Statistical Methods for Evidence Synthesis

Statistical Methods for Evidence Synthesis

a dissertation presented
by
Maya B. Mathur
to
The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts
August 2018

Dissertation advisor: Prof. Tyler J. VanderWeele Maya B. Mathur

Statistical Methods for Evidence Synthesis

Abstract

In many empirical disciplines, scientific discovery is modularized into discrete papers each investigating one or more hypotheses. Synthesizing these modules of evidence is critical to inform a balanced and appropriately evolving view of the overall evidence on a topic as well as to identify where substantial uncertainty remains. This dissertation considers three realms in which such synthesis can occur: (1) when meta-analyzing multiple studies; (2) when subjecting a single study to independent replications; and (3) when testing related hypotheses within a study. We consider specific methodological challenges within each of these realms and propose statistical methods to address each. All proposed methods are implemented in R packages.

Contents

List of Figures

List of Tables

To my parents, for all the science then and now.

Acknowledgments

I am deeply grateful to my doctoral advisor, Tyler VanderWeele, whose outstanding mentorship has been formative for me. His intellectual bravery and scientific integrity are inspiring. It has been an honor to work with my committee members, Sebastien Haneuse and Nan Laird, with whom I had interesting and illuminating discussions that shaped this dissertation and deepened my understanding. Tom Chen's incisive comments improved Chapter 3. Ying Chen provided advice and analysis code for the applied example in Chapter 3.

I was supported by National Defense Science and Engineering Graduate Fellowship 32 CFR 168a. Tyler VanderWeele was supported by NIH grants ES017876 and CA222147. The MIDUS study was funded by the John D. and Catherine T. MacArthur Foundation Research Network and the National Institute on Aging (P01-AG020166 and U19-AG051426). The funders had no role in the design, conduct, or reporting of this research.

1

Sensitivity Analysis for Unmeasured Confounding in Meta-Analyses

1.1 Abstract

Random-effects meta-analyses of observational studies can produce biased estimates if the synthesized studies are subject to unmeasured confounding. We propose sensitivity analyses quantifying the extent to which unmeasured confounding of specified magnitude could reduce to below a certain threshold the proportion of true effect sizes that are scientifically meaningful. We also develop converse methods to estimate the strength of confounding capable of reducing the proportion of scientifically meaningful true effects to below a chosen threshold. These methods apply when a “bias factor” is assumed to be normally distributed across studies or is assessed across a range of fixed values. Our estimators are derived using recently proposed sharp bounds on confounding bias within a single study that do not make assumptions regarding the unmeasured confounders themselves or the functional form of their relationships to the exposure and outcome of interest. We provide an R package, EValue, and a free website that compute point estimates and inference and produce plots for conducting such sensitivity analyses. These methods facilitate principled use of random-effects meta-analyses of observational studies to assess the strength of causal evidence for a hypothesis.

1.2 Introduction

Meta-analyses can be indispensable for assessing the overall strength of evidence for a hypothesis and for precisely estimating effect sizes through aggregation of multiple estimates. Meta-analysis is often used not only for randomized trials, but also for observational studies. When the hypothesis of interest is about causation (for example, of an exposure on a health outcome), evidence strength depends critically not only on the size and statistical uncertainty of the meta-analytic point estimate, but also on the extent to which these apparent effects are robust to unmeasured confounding [34, 92, 108]. However, when well-designed randomized studies do not exist because the exposure cannot be randomized, meta-analyses often comprise potentially confounded observational studies. Therefore, in practice, meta-analyses of observational studies are often met with concerns about the potential for unmeasured confounding to attenuate – or possibly even reverse the direction of – the estimated effects (e.g., [18], [6], and [99] with critiques on the latter by [100]). Yet such considerations rarely proceed beyond qualitative speculation given the limited availability of quantitative methods to assess the impact of unmeasured confounding in a meta-analysis.

Our focus in this paper is therefore on conducting sensitivity analyses assessing the extent to which unmeasured confounding of varying magnitudes could have compromised the results of the meta-analysis. Existing sensitivity analyses for confounding bias or other internal biases in meta-analysis estimate a bias-corrected pooled point estimate by directly incorporating one or more bias parameters in the likelihood and placing a Bayesian prior on the distribution of these parameters [67, 119]. An alternative frequentist approach models bias as additive or multiplicative within each study and then uses subjective assessment to elicit study-specific bias parameters [107]. Although useful, these approaches typically require strong assumptions on the nature of unmeasured confounding (for example, requiring a single binary confounder), rely on the arbitrary specification of additive or multiplicative effects of bias, or require study-level estimates rather than only meta-analytic pooled estimates. Furthermore, the specified bias parameters do not necessarily lead to precise practical interpretations.

An alternative approach is to analytically bound the effect of unmeasured confounding on the results of a meta-analysis. To this end, bounding methods are currently available for point estimates of individual studies. We focus on sharp bounds derived by [29] because of their generality and freedom from assumptions regarding the nature of the unmeasured confounders or the functional forms of their relationships with the exposure of interest and outcome. This approach subsumes several earlier approaches [22, 37, 87] and, in contrast to [63] and [114], does not make any no-interaction assumptions between the exposure and the unmeasured confounder(s).

The present paper extends these analytic bounds for single studies to the meta-analytic setting. Using standard estimates from a random-effects meta-analysis and intuitively interpretable sensitivity parameters on the magnitude of confounding, these results enable inference about the strength of causal evidence in a potentially heterogeneous population of studies. Broadly, our approach proceeds as follows. First, we select an effect size representing a minimum threshold of scientific importance for the true causal effect in any given study. Second, we use the confounded effect estimates from the meta-analyzed studies, along with simple sensitivity parameters, to make inference to the population distribution of true causal effects (the quantities of ultimate scientific interest). Lastly, we use this estimated distribution in turn to estimate the proportion of true causal effects in the population that are of scientifically meaningful size (that is, those stronger than the chosen threshold). As we will discuss, the proportion of scientifically meaningful effect sizes in a meta-analysis is a useful characterization of evidence strength when the effects may be heterogeneous. Conversely, we also solve for the sensitivity parameters on the bias that would be capable of "explaining away" the results of the meta-analysis by substantially reducing the proportion of strong causal effects. We also discuss sensitivity analysis for the pooled estimate of the mean effect.

If sensitivity analysis for unmeasured confounding indicates that only a small proportion of true causal effects are stronger than the chosen threshold of scientific importance, then arguably the results of the meta-analysis are not robust to unmeasured confounding in a meaningful way regardless of the "statistical significance" of the observed point estimate. To this end, we develop estimators that answer the questions: "In the presence of unmeasured confounding of specified strength, what proportion of studies would have true causal effects of scientifically meaningful size?" and "How severe would unmeasured confounding need to be 'explain away' the results; that is, to imply that very few causal effects are of scientifically meaningful size?" This approach to sensitivity analysis is essentially a meta-analytic extension of a recently proposed metric (the E-value) that quantifies, for a single study, the minimum confounding bias capable of reducing the true effect to a chosen threshold [111]. We provide and demonstrate use of an R package (EValue) and a free website for conducting such analyses and creating plots.

1.3 Existing bounds on confounding bias in a single study

[29] developed bounds for a single study as follows. Let X denote a binary exposure, Y a binary outcome, Z a vector of measured confounders, and U one or more unmeasured confounders. Let:

$$RR^{c}_{XY|z} = \frac{P\left(Y=1 \mid X=1, Z=z\right)}{P\left(Y=1 \mid X=0, Z=z\right)}$$

be the confounded relative risk (RR) of Y for $X=1$ versus $X=0$ conditional or stratified on the measured confounders $Z=z$.

Let its true, unconfounded counterpart standardized to the population be:

$$RR^{t}_{XY|z} = \frac{\sum_u P\left(Y=1 \mid X=1, Z=z, U=u\right) P\left(U=u \mid Z=z\right)}{\sum_u P\left(Y \mid X=0, Z=z, U=u\right) P\left(U=u \mid Z=z\right)}$$

(Throughout, we use the term “true” as a synonym for “unconfounded” or “causal” when referring to both sample and population quantities. Also, henceforth, we condition implicitly on $Z=z$, dropping the explicit notation for brevity.) Define the ratio of the confounded to the true relative risks as $B = RR^{c}_{XY} / RR^{t}_{XY}$.

Let $RR_{Xu} = P\left(U=u \mid X=1\right) / P\left(U=u \mid X=0\right)$. Define the first sensitivity parameter as $RR_{XU} = \max_u \left(RR_{Xu}\right)$; that is, the maximal relative risk of $U=u$ for $X=1$ versus $X=0$ across strata of U. (If U is binary, this is just the relative risk relating X and U.) Next, for each stratum x of X, define a relative risk of Y on U, maximized across all possible contrasts of U:

$$RR_{UY|X=x} = \frac{\max_u P\left(Y = 1|X = x, U = u\right)}{\min_u P\left(Y = 1|X = x, U = u\right)}, x \in \{0, 1\}$$

Define the second sensitivity parameter as $RR_{UY} = \max\left(RR_{UY|X=0}, RR_{UY|X=1}\right)$. That is, considering both strata of X, it is the largest of the maximal relative risks of Y on U conditional on X. Then, Ding and VanderWeele [29] showed that when $B \geq 1$, it is bounded above by:

$$B \leq \frac{RR_{XU} \cdot RR_{UY}}{RR_{XU} + RR_{UY} - 1}$$

and that when $B \leq 1$, the same bound holds for $1/B$. Thus, defining the "worst-case" bias factor as $B^+ = \frac{RR_{XU} \cdot RR_{UY}}{RR_{XU}+RR_{UY}-1}$, a sharp bound for the true effect is:

$$RR^t_{XY} \geq RR^c_{XY} \Big/ B^+ \tag{1.1}$$

This bound on the bias factor applies when examining the extent to which unmeasured confounding might have shifted the observed estimate RR^c_{XY} away from the null. Thus, Equation (1.1) indicates that RR^t_{XY} is at least as strong as a bound constructed by attenuating RR^c_{XY} toward the null by a factor of B^+. The factor B^+ is larger, indicating greater potential bias, when U is strongly associated with both X and Y (i.e., RR_{XU} and RR_{UY} are large) and is equal to 1, indicating no potential for bias, if U is unassociated with either X or Y (i.e., $RR_{XU} = 1$ or $RR_{UY} = 1$).

If the two sensitivity parameters are equal ($RR_{XU} = RR_{UY}$), then to produce a worst-case bias factor B^+, each must exceed $B^+ + \sqrt{B^+\left(B^+ - 1\right)}$ (which VanderWeele and Ding [111] call the "E-value"). Thus, a useful transformation of B^+ is the "confounding strength scale", g, which is the minimum size of RR_{XU} and RR_{UY} under the assumption that they are equal:

$$g = B^+ + \sqrt{B^+ (B^+ - 1)} \quad \Leftrightarrow \quad B^+ = \frac{g^2}{2g - 1} \tag{1.2}$$

If $RR^c_{XY} < 1$ (henceforth the "apparently preventive case"), then Equation (1.1) becomes [29]:

$$RR^t_{XY} \leq RR^c_{XY} \cdot \frac{RR^*_{XU} \cdot RR_{UY}}{RR^*_{XU} + RR_{UY} - 1}$$

where $RR^*_{XU} = \max_u \left(RR^{-1}_{Xu} \right)$, i.e., the maximum of the inverse relative risks, rather than the relative risks themselves. Thus, B^+ remains ≥ 1, and we have $RR^t_{XY} \geq RR^c_{XY}$.

Although these results hold for multiple confounders, in the development to follow, we will use a single, categorical unmeasured confounder for clarity. However, all results can easily be interpreted without assumptions on the type of exposure and unmeasured confounders, for instance by interpreting the relative risks defined above as "mean ratios" [29].

1.4 Random-effects meta-analysis setting

In this paper, we use the aforementioned analytic bounds to derive counterparts for random-effects meta-analysis. Under standard parametric assumptions [103], each of k studies measures a potentially unique effect size M_i, such that $M_i \sim_{iid} N(\mu, V)$ for a grand mean μ and variance V. Let y_i be the point estimate of the i^{th} study and σ^2_i be the within-study variance (with the latter assumed fixed and known), such that $y_i \mid M_i \ \sim N(M_i, \sigma^2_i)$. Thus, marginally, $y_i \sim N(M_i, V + \sigma^2_i)$.

Analysis proceeds by first estimating V via one of many possible estimators, denoted τ^2. Heterogeneity estimation approaches include, for example, maximum likelihood and restricted maximum likelihood as well as approaches proposed by [78], [93], [44], and [46]; see [116] for a review. We will denote an estimator of μ by $\widehat{y}_R$, which, for many estimators, will also be a function of τ^2. For example, a common approach is to use the maximum likelihood solutions for the two parameters[1]:

$$\widehat{y}_R = \frac{\sum_{i=1}^{k} w_i \, y_i}{\sum_{i=1}^{k} w_i} \tag{1.3}$$

$$\tau^2 = \max\left\{0, \frac{\sum_{i=1}^{k} w_i^2 [(y_i - \widehat{y}_R)^2 - \sigma_i^2]}{\sum_{i=1}^{k} w_i^2}\right\} \tag{1.4}$$

The weights, w_i, are inversely proportional to the total variance of each study (a sum of the between-study variance and the within-study variance), such that $w_i = 1/\left(\tau^2 + \sigma_i^2\right)$. Estimation can then proceed by first initializing $\widehat{y}_R$ and τ^2 to, for example, the weighted mean assuming $\tau^2 = 0$ and the method of moments estimators, respectively, and then by iterating between (1.3) and (1.4) to reach the maximum likelihood solutions [116]. Other estimation procedures exist (see [116] for a review), and our methods apply regardless of estimation procedure as long as $\widehat{y}_R$ and τ^2 are consistent and unbiased, asymptotically normal, and asymptotically independent.

1.5 Main results

Consider k studies measuring relative risks with confounded population effect sizes on the log-RR scale, denoted M^c, such that $M^c \sim N(\mu^c, V^c)$. (Other outcome measures are considered briefly in the

[1] The maximum likelihood solution for $\widehat{y}_R$ coincides with the classical moments estimator [28], so in practice, widespread methods for random-effects meta-analysis differ primarily in estimation of τ^2.

Discussion.) For studies in which some confounders are measured and adjusted in analysis, we define M^c as the population effect sizes after adjusting for these measured confounders, but without adjusting for any unmeasured confounders. Let the corresponding true effects be M^t with expectation μ^t and variance V^t. Let $\widehat{y_R^c}$ be the pooled point estimate and τ_c^2 be a heterogeneity estimate, both computed from the confounded point study estimates (for example, from Equations (1.3) and (1.4)).

Consider the bias factor on the log scale, $B^* = \log\left(\frac{RR_{XU} \cdot RR_{UY}}{RR_{XU}+RR_{UY}-1}\right)$, and allow it to vary across studies under the assumption that $B^* \sim N\left(\mu_{B^*}, \sigma^2_{B^*}\right)$, with B^* independent of M^t. That is, we assume that the bias factor is independent of the true effects but not the confounded effects: naturally, studies with larger bias factors will tend to obtain larger effect sizes. For studies in which analyses conditioned on one or more measured confounders, B^* represents additional bias produced by unmeasured confounding, above and beyond the measured confounders. Hence, studies with better existing control of confounding are likely to have a smaller value of B^* than studies with poor confounding control. The normality assumption on the bias factor holds approximately if, for example, its components (RR_{XU} and RR_{UY}) are identically and independently normal with relatively small variance (Appendix). We now develop three estimators enabling sensitivity analyses.

1.5.1 Proportion of studies with large effect sizes as a function of the bias factor

For an apparently causative relative risk ($\widehat{y_R^c} > 0$, or equivalently the confounded pooled RR is greater than 1), define $p(q) = P\left(M^t > q\right)$ for any threshold q, i.e., the proportion of studies with true effect sizes larger than q. Then a consistent estimator of $p(q)$ is:

$$\widehat{p}(q) = 1 - \Phi\left(\frac{q + \mu_{B^*} - \widehat{y}_R^c}{\sqrt{\tau_c^2 - \sigma_{B^*}^2}}\right), \ \tau_c^2 > \sigma_{B^*}^2$$

where Φ denotes the standard normal cumulative distribution function. In the special case in which the bias factor is fixed to μ_{B^*} across all studies, the same formula applies with $\sigma_{B^*}^2 = 0$.

Many common choices of heterogeneity estimators, τ_c^2, are asymptotically independent of $\widehat{y}_R^c$ (Appendix), an assumption used for all standard errors in the main text. Results relaxing this assumption appear throughout the Appendix. An application of the delta method thus yields an approximate standard error:

$$\widehat{\text{SE}}\left(\widehat{p}(q)\right) \approx \sqrt{\frac{\widehat{\text{Var}}\left(\widehat{y}_R^c\right)}{\tau_c^2 - \sigma_{B^*}^2} + \frac{\widehat{\text{Var}}\left(\tau_c^2\right)\left(q + \mu_{B^*} - \widehat{y}_R^c\right)^2}{4\left(\tau_c^2 - \sigma_{B^*}^2\right)^3}} \cdot \varphi\left(\frac{q + \mu_{B^*} - \widehat{y}_R^c}{\sqrt{\tau_c^2 - \sigma_{B^*}^2}}\right)$$

where φ denotes the standard normal density function. (If $\tau_c^2 \leq \sigma_{B^*}^2$, leaving one of the denominators undefined, this indicates that there is so little observed heterogeneity in the confounded effect sizes that, given the specified bias distribution, V^t is estimated to be less than 0. Therefore, attention should be limited to a range of values of $\sigma_{B^*}^2$ such that $\tau_c^2 > \sigma_{B^*}^2$.)

For an apparently preventive relative risk ($\widehat{y}_R^c < 0$ or the confounded pooled RR is less than 1), define instead $p(q) = P\left(M^t < q\right)$, i.e., the proportion of studies with true effect sizes less than q. Then a consistent estimator is:

$$\widehat{p}(q) = \Phi\left(\frac{q - \mu_{B^*} - \widehat{y}_R^c}{\sqrt{\tau_c^2 - \sigma_{B^*}^2}}\right), \ \tau_c^2 > \sigma_{B^*}^2$$

with approximate standard error:

$$\widehat{\mathrm{SE}}\left(\widehat{p}(q)\right) = \sqrt{\frac{\widehat{\mathrm{Var}}\left(\widehat{y}^c_R\right)}{\tau_c^2 - \sigma^2_{B^*}} + \frac{\widehat{\mathrm{Var}}\left(\tau_c^2\right)\left(q - \mu_{B^*} - \widehat{y}^c_R\right)^2}{4\left(\tau_c^2 - \sigma^2_{B^*}\right)^3}} \cdot \varphi\left(\frac{q - \mu_{B^*} - \widehat{y}^c_R}{\sqrt{\tau_c^2 - \sigma^2_{B^*}}}\right)$$

Because $\widehat{p}(q)$ is monotonic in $\sigma^2_{B^*}$, the homogeneous bias case (i.e., $\sigma^2_{B^*} = 0$) provides either an upper or lower bound on $\widehat{p}(q)$ (Table 1.6.1). We later return to the practical utility of these results.

1.5.2 Bias factor required to reduce proportion of large effect sizes to a threshold

Conversely, we might consider the minimum common bias factor (on the RR scale) capable of reducing to less than r the proportion of studies with true effect exceeding q. We accordingly define $T(r,q) = B^+ : P\left(M^t > q\right) = r$ to be this quantity, with B^+ taken to be constant across studies. (Note that taking B^+ to be constant does not necessarily imply that the unmeasured confounders themselves are identical across studies.) Then for an apparently causative relative risk, a consistent estimator for the minimum common bias capable of reducing to less than r the proportion of studies with effects surpassing q is:

$$\widehat{T}(r,q) = \exp\left\{\Phi^{-1}(1-r)\sqrt{\tau_c^2} - q + \widehat{y}^c_R\right\}$$

with approximate standard error:

$$\widehat{\mathrm{SE}}\left(\widehat{T}(r,q)\right) = \exp\left\{\sqrt{\tau_c^2}\left(\Phi^{-1}(1-r)\right) - q + \widehat{y}^c_R\right\}\sqrt{\widehat{\mathrm{Var}}\left(\widehat{y}^c_R\right) + \frac{\widehat{\mathrm{Var}}\left(\tau_c^2\right)\left(\Phi^{-1}(1-r)\right)^2}{4\tau_c^2}}$$

For an apparently preventive relative risk, we can instead consider the minimum common bias factor (on the RR scale) capable of reducing to less than r the proportion of studies with true effect less than q, thus defining $T(r,q) = B^{+} : P\left(M^{t} < q\right) = r$. Then a consistent estimator is:

$$\widehat{T}(r,q) = \exp\left\{q - \widehat{y}_R^c - \Phi^{-1}(r)\sqrt{\tau_c^2}\right\}$$

with approximate standard error:

$$\widehat{\text{SE}}\left(\widehat{T}(r,q)\right) = \exp\left\{q - \widehat{y}_R^c - \sqrt{\tau_c^2}\left(\Phi^{-1}(r)\right)\right\}\sqrt{\widehat{\text{Var}}\left(\widehat{y}_R^c\right) + \frac{\widehat{\text{Var}}\left(\tau_c^2\right)\left(\Phi^{-1}(r)\right)^2}{4\tau_c^2}}$$

1.5.3 Confounding strength required to reduce proportion of large effect sizes to a threshold

Under the assumption that the two components of the common bias factor are equal as in Equation (1.2), such that $g = RR_{XU} = RR_{UY}$, the bias can alternatively be parameterized on the confounding strength scale. Consider the minimum confounding strength required to lower to less than r the proportion of studies with true effect exceeding q and accordingly define $G(r,q) = g : P\left(M^{t} > q\right) = r$. For both the apparently causative and the apparently preventive cases, an application of Equation (1.2) yields:

$$\widehat{G}(r,q) = \widehat{T}(r,q) + \sqrt{\left(\widehat{T}(r,q)\right)^2 - \widehat{T}(r,q)}$$

with approximate standard error:

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

1.6 Practical use and interpretation

1.6.1 Interpreting $\widehat{p}(q)$

To conduct our first proposed sensitivity analysis, one first assumes a simple distribution on the amount of confounding bias in the meta-analyzed studies, leading to the specification of a pair of sensitivity parameters, μ_{B^*} and $\sigma^2_{B^*}$. Then, one computes $\widehat{p}(q)$ to gauge the strength of evidence for causation if confounding bias indeed follows the specified distribution. As mentioned in the Introduction, we consider the proportion of true effects above a chosen threshold of scientific importance because this metric characterizes evidence strength while taking into account the effect heterogeneity that is central to the random-effects meta-analysis framework. That is, a large proportion of true effect sizes stronger than a threshold of scientific importance in a meta-analysis (e.g., 70% of true effects stronger than the threshold $RR = 1.10$, i.e. $q = \log 1.10$) suggests that, although the true causal effects may be heterogeneous across studies, there is evidence that overall, many of these effects are strong enough to merit scientific interest. If $\widehat{p}(q)$ remains large for even large values of μ_{B^*}, this indicates that even if the influence of unmeasured confounding were substantial, a large proportion of true effects in the population distribution would remain of scientifically meaningful magnitude. Thus, the results of the meta-analysis might be considered relatively robust to unmeasured confounding.

1.6.2 How to choose q, μ_{B^*}, and $\sigma^2_{B^*}$ when computing $\widehat{p}(q)$

The threshold q allows the investigator to flexibly define how much attenuation in effect size due to confounding bias would render a causal effect too weak to be considered scientifically meaningful. A general guideline might be to use $q = \log 1.10$ as a minimum threshold for an apparently causative relative risk or $q = \log 0.90$ for an apparently preventive relative risk. Because μ_{B^*} and $\sigma^2_{B^*}$ are sensitivity parameters that are not estimable from the data, we would recommend reporting $\widehat{p}(q)$ for a wide range of values of μ_{B^*} (including large values, representing substantial confounding bias) and with $\sigma^2_{B^*}$ ranging from 0 to somewhat less than τ^2_c.

To provide intuition for what values of μ_{B^*} and $\sigma^2_{B^*}$ might be plausible in a given setting, it can be useful to consider the implied range of bias factors across studies for a given pair of sensitivity parameters. For example, if $\mu_{B^*} = \log 1.20$ and $\sigma^2_{B^*} = 0.01$, so that the standard deviation of the bias on the log scale is 0.10, these choices of sensitivity parameters imply that 95% of the studies have B (on the risk ratio scale) between $\exp\left(\mu_{B^*} - \Phi^{-1}(0.975) \times \sigma_{B^*}\right) = 0.98$ and $\exp\left(\mu_{B^*} + \Phi^{-1}(0.975) \times \sigma_{B^*}\right) = 1.46$. This choice of sensitivity parameters may be reasonable, then, if one is willing to assume that studies very rarely (with approximately 2.5% probability) obtain point estimates that are inflated by more than 1.46-fold due to unmeasured confounding, and furthermore that studies very rarely obtain point estimates that are biased toward, instead of away from, the null (which requires $B < 1$). If, in contrast, an assessment of study design quality suggests that some studies in the meta-amalysis might have more severely biased point estimates, then one might consider increasing μ_{B^*} or $\sigma^2_{B^*}$. The choice of $\sigma^2_{B^*}$ can also be informed by the extent to which the meta-analyzed studies differ with respect to existing confounding control. When some studies have much better confounding control than others, then B^* may vary substantially, so a larger $\sigma^2_{B^*}$ may be reasonable. When all studies adjust for similar sets of confounders and use similar populations, then a small $\sigma^2_{B^*}$ may be reasonable.

Lastly, bounds achieved when $\sigma^2_{B^*} = 0$ can provide useful conservative analyses. Table 1.6.1 shows that setting $\sigma^2_{B^*} = 0$ yields either an upper or lower bound on $\widehat{p}(q)$, where the latter allows $\sigma^2_{B^*} > 0$. The direction of the bound depends on whether $\widehat{y}^c_R$ is apparently causative or preventive and on whether q is chosen to be on the lower or upper tail of the bias-corrected pooled point estimate, defined as $\widehat{y}^t_R = \widehat{y}^c_R - \mu^*_B$ for the apparently causative case and $\widehat{y}^t_R = \widehat{y}^c_R + \mu^*_B$ for the apparently preventive case. For example, for $\widehat{y}^c_R > 0$ and $q > \widehat{y}^c_R - \mu^*_B$, the $\sigma^2_{B^*} = 0$ case provides an upper bound on $\widehat{p}(q)$. When concluding that results are not robust to unmeasured confounding, the analysis with $\sigma^2_{B^*} = 0$ is therefore conservative in that fewer true effect sizes would surpass q under heterogeneous bias. For example, if we calculated $\widehat{p}\left(q = \log 1.10\right) = 0.15$ with $\mu_{B^*} = \log 1.20$ and $\sigma^2_{B^*} = 0$, then an analysis like this would yield conclusions such as: "The results of this meta-analysis are relatively sensitive to unmeasured confounding. Even a bias factor as small as 1.20 in each study would reduce to only 15% the proportion of studies with true relative risks greater than 1.10, and if the bias in fact varied across studies, then even fewer studies would surpass this effect size threshold."

Table 1.6.1: Bounds on $\widehat{p}(q)$ provided by homogeneous bias with an apparently causative or preventive pooled effect. $\widehat{y}^t_R$ estimates μ^t and is equal to $\widehat{y}^c_R - \mu_{B^*}$ for $\widehat{y}^c_R > 0$ or $\widehat{y}^c_R + \mu_{B^*}$ for $\widehat{y}^c_R < 0$.

	$q > \widehat{y}^t_R$	$q < \widehat{y}^t_R$
$\widehat{y}^c_R > 0$	Upper bound	Lower bound
$\widehat{y}^c_R < 0$	Lower bound	Upper bound

1.6.3 Interpreting $\widehat{T}(r, q)$ and $\widehat{G}(r, q)$

In contrast to $\widehat{p}(q)$, the metrics $\widehat{T}(r, q)$ and $\widehat{G}(r, q)$ do not require specification of a range of sensitivity parameters regarding the bias distribution. Instead, they solve for the minimum amount of bias that, if constant across all studies, would "explain away" the effect in a manner specified through q (the minimum threshold of scientific importance) and r (the minimum proportion of true effects above q). That is, we might say that unmeasured confounding has, for practical purposes, "explained away" the results of a meta-analysis if fewer than, for example, 10% of the true effects are stronger than a threshold of $RR = 1.10$, in which case we would set $r = 0.10$ and $q = \log 1.10$.

A large value of either $\widehat{T}(r,q)$ or $\widehat{G}(r,q)$ indicates that it would take substantial unmeasured confounding (i.e., a large bias factor as parameterized by $\widehat{T}(r,q)$ or a large strength of confounding as parameterized by $\widehat{G}(r,q)$) to "explain away" the results of the meta-analysis in this sense, and that weaker unmeasured confounding could not do so. Thus, the results may be considered relatively robust to unmeasured confounding. For example, by choosing $q = \log(1.10)$ and $r = 0.20$ and computing $\widehat{T}(r,q) = 2.50$ (equivalently, $\widehat{G}(r,q) = 4.44$), one might conclude: "The results of this meta-analysis are relatively robust to unmeasured confounding, insofar as a bias factor of 2.50 on the relative risk scale (e.g., a confounder associated with the exposure and outcome by risk ratios of 4.44 each) in each study would be capable of reducing to less than 20% the proportion of studies with true relative risks greater than 1.10, but weaker confounding could not do so." On the other hand, small values of $\widehat{T}(r,q)$ and $\widehat{G}(r,q)$ indicate that only weak unmeasured confounding would be required to reduce the effects to a scientifically unimportant level; the meta-analysis would therefore not warrant strong scientific conclusions regarding causation.

1.6.4 How to choose q and r when computing $\widehat{T}(r,q)$ and $\widehat{G}(r,q)$

When computing $\widehat{T}(r,q)$ and $\widehat{G}(r,q)$, one can use the same effect size threshold q as discussed above for computing $\widehat{p}(q)$. When the number of studies, k, is large (for example, ≥ 10), one might require at least 10% of studies ($r = 0.10$) to have effect sizes above q for results to be of scientific interest. For $k < 10$, one might select a higher threshold, such as $r = 0.20$ (thus requiring at least 20% of studies to have effects more extreme than, for example, $\log 1.10$). Of course, these guidelines can and should be adapted based on the substantive application. Furthermore, note that the amount of bias that would be considered "implausible" must be determined with attention to the design quality of the synthesized studies: a large bias factor may be plausible for a set of studies with poor confounding control and with high potential for unmeasured confounding, but not for a set of better-designed studies in which the measured covariates already provide good control of confounding.

1.7 Further remarks on heterogeneity

We operationalized "robustness to unmeasured confounding" as the proportion of true effects surpassing a threshold, an approach that focuses on the upper tail (for an apparently causative RR^c_{XY}) of the distribution of true effect sizes. Potentially, under substantial heterogeneity, a high proportion of true effect sizes could satisfy, for example, $RR^t_{XY} > 1.10$ while, simultaneously, a non-negligible proportion could be comparably strong in the opposite direction ($RR^t_{XY} < 0.90$). Such situations are intrinsic to the meta-analysis of heterogeneous effects, and in such settings, we recommend reporting the proportion of effect sizes below another threshold on the opposite side of the null (e.g., $\log 1/1.20 \approx \log 0.80$) both for the confounded distribution of effect sizes and for the distribution adjusted based on chosen bias parameters. For example, a meta-analysis that is potentially subject to unmeasured confounding and that estimates $\widehat{y}^c_R = \log 1.15$ and $\tau^2_c = 0.10$ would indicate that 45% of the effects RR^c_{XY} surpass 1.20, while 13% are less than 0.80. For a common $B^* = \log 1.10$ (equivalently, $g = 1.43$), we find that $\left(1 - \Phi\left(\frac{\log 1.20 - \log 1.15 + \log 1.10}{\sqrt{0.10}}\right)\right) \cdot 100\% = 33\%$ of the true effects surpass $RR^c_{XY} = 1.20$, while 20% are less than $RR^c_{XY} = 0.80$. More generally, random-effects meta-analyses could report the estimated proportion of effects above the null or above a specific threshold (along with a confidence interval for this proportion) as a continuous summary measure to supplement the standard pooled estimate and inference. Together, these reporting practices could facilitate overall assessment of evidence strength and robustness to unmeasured confounding under effect heterogeneity.

1.8 Sensitivity analysis for the point estimate

As discussed above, the proportion of effects stronger than a threshold can be a useful measure of evidence strength across heterogeneous effects in addition to pooled point estimate alone, and hence our sensitivity analysis techniques have emphasized the former. However, it is also possible to conduct sensitivity analysis on the pooled point estimate itself to assess the extent to which unmeasured confounding could compromise estimation of μ^t. The following development proceeds analogously to that of Section 1.5.

1.8.1 An adjusted point estimate as a function of the bias factor

For an apparently causative relative risk and a specified μ_{B^*}, an unbiased estimate of the true mean, μ^t, is simply $\widehat{y^t_R} = \widehat{y^c_R} - \mu_{B^*}$. For an apparently preventive relative risk, it is $\widehat{y^t_R} = \widehat{y^c_R} + \mu_{B^*}$. Because these expressions consider the average true effect only, they do not involve bias correction of τ^2_c, so are independent of $\sigma^2_{B^*}$. Since $\text{Var}\left(\widehat{y^t_R}\right) = \text{Var}\left(\widehat{y^c_R}\right)$, inference on $\widehat{y^t_R}$ can use without modification the standard error estimate for $\widehat{y^c_R}$ computed through standard meta-analysis of the confounded data. For example, [45]'s estimation approach yields:

$$\widehat{\text{SE}}\left(\widehat{y}_R^t\right) = \sqrt{\frac{\sum_{i=1}^k \frac{1}{\tau_c^2+\sigma_i^2}\left(y_i^c - \widehat{y}_R^c\right)^2}{(k-1)\sum_{i=1}^k \frac{1}{\tau_c^2+\sigma_i^2}}}$$

where y_i^c is the confounded log-relative risk estimate in the i^{th} study.

1.8.2 Bias factor and confounding strength required to shift the point estimate to the null

One could instead consider the value of μ_{B^*} that would be required to "explain away" the point estimate. That is, to completely shift the point estimate to the null (i.e., $\mu^t = 0$, implying an average risk ratio of 1) would require $\mu_{B^*} = \widehat{y}_R^c$. As in Section 1.5.3, the bias factor can be converted to the more intuitive confounding strength scale via Equation 1.2. Thus, the minimum confounding strength to completely shift the point estimate to the null is, for the apparently causative case:

$$\exp\left(\widehat{y}_R^c\right) + \sqrt{\exp\left(\widehat{y}_R^c\right)\left[\exp\left(\widehat{y}_R^c\right) - 1\right]} \tag{1.5}$$

Additionally, one can consider the confounding strength required to shift the confidence interval for $\widehat{y}_R^c$ to include the null; to do so, $\widehat{y}_R^c$ in the above expression would simply be replaced with the confidence bound closer to the null. (For the apparently preventive case, whether considering the point estimate or the confidence interval bound, each exponentiated term in Equation 1.5 would be replaced by its inverse.) As above, these measures do not describe heterogeneity. Thus, Equation 1.5 is in fact equivalent to [111]'s E-value (as discussed in Section 1.3) applied directly to $\widehat{y}_R^c$, as illustrated in the next section.

1.9 Software and applied example

The proposed methods (as well as those discussed in Section 1.8 above) are implemented in an R package, EValue, which produces point estimates and inference for sensitivity analyses, tables across a user-specified grid of sensitivity parameters, and various plots. Descriptions of each function with working examples are provided in the Appendix and standard R documentation. A website implementing the main functions is freely available (https://mmathur.shinyapps.io/meta_gui_2/).

We illustrate the package's basic capabilities using an existing meta-analysis assessing, among several outcomes, the association of high versus low daily intake of soy protein with breast cancer risk among women [106]. The analysis comprised 20 observational studies that varied in their degree of adjustment for suspected confounders, such as age, body mass index (BMI), and other risk factors. To obtain τ_c^2 and $\widehat{\text{Var}}(\tau_c^2)$ (which were not reported), we obtained study-level summary measures as reported in a table from [106], approximating odds ratios with risk ratios given the rare outcome. This process is automated in the function EValue::scrape_meta. We estimated $\widehat{y_R^c} = \log 0.82$, $\widehat{\text{SE}}\left(\widehat{y_R^c}\right) = 8.8 \times 10^{-2}$ via the [45] adjustment (whose advantages were demonstrated by [50]), $\tau_c^2 = 0.10$ via the [78] method, and $\widehat{\text{SE}}\left(\tau_c^2\right) = 5.0 \times 10^{-2}$.

Figure 1.9.1 (produced by EValue::sens_plot) displays the estimated proportion of studies with true relative risks < 0.90 as a function of either the bias factor or the confounding strength, holding constant $\sigma_{B^*}^2 = 0.01$. Table 1.9.1 (produced by EValue::sens_table) displays $\widehat{T}(r,q)$ and $\widehat{G}(r,q)$ across a grid of values for r and q. For example, only a bias factor exceeding 1.63 on the relative risk scale (equivalently, confounding association strengths of 2.64) could reduce

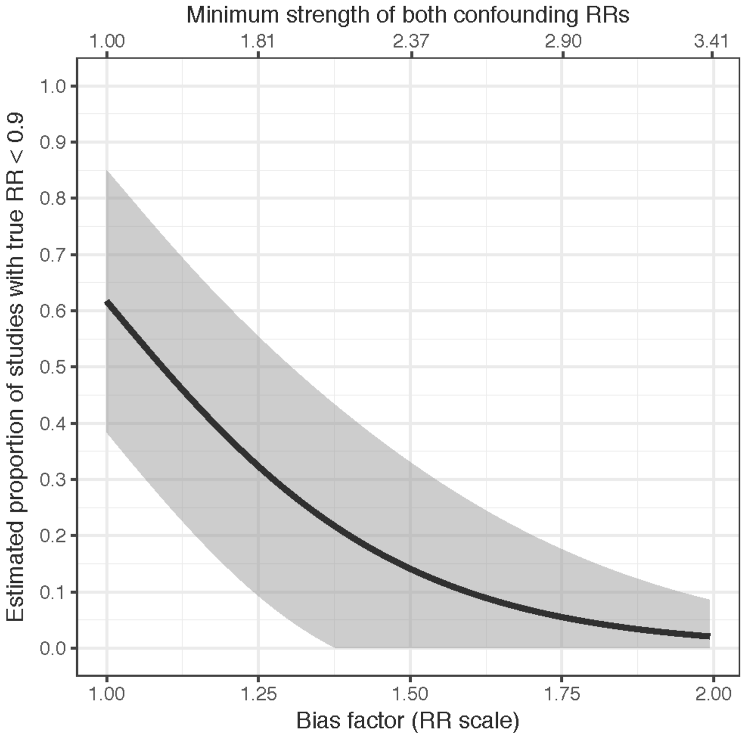

Figure 1.9.1: Impact of varying degrees of unmeasured confounding bias on proportion of true relative risks < 0.90

to less than 10% the proportion of studies with true relative risks < 0.90. However, variable bias across studies would reduce this proportion, and the confidence interval is wide.

We now briefly illustrate the sensitivity analysis techniques for $\widehat{y}_R^c$ described in Section 1.8. For example, applying Equation (1.5) indicates that an unmeasured confounder associated with both soy intake and breast cancer by risk ratios of at least 1.72 could be sufficient to shift the point estimate ($RR_{XY}^c = 0.82$) to 1, but weaker confounding could not do so [111]. To reiterate the remarks made in Section 1.7 regarding heterogeneity, note that our proposed sensitivity analyses found $\widehat{G}(r = 0.10, q = \log 0.90) = 2.64$. This is considerably larger than the E-value of 1.72 for the point estimate, demonstrating that even in the presence of unmeasured confounding strong enough to shift the point estimate to the null, more than 10% of the true relative risks would nevertheless remain stronger than 0.90.

Other methods developed for a single study could similarly be applied to the meta-analytic point estimate, but they require specification of many more sensitivity parameters or make more assumptions about the underlying unmeasured confounder (e.g., [87]; [49]; [63, 114]). To apply these methods directly, we use a simplified form assuming that U is binary, that the prevalences $P(U = 1 \mid X = 1, Z) = 0.65$ and $P(U = 1 \mid X = 0, Z) = 0.35$ are in fact known, and that the relationship between U and Y is identical for $X = 1$ and $X = 0$. Under this more restrictive specification on unmeasured confounding, an application of [87]'s method (or an application of a special case of Theorem 2 by [114]) finds that such a confounder would exactly shift the point estimate to the null if were associated with both soy intake and breast cancer by risk ratios of 1.94.

Table 1.9.1: $\widehat{T}(r,q)$ and $\widehat{G}(r,q)$ (in parentheses) for varying r and q. Blank cells indicate combinations for which no bias would be required.

	q		
r	0.70	0.80	0.90
0.1	1.27 (1.85)	1.45 (2.25)	1.63 (2.64)
0.2	1.10 (1.44)	1.26 (1.84)	1.42 (2.19)
0.3		1.14 (1.55)	1.29 (1.89)
0.4		1.05 (1.28)	1.18 (1.64)
0.5			1.09 (1.41)

1.10 Simulation study

We assessed finite-sample performance of inference on $\widehat{p}(q)$ in a simple simulation study. While fixing the mean and variance of the true effects to $\mu^t = \log 1.4$ and $V^t = 0.15$ and the bias parameters to $\mu_{B^*} = \log 1.6$ and $\sigma^2_{B^*} = 0.01$, we varied the number of studies ($k \in \{15, 25, 50, 200\}$) and the average sample size N within each study ($E[N] \in \{300, 500, 1000\}$). The fixed parameters were chosen to minimize artifacts from discarding pathological samples with $\tau^2_c < \sigma^2_{B^*}$ or with truncated outcome probabilities due to extreme values of RR^c_{XY}; theoretically, $\widehat{p}(q)$ is unbiased regardless of these parameters. We set the threshold of scientific significance at $q = \log 1.4$ to match μ^t, such that, theoretically, 50% of true effects exceed q. We ran 1000 simulations for each possible combination of k and $E[N]$, primarily assessing coverage of nominal 95% confidence intervals and secondarily assessing their precision (total width) and bias in $\widehat{p}(q = \log 1.4)$ versus the theoretically expected 50%. Additionally, we assessed agreement between $\widehat{p}(q)$ and results obtained from an unconfounded meta-analysis (one in which all meta-analyzed studies adjust fully for confounding through stratification).

For each study, we drew $N \sim \text{Unif}\,(150, 2E[N] - 150)$, using 150 as a minimum sample size to prevent model convergence failures, and drew the study's true effect size as $M^t \sim N(\mu^t, V^t)$. We simulated data for each subject under a model with a binary exposure ($X \sim \text{Bern}(0.5)$), a single binary unmeasured confounder, and a binary outcome. We set the two bias components equal to one another ($g = RR_{XU} = RR_{UY}$) and fixed $P(U = 1|X = 1) = 1$, allowing closed-form computation of:

$$P(U = 1|X = 0) = \frac{\exp(M^t)[1 + (g - 1)] - \exp(M^c)}{(g - 1)\exp(M^c)}$$

as in [29]. Within each stratum $X = x$, we simulated $U \sim \text{Bern}\left(P(U = 1|X = x)\right)$. We simulated outcomes as $Y \sim \text{Bern}\left(\exp\{\log 0.05 + \log(g)U + M^t X\}\right)$. Finally, we computed effect sizes and fit the random-effects model using the metafor package in R [117], estimating τ_c^2 per [78] and $\widehat{\text{Var}}\left(\widehat{y_R^c}\right)$ with the [45] adjustment.

To compare results of our estimators to estimates from unconfounded meta-analyses, we also computed unconfounded effect sizes for each study using the Mantel-Haenszel risk ratio stratifying on U [86]. (This approach is used only for theoretical comparison, since in practice we are concerned with confounders that are unmeasured and therefore cannot be incorporated in analysis.) We then meta-analyzed these unconfounded point estimates and estimated, with no adjustment for bias, the proportion of effects in the population stronger than q.

Results (Table 1.10.1) indicated approximately nominal performance for all combinations of k and $E[N]$, with precision appearing to depend more strongly on k than $E[N]$. As expected theoretically, $\widehat{p}(q)$ was approximately unbiased. Compared to theoretical expectation, the proposed estimators appeared to perform slightly better than meta-analyses of unconfounded point estimates obtained through stratification on U. The latter method may have been compromised under strong confounding, which often induced

Table 1.10.1: For varying numbers of studies (k) and mean sample sizes within each study (Mean N), displays the estimated proportion ($\widehat{p}(q)$) of true effects above $RR = 1.4$ with its bias vs. theoretically expected 50% ($\widehat{p}(q)$ bias), coverage of 95% confidence intervals for $\widehat{p}(q)$ (CI coverage), and mean width of 95% confidence intervals (CI width). $\widehat{p}_{MH}$ is the estimated proportion of effects above $RR = 1.4$ in unconfounded analyses stratifying on U.

k	Mean N	$\widehat{p}(q)$	$\widehat{p}(q)$ bias	CI coverage	CI width	$\widehat{p}_{MH}$
15	300	0.530	0.030	0.970	0.575	0.585
25	300	0.533	0.033	0.965	0.459	0.582
50	300	0.527	0.027	0.975	0.316	0.572
200	300	0.528	0.028	0.917	0.154	0.568
15	500	0.523	0.023	0.981	0.522	0.558
25	500	0.527	0.027	0.982	0.409	0.561
50	500	0.522	0.022	0.973	0.283	0.554
200	500	0.523	0.023	0.945	0.140	0.553
15	1000	0.518	0.018	0.976	0.475	0.540
25	1000	0.516	0.016	0.983	0.370	0.537
50	1000	0.521	0.021	0.983	0.259	0.541
200	1000	0.515	0.015	0.971	0.129	0.536

zero cells in confounder-stratified analyses due to near collinearity of U with X and Y.

1.11 Discussion

This paper has developed sensitivity analyses for unmeasured confounding in a random-effects meta-analysis of a relative risk outcome measure. Specifically, we have presented estimators for the proportion, $\widehat{p}(q)$, of studies with true effect sizes surpassing a threshold and for the minimum bias, $\widehat{T}(r, q)$, or confounding association strength, $\widehat{G}(r, q)$, in all studies that would be required to reduce to a threshold the proportion of studies with effect sizes less than q. Such analyses quantify the amount of confounding bias in terms of intuitively tractable sensitivity parameters. Computation of $\widehat{p}(q)$ uses two sensitivity parameters, namely the mean and variance across studies of a joint bias factor on the log-relative risk scale. Estimators $\widehat{T}(r, q)$ and $\widehat{G}(r, q)$ make reference to, and provide conclusions for, a single sensitivity parameter, chosen as either the common joint bias factor across studies or the strength of confounding associations on the relative risk scale. These methods assume that the bias factor is normally distributed or fixed across studies, but do not make further assumptions regarding the nature of unmeasured confounding.

Assessing sensitivity to unmeasured confounding is particularly important in meta-analyses of observational studies, where a central goal is to assess the current quality of evidence and to inform future research directions. If a well-designed meta-analysis yields a low value of $\widehat{T}(r, q)$ or $\widehat{G}(r, q)$ and thus is relatively sensitive to unmeasured confounding, this indicates that future research on the topic should prioritize randomized trials or designs and data collection that reduce unmeasured confounding. On the other hand, individual studies measuring moderate effect sizes with relatively wide confidence intervals may not, when considered individually, appear highly robust to unmeasured confounding; however, a meta-analysis aggregating their results may nevertheless suggest that a substantial proportion of the true effects are above a threshold of scientific importance even in the presence of some unmeasured confounding. Thus, conclusions of the meta-analysis may in fact be robust to moderate degrees of unmeasured confounding.

We focused on relative risk outcomes because of their frequency in biomedical meta-analyses and their mathematical tractability, which allows closed-form solutions with the introduction of only one assumption (on the distribution of the bias factor). To allow application of the present methods, an odds ratio outcome can be approximated as a relative risk if the outcome is rare. If the outcome is not rare, the odds ratio can be approximately converted to a relative risk by taking its square root; provided that the outcome probabilities are between 0.2 and 0.8, this transformation is always within 25% of the true relative risk [112]. Comparable sensitivity analyses for other types of outcomes, such as mean differences for continuous outcome variables, would require study-level summary measures (for example, of within-group means and variances) and in some cases would yield closed-form solutions only at the price of more stringent assumptions. Under the assumption of an underlying binary outcome with high prevalence, such measures could be converted to log-odds ratios [10] and then to relative risks [112] as described above (see [111]). It is important to note that, in circumstances discussed elsewhere [104, 105], relative risk outcomes can produce biased meta-analytic estimates. When such biases in pooled point estimates or heterogeneity estimators are likely, sensitivity analyses will also be biased.

For existing meta-analyses that report estimates of the pooled effect, the heterogeneity, and their standard errors or confidence intervals, one could conduct the proposed sensitivity analyses using only these four summary measures (that is, simply using existing summary statistics and without re-analyzing study-level point estimates). However, in practice, we find that reporting of τ_c^2 and $\widehat{\text{Var}}\left(\tau_c^2\right)$ is sporadic in the biomedical literature. Besides their utility for conducting sensitivity analyses, we consider τ_c^2 and $\widehat{\text{Var}}\left(\tau_c^2\right)$ to be inherently valuable to the scientific interpretation of heterogeneous effects. We therefore recommend that they be reported routinely for random-effects meta-analyses, even when related measures, such as the proportion of total variance attributable to effect heterogeneity (I^2), are also reported. To enable sensitivity analyses of existing meta-analyses that do not report the needed summary measures, the package EValue helps automate the process of obtaining and drawing inferences from study-level data from a published forest plot or table. The user can then simply fit a random-effects model of choice to obtain the required summary measures.

Our framework assumes that the bias factor is normally distributed or taken to be fixed across studies. Normality is approximately justified if, for example, $\log RR_{XU}$ and $\log RR_{UY}$ are approximately identically and independently normal with relatively small variance. Since RR_{UY} is in fact a maximum over strata of X and the range of U, future work could potentially consider an extreme-value distribution for this component, but such a specification would appear to require a computational, rather than closed-form, approach. Perhaps a more useful, conservative approach to assessing sensitivity to bias that may be highly skewed is to report $\widehat{T}(r, q)$ and $\widehat{G}(r, q)$ for a wide range of fixed values B^*, including those much larger than a plausible mean.

An alternative sensitivity analysis approach would be to directly apply existing analytic bounds [29] to each individual study in order to compute the proportion of studies with effect sizes more extreme than q given a particular bias factor. This has the downside of requiring access to study-level summary measures (rather than pooled estimates). Moreover, the confidence interval of each study may be relatively wide, such that no individual study appears robust to unmeasured confounding, while nevertheless a meta-analytic estimate that takes into account the distribution of effects may in fact indicate that some of these effects are likely robust. One could also alternatively conduct sensitivity analyses on the pooled point estimate itself, but such an approach is naïve to heterogeneity: when the true effects are highly variable, a non-negligible proportion of large true effects may remain even with the introduction of enough bias to attenuate the pooled estimate to a scientifically unimportant level.

In summary, our results have shown that sensitivity analyses for unmeasured confounding in meta-analyses can be conducted easily by extending results for individual studies. These methods are straightforward to implement through either our R package EValue or website and ultimately help inform principled causal conclusions from meta-analyses.

1.12 Reproducibility

All code required to reproduce the applied example and simulation study is publicly available (https://osf.io/2r3gm/).

2

New Statistical Metrics for Multisite Replication Projects

2.1 Abstract

Increasing interest in replicability in the social sciences has engendered novel designs for replication projects in which multiple sites replicate an original study. At least 134 such “many-to-one” replications have been completed since 2014 or are currently ongoing. These designs have unique potential to help estimate whether the original study is statistically consistent with the replications and to re-assess the strength of evidence for the scientific effect of interest. However, existing statistical analyses generally focus on single replications; when applied to many-to-one designs, they provide an incomplete view of aggregate evidence and can lead to unduly pessimistic conclusions about replication success. We therefore propose new statistical metrics representing: (1) the probability that the original study's estimated effect size would be as extreme or more extreme than it actually was, if in fact the original study is statistically consistent with the replications; (2) the proportion of true effects agreeing in direction with the original study. Generalized versions of the second metric allow consideration only of true effects of non-negligible size; they estimate the proportion of true effects of scientifically meaningful size in the same direction as the estimate of the original study and, secondly, the proportion of effects of meaningful size in the direction opposite the original study's estimate. We provide an R package ("Replicate").

2.2 Introduction

Several social science disciplines have recently moved to empirically assess replicability of the published literature through systematic, third-party replications. Investigators conducting replications often seek to assess, firstly, how similar the results of the replication studies are to those of the original studies, that is, the extent to which the original studies are statistically consistent or inconsistent with their replications [3]. Second, investigators often aim to use replications to re-assess evidence strength for the scientific effect under investigation [3], ideally while minimizing bias (e.g., through protocol and analysis preregistration and a priori editorial approval [97]) and while ensuring high statistical power.

Novel designs for reproducibility research now exist to address these objectives with more sophistication than simple designs involving a single replication of an original study. Some high-impact experimental psychology journals now encourage projects in which multiple independent sites attempt to replicate a single, published original study using a standardized experimental protocol closely approximating the original and developed with input from the original authors [97]. Extensions (sometimes called “Many Labs” projects) select multiple original studies and subject each to a multisite replication [31, 57], and others have applied a similar approach to replicate new original research prior to its publication [89]. We use the term “many-to-one design” to refer generically to any design in which an original study is replicated in multiple sites. Many-to-one replication research is a nascent, but rapidly expanding, field: we are aware of at least 79 completed and 55 ongoing many-to-one replication studies to date, all completed or initiated since 2014 and in experimental psychology and experimental philosophy alone (completed: [2, 11, 16, 23, 31, 33, 42, 57, 89, 118]; ongoing: [5, 32, 58, 88]).

However, the adoption of many-to-one designs in the social sciences has outpaced development of corresponding statistical analyses. Existing work [4, 36, 77, 98, 115] has proposed analytic approaches for a single replication of a single study or designs in which numerous original studies across a discipline or domain are each replicated once (here termed “one-to-one designs”), as in [76] and in [14]. However, many-to-one designs pose unique statistical challenges and opportunities. Results of many-to-one replications often suggest effect heterogeneity across sites despite use of standardized protocols (for example, 8 of 16 replications in [57] suggested “statistically significant” evidence of heterogeneity), yet current analysis approaches do not adequately account for heterogeneity. As we will discuss, this can lead to unduly pessimistic assessments of consistency between the original study and the replications and to misleading re-assessments of the strength of evidence for the effect under investigation. Additionally, results of many-to-one designs often lead to unresolved debates regarding the extent to which the original study "replicated" or "did not replicate", but these debates remain highly speculative, perhaps partly because few directly relevant quantitative metrics are currently available.

We therefore propose new statistical metrics specifically designed for many-to-one designs. To assess statistical consistency, we provide a metric (P_{orig}) representing the probability that the effect estimate from the original study would be as extreme or more extreme than it actually was if, in fact, the original study and the replications were statistically consistent in the sense of being drawn from the same distribution. To assess evidence strength, we provide a metric estimating the proportion of true effects agreeing in direction with the original effect estimate ($P_{>0}$). Because replication effects that agree in direction with the original, but are very weak, may in fact be considered insufficient evidence to support the original effect, we also demonstrate how to generalize this metric to consider the proportion of true effects that not only agree in direction with the original, but are also stronger than a user-chosen threshold of scientifically meaningful size ($P_{>q}$). Lastly, we also provide a counterpart metric estimating the proportion of true effects in the opposite direction of the original ($P_{<q^*}$). In contrast to existing metrics, the proposed metrics account for all relevant sources of statistical uncertainty in many-to-one replication designs, including heterogeneity [53], and they harness the specific strengths of many-to-one designs. These metrics are mathematically very straightforward, but to the best of our knowledge, have not yet been reported in any published many-to-one replication. We provide an R package, "Replicate", to conduct all proposed analyses.

2.3 Applied example

As a running example, we will consider one of several many-to-one replication attempts conducted by [31]. Specifically, each of 21 independent labs used a common protocol to replicate a classic psychology experiment (Experiment 1 of [69]) on “moral credentialing” theory, which proposes that people given an initial opportunity to demonstrate that they are not prejudiced (and thus establish “moral credentials”) are more likely to display apparently prejudiced attitudes in subsequent tasks (having licensed themselves to do so because of their previously established credentials). In the replicated experiment, the initial task required subjects to agree or disagree with potentially sexist statements. In the initial task, subjects were randomized to a credentialing condition in which the statements described “most women” (e.g., “Most women need a man to protect them”) or to a control condition, in which the same statements described only “some women”. Thus, credentialing statements were designed to induce higher disagreement than control statements, allowing subjects in the former condition to more clearly establish themselves as non-sexists. The dependent variable was subjects’ degree of preference for male candidates in an imagined hiring scenario. As predicted, subjects in the credentialing condition more strongly preferred to hire male candidates than did control subjects (corresponding to an effect size of $r = 0.21$ on Pearson’s correlation scale, $p = 7 \times 10^{-4}$, 95% CI: 0.09, 0.32). [69] also reported an unexpected interaction of credentialing condition with the subject’s sex, and [31] (2016) attempted to replicate both the main effect and the interaction. For brevity, we focus only on the main effect.

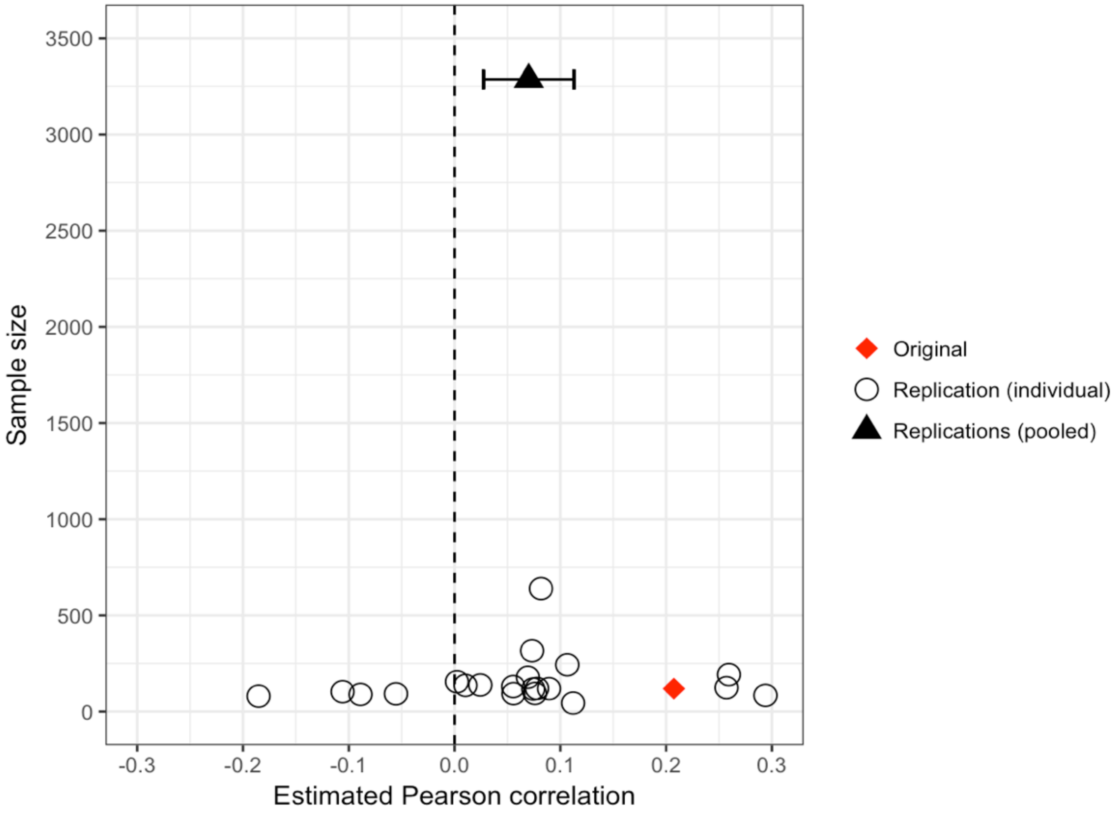

Figure 2.3.1: Estimated correlation in [69]'s original study, in each of [31]'s replications, and in a meta-analytic pooled estimate across the replications.

2.4 Existing metrics

We first review metrics commonly reported in many-to-one designs as well as those developed for other designs, but that are frequently reported in many-to-one designs. First, nearly all many-to-one designs report a pooled estimate of the effect size in the replications. The pooled estimate is usually estimated by meta-analyzing effect sizes from the replications or by fitting a mixed model to individual subject data. For example, fitting a random-effects meta-analysis model to [31]'s replication studies on moral credentialing estimates an average effect size of 0.07 (95% CI: 0.03, 0.11) on the Pearson's correlation scale; both the replicators and the lead author of the original study [68] interpreted this finding as a successful replication supporting moral credentialing. Regardless of modeling approach, this metric estimates the average true effect size across the replications. This is adequate if replications exhibit little heterogeneity but provides an incomplete picture in the presence of heterogeneity across replication studies. Such heterogeneity may occur, for example, if replication studies differ with respect to subjects' demographic characteristics (e.g., age, sex, race, or geographic region) or the setting in which the study is conducted (e.g., time of day, physical setting, etc.). As the proposed metrics will formalize, Figure 2.3.1 suggests heuristically that although a group of replication point estimates were clustered around the pooled point estimate, several point estimates were in fact in the direction opposite the original, and several were even larger than the original.

As discussed elsewhere in the context of meta-analyses rather than replications [65], under moderate or substantial heterogeneity, a pooled estimate near the null can belie the existence of strong effects in some replication settings. Thus, due to heterogeneity, a many-to-one replication design whose pooled estimate appears not to support the hypothesized effect may nevertheless provide evidence of meaningfully large effects in favor of the original hypothesis in some contexts (for example, locations, subject demographics, variations in protocol administration, etc.). Conversely, if the pooled estimate is in the same direction as the original estimate, but is smaller, we cannot directly discern whether the true effect is never as large as originally reported (and perhaps is too small to warrant scientific interest) or whether it may, in fact, be as large as or larger than the original estimate in some settings. For these reasons, we will recommend supplementing the pooled point estimate with new metrics that additionally characterize heterogeneity.

A widespread metric of statistical consistency assesses whether the replication study obtains a "statistically significant" p-value and an effect estimate in the same direction as in the original study (assuming that the original study itself obtained a "significant" p-value). This "significance agreement" metric is widely reported in single replications [3], in one-to-one designs [14, 76], and in many-to-one designs. However, as others have noted [77, 98], "significance agreement" is challenging to interpret because it is a function not only of the nominal α-level (e.g., 0.05), but also of power in both the original and the replication study. Thus, the expected probability of "significance agreement" may be quite low [4, 77], though it can be simulated [77] or derived (Appendix) for a given original and replication study and then compared to the observed probability. In our running example regarding moral credentialing, 24% of replications (5 of 21) obtained results agreeing in "statistical significance" and effect direction with the original, which appears much lower than the 62% that we would expect theoretically (based on the original effect size and its standard error, as well as the standard error of each replication).

A variety of more interpretable metrics have been developed for one-to-one replications, and some have also been reported in many-to-one designs. [77] proposed using the original study to construct a prediction interval representing a plausible range for the effect estimate in the replication study, assuming that the replication and the original study are generated from the same distribution (i.e., they are statistically "consistent"). If indeed the two studies are generated from the same distribution, then regardless of power in either study, there is, by construction, a 95% probability that the replication effect estimate will fall inside the prediction interval. [98] proposed a hypothesis test of the replication estimate versus a nonzero null value chosen as the smallest effect size that the original study would have had an estimated 33% power to detect. [4] developed a sophisticated, general statistical model for median-unbiased effect size estimation in one-to-one replication designs such as [76]. Several authors (e.g., [36, 115]) recommend using Bayes factors to quantify evidence for and against the null hypothesis.

In a many-to-one design, some of these metrics can be applied individually to each replication study or to the pooled estimate. The former analysis can be informative, but does not aggregate evidence and statistical power across all replications. The latter analysis is subject to the same limitations as the pooled estimate itself, namely that it summarizes a potentially heterogeneous distribution of replication effects by only its mean. In fact, as we illustrate below, analyses that fail to account for heterogeneity can underestimate consistency when there is in fact heterogeneity, leading to conclusions that are unduly unfavorable to the original study (see Appendix for proof).

2.5 Proposed new analyses

As discussed above, few statistical methods have been developed specifically for many-to-one designs, and those that were developed for other replication designs have limitations when applied to many-to-one designs, particularly in the presence of heterogeneous effects. We therefore propose new metrics to address central objectives of replication research while accounting for all relevant sources of statistical uncertainty, namely statistical error in the original, statistical error in the replication, and heterogeneity. All proposed analyses are easy to compute manually or using the R package Replicate, whose capabilities are summarized in the Appendix [66].

Consistency of original with replications (P_{orig})

Our first proposed metric assesses statistical consistency. Rather than assuming that the replications and the original measure exactly the same underlying effect size – an assumption implicit to most metrics for single replications – we instead assume that they measure potentially heterogeneous, normally distributed effects. We will then say that the original study is "consistent" with the replications if it is generated from the same underlying distribution as the replications; that is, its true effect size comes from the same distribution as those of the replications. Then, we define the first proposed metric, called P_{orig}, as the probability that, if indeed the original is consistent with the replications in this sense, its estimate would be as extreme or more extreme than it actually was. A small value of P_{orig} would indicate strong evidence that the original study is inconsistent with the replications, whereas a large value would suggest relatively good consistency. In practice, if the original study is highly inconsistent with the replications, even accounting for heterogeneity, then we might consider it an anomaly. Future meta-analyses of the published literature might then present analyses both including and excluding such potentially anomalous studies. Additionally, others describe meta-analytically pooling results of an original study with those of a replication [3, 76]; high inconsistency would suggest interpreting such analyses with greater caution.

To estimate P_{orig}, we first define (as before) $\widehat{\theta}_{orig}$ and $\widehat{SE}_{orig}$ as the original effect estimate and its standard error, $\widehat{\mu}$ and $\widehat{SE}\left(\widehat{\mu}\right)$ as an estimate of the average true effect size in the replications and its standard error, and $\widehat{V}$ as an estimate of the variance of the true effect sizes across replications. The effect sizes should be estimated on a scale for which the normality assumption is plausible. In practice, $\widehat{\mu}$ and $\widehat{V}$ are most commonly estimated using the pooled estimate and heterogeneity estimate, often denoted τ^2, from a random-effects meta-analysis of the replication sites' estimates. Alternatively, they could be estimated by fitting a mixed model to the individual observations themselves (also known as an "individual patient data meta-analysis" [101]); both approaches are further discussed in the Appendix. In the main text, for simplicity, we illustrate the common meta-analytic approach, but all analyses can be conducted using any unbiased estimates $\widehat{\mu}$ and $\widehat{V}$ arising from a model with the given distributional assumptions (Appendix).

Then, if the original study is in fact consistent with the replications, the probability that its estimate would be as extreme as we observe it to be is approximately:

$$P_{orig} = 2 \times \left(1 - \Phi\left(\frac{|\widehat{\theta}_{orig} - \widehat{\mu}|}{\sqrt{\widehat{V} + \widehat{SE}^2_{orig} + \widehat{SE}^2\left(\widehat{\mu}\right)}}\right)\right) \tag{2.1}$$

For example, we fit a random-effects meta-analysis to [31]'s site-level data to estimate (on the Fisher's z scale) $\widehat{\mu} = 0.07$, $\widehat{SE}\left(\widehat{\mu}\right) =$

0.02, and $\widehat{V} = 2.7 \times 10^{-3}$. We computed $\widehat{\theta}_{orig} = 0.21$ and $\widehat{\text{SE}}_{orig} = 0.06$ for the original study by converting the reported η^2 scale to Fisher's z [60]. Then, we applied Equation 2.1 to compute that if the true effect in the original study indeed arose from the same estimated distribution as those in the replications, there would be a 10% chance that the original effect estimate would be as extreme or more extreme than the observed 0.21. We can interpret this fairly low, but nonnegligible, probability as being only weakly suggestive of inconsistency.

In contrast, previously discussed metrics indicating a low proportion of replications agreeing in "statistical significance" (24% versus 62% expected) and falling within the original prediction interval (76% versus 95% expected) might appear to more strongly suggest inconsistency. These relatively more pessimistic conclusions (compared to the conclusions we might draw from P_{orig}) reflects their failure to account for heterogeneity in the effects across replications. To illustrate quantitatively, we can re-compute P_{orig}, but this time setting $\widehat{V} = 0$ to assume no heterogeneity in the effects across replications. We then obtain a probability of only 3%. This is considerably lower than the 10% obtained by properly accounting for heterogeneity: a heterogeneous distribution of effects in the replications allows a higher chance that any given study would measure a very large or very small effect size (as shown mathematically in the Appendix).

Proportion of true effects agreeing in direction with the original ($P_{>0}$)

To address a second central objective of replication – re-assessing evidence strength for the scientific effect of interest – we propose a metric ($P_{>0}$) to supplement the usual pooled effect estimate and its confidence interval. Unlike these existing metrics, which characterize only the mean of the distribution of true effects in the replications, $P_{>0}$ characterizes both the mean and the heterogeneity of this distribution, and it addresses effect size rather than "statistical significance". Specifically, $P_{>0}$ represents the proportion of true effects, among the potentially heterogeneous population from which the replications are a sample, that agree in direction with the original. That is, any nonzero true effect agreeing in direction can be interpreted as a "real" effect supporting the original study's theory (albeit potentially of a smaller effect size).

To estimate $P_{>0}$, it is not sufficient to simply compute the observed proportion of replication estimates agreeing in direction with the original; such an approach would fail to account for statistical error in the replication estimates. That is, the challenge is to use the distribution of the replication estimates (which has variability due to both heterogeneity and statistical error) to estimate the distribution of true effects (which has variability due only to heterogeneity). Thus, we can estimate the proportion of true effects above 0 as:

$$P_{>0} = \Phi\left(\frac{\widehat{\mu}}{\sqrt{\widehat{V}}}\right) \tag{2.2}$$

where Φ denotes the standard normal cumulative distribution function.

(We assume for simplicity that the original effect estimate is positive, such that Equation 2.2 represents effect sizes in the same direction as the original study. If instead the original effect estimate is negative, simply use the subsequent Equation 2.4 with $q^* = 0$ to assess effect sizes agreeing in direction with the original estimate. Additionally, we assume the null hypothesis $\mu = 0$; for other null hypotheses, use Equation 2.3 with q set to the null value.) When there are approximately 10 or more replications, $P_{>0}$ is approximately normal with estimated standard error:

$$\widehat{SE} = \sqrt{\frac{\widehat{SE}^2\left(\widehat{\mu}\right)}{\widehat{V}} + \frac{\widehat{SE}^2\left(\widehat{V}\right)\widehat{\mu}^2}{4\widehat{V}^3}} \cdot \varphi\left(\frac{\widehat{\mu}}{\sqrt{\widehat{V}}}\right)$$

where φ denotes the standard normal density function and $\widehat{SE}\left(\widehat{V}\right)$ the estimated standard error of the heterogeneity estimate. Thus, an approximate 95% CI is $P_{>0} \pm 1.96 \times \widehat{SE}$. (This expression applies for estimators $\widehat{V}$ that are asymptotically normal and independent of $\widehat{\mu}$, which holds for many common choices (Mathur and VanderWeele, 2017b, Appendix).)

Proportion of true effects of scientifically meaningful size ($P_{>q}$ and $P_{<q^*}$)

The aforementioned $P_{>0}$ treats all effects that agree in direction with the original estimate, even those that are very close to the null, as evidence in favor of the scientific effect under investigation. This is generous toward the original study, and therefore might serve as a useful default analysis. Alternatively, as a more stringent measure of evidence strength, it can also be useful to consider a generalized metric ($P_{>q}$) representing the proportion of effects stronger than a non-null threshold, q. This approach is similar to equivalence testing and minimal effects testing, which compare a point estimate to null values other than 0 [61]. An extensive interdisciplinary literature has provided recommendations on how to choose thresholds for scientifically meaningful effect sizes, which we summarize briefly in the Appendix. For example, suppose that through comparison to well-established effects on similar dependent variables (Appendix)}, one selects a threshold at a effect size of Cohen's $d = 0.20$, or equivalently, an approximate correlation of $r = 0.10$ [20]. If $P_{>q}$ is large (e.g., 85%), this suggests that, when drawing from the population distribution of effect sizes underlying the replications, a high proportion of true effects are large enough to warrant scientific interest (e.g., larger than Cohen's $d = 0.20$). We might therefore conclude that the replications provide strong evidence that the scientific effect of interest is meaningfully strong in many settings. In contrast, if $P_{>q}$ is small, we might instead conclude that the replications fail to support scientifically meaningful effects in most contexts.

Conversely, it can also be useful to consider effects in the direction opposite the original estimate using a second threshold-based metric, $P_{<q^*}$. That is, one could select a second threshold representing a scientifically meaningful effect size in the opposite direction (e.g., Cohen's $d = -0.20$) and estimating the proportion of true effects below this threshold. If the pooled estimate is fairly close to the null or if heterogeneity is substantial, this probability may be nonneglible, suggesting that the experimental manipulation may (perhaps unexpectedly) induce meaningful effects in the opposite direction in some replication settings. Such a finding may help stimulate hypotheses regarding important moderators or boundary conditions on the effect of interest. Additionally, effects in the opposite direction from theoretical predictions may actively support competing theories. Indeed, when evaluating competing theories, researchers sometimes deliberately design experimental manipulations that are expected to induce opposing effects under each candidate theory. Returning to moral credentialing, the theory under investigation predicts that credentialing opportunities would increase subsequent attitudes consistent with prejudice; however, other theories suggest that credentialing opportunities might sometimes decrease such attitudes by prompting self-consistency or by priming personal values that discourage prejudice [72]. Using $P_{<q^*}$ to explicitly characterize effects in the opposite direction (rather than simply allowing them to dilute the pooled estimate without additional consideration) may help

identify situations, possibly supported by alternative theories, in which such competing effects occur.

These threshold-based metrics are particularly informative when the pooled estimate in the replications is smaller than that of the original study, as is often the case (e.g., [31]). The proportion of true effects above a threshold ($P_{>q}$) may then help identify whether: (1) the true effects are closely clustered around a small average effect size, providing little evidence for effects of scientifically meaningful magnitude; versus (2) the true effects are quite variable around a small average effect size, such that there is in fact compelling evidence that effects of scientifically meaningful magnitude occur in some settings (and thus suggesting the importance of examining possible moderators). For example, suppose the original study estimates an effect size of $d = 0.85$, but the replications estimate a much smaller pooled effect size of $d = 0.40$. Exclusive focus on the existing metrics may then mislead us into considering the replication effort to have succeeded completely (if the pooled point estimate is also "statistically significant") or to have failed completely (if the pooled point estimate is not "statistically significant"). However, if we additionally choose a threshold of scientific importance at, for example, $d = 0.20$ and estimate a reasonably high percent (e.g., 25%) of true effects exceeding this threshold, then we might instead consider the replications to provide moderately strong evidence for meaningful effect sizes in some replication settings, warranting an assessment of possible moderators. In contrast, if we instead find that only, for example, 8% of true effects exceed $d = 0.20$, then we might instead conclude that the replications provide little evidence to support scientifically meaningful effect sizes (even if the pooled point estimate is "statistically significant").

To estimate $P_{>q}$ and $P_{<q^*}$, first let q be a chosen effect size threshold of scientific importance. Then we can estimate the proportion of true

effects above q as:

$$P_{>q} = 1 - \Phi\left(\frac{q - \widehat{\mu}}{\sqrt{\widehat{V}}}\right) \tag{2.3}$$

For the second metric, we can estimate the proportion of true effects below a second threshold, q^*, (e.g., Cohen's $d = -0.20$) as:

$$P_{<q^*} = \Phi\left(\frac{q^* - \widehat{\mu}}{\sqrt{\widehat{V}}}\right) \tag{2.4}$$

(Again, we assume that the original effect estimate is above the null; if instead the original effect estimate is below the null, simply reverse the two equations, using Equation 2.4 to assess effect sizes supporting the original theory and Equation 2.3 to represent those in the opposite direction.) Both $P_{>q}$ and $P_{<q^*}$ have approximate standard error:

$$\widehat{SE} = \sqrt{\frac{\widehat{SE}^2\left(\widehat{\mu}\right)}{\widehat{V}} + \frac{\widehat{SE}^2\left(\widehat{V}\right)\left(q - \widehat{\mu}\right)^2}{4\widehat{V}^3}} \cdot \varphi\left(\frac{q - \widehat{\mu}}{\sqrt{\widehat{V}}}\right)$$

where q^* can simply be substituted for q when considering $P_{<q^*}$.

Proportion of replication effects supporting moral credentialing

In the moral credentialing example, the original study estimated an effect size of 0.21 (95% CI: 0.09, 0.32) on the Pearson's correlation scale, whereas our meta-analysis of [31]'s replications estimates a pooled effect size of 0.07 (95% CI: 0.03, 0.11) with estimated heterogeneity $\widehat{V} = 2.7 \times 10^{-3}$. As discussed previously, the "statistically significant" result in the replications might appear to suggest that the replication effort was successful. But does the small pooled estimate in the replications, despite its “statistical significance”, correspond to a high proportion of replication effects supporting credentialing theory? First, we can use $P_{>0}$ to estimate the proportion of true effects above 0 (91% with 95% CI: 64%, 100%). Alternatively, suppose we select a threshold of $r = 0.20$ as a minimum effect size of scientific importance, which is similar to the original estimate and is more conservative than well-established effects of experimentally-induced intergroup contact on prejudice (see Appendix). Then we can estimate via $P_{>q}$ that almost no effects (1% with 95% CI: 0%, 5%) surpass $r = 0.20$. If we select a less stringent effect size threshold of $r = 0.10$ (approximately equal to Cohen’s $d = 0.20$), we would estimate that approximately 28% (95% CI: 0%, 63%) of true effects surpass $r = 0.10$. We can then also estimate $P_{<q^*}$, that is, the proportion of true effects of scientifically meaningful magnitude in the direction opposite [69]’s original findings. We might, for example, choose a conservative second threshold at, for example, $r = -0.10$ and use Equation 2.4 to estimate that almost no inverse effects (0% with 95% CI: 0%, 1%) are more negative than this threshold.

Ultimately, although these replications produce a “statistically significant” point estimate in the same direction as the original study’s estimate, we might nevertheless caution that they provide little evidence for effect sizes of comparable strength to the original estimate across replication settings. In the distribution of true effects, there is a high proportion of a nonzero effects in the direction of the original estimate, but most of these effects are considerably smaller than the original estimate. Considering these results along with the previously discussed consistency metric (P_{orig} = 10%), we might say, overall, that the moral credentialing main effect “replicated” in the sense that there is not compelling evidence for inconsistency between the original study and replications (once we account for heterogeneity), yet evidence strength for scientifically meaningful effect sizes of moral credentialing is considerably weaker than suggested by the original study. These complementary findings further illustrate the conceptual distinction between statistical consistency and evidence strength for scientifically meaningful effects of interest.

2.6 Statistical assumptions

Our proposed metrics assume that the true effect sizes in the replication studies are normally distributed. In most many-to-one designs, which use mixed modelling or parametric random effects meta-analysis to estimate the pooled effect, this assumption is already implicit. Nevertheless, investigators should assess whether the normality assumption is plausible by checking for approximate normality of the replication estimates. (Although the replication estimates are not themselves true effects, normal true effects would typically produce approximately normal replication estimates, and nonnormal true effects would produce nonnormal replication estimates.) To allow assessment of normality and accurate heterogeneity estimation, these metrics should generally be applied only when there are at least 10 replication studies (which, to the best of our knowledge, was true in each of the 79 completed many-to-one designs discussed in the Introduction). An exception to this rule of thumb is when there is no heterogeneity, as discussed below.

2.7 Applications to other replication designs

We have primarily discussed our metrics in the context of many-to-one designs conducted under a shared replication protocol and in which true effects are heterogeneous. Here, we discuss other designs and settings to which the proposed metrics apply without modification.

2.7.1 Replications without heterogeneity

In many-to-one designs yielding a negligible statistical estimate of heterogeneity, in one-to-one replication designs, or in a single replication of a single original study, P_{orig} can still be informative to assess consistency. Without heterogeneity, P_{orig} does not require a normality assumption and can be reported with as few as one replication study, and it becomes a continuous counterpart to a prediction interval in which all replication data are analyzed in aggregate, without regard to site (Appendix). When there is no heterogeneity or when it is not estimable (in single replications or one-to-one replications), $P_{>0}$, $P_{>q}$, and $P_{<q^*}$ are no longer relevant because all true effects are taken to be identical.

2.7.2 "Many Labs" designs

In designs in which multiple original studies are each replicated in many sites (e.g., [31, 57, 89]), the proposed metrics permit direct comparison or aggregation of results across many-to-one replications of multiple original studies. For example, one could estimate the proposed metrics for each original study and report the average consistency (P_{orig}) as a global summary measure of replication success. The average $P_{>0}$ could also be reported as a global summary of replication evidence strength across numerous scientific effects.

2.7.3 Conceptual replications

We have so far considered contexts in which all replications share a single protocol closely approximating that of the original study (sometimes called "direct replications"). However, some researchers question using only direct replications in many-to-one designs, arguing that these designs assess replicability of a specific operationalization of a theory, rather than of the theory itself [7].

Others advocate supplementing direct replications with "conceptual replications" that assess the same theory as the original study, but using a different operationalization ([24, 64, 70]; see also dissent by [96] and [74]). For example, replication sites in a conceptual many-to-one design could implement different experimental protocols, each approved by the original authors. Conceptual replications create heterogeneity by design, which exacerbates problems with the metrics proposed prior to this paper (e.g., leading to particularly unfavorable assessments of consistency and inadequately characterizing evidence strength). In contrast, our proposed metrics could simply be applied without modification as they take into account heterogeneity across replications. They would retain their original interpretations, but $P_{>0}$ could then additionally be interpreted as the probability that a new operationalization of the theory at stake would yield a true effect size either in the same direction as the theoretical prediction. Such an interpretation holds only when the new operationalization under consideration can be treated as comparable to the range of protocols considered in the conceptual replications.

2.8 Conclusion

We have proposed intuitively tractable metrics (implemented in the R package "Replicate") for statistical consistency between the original study and replications and for evidence strength in many-to-one replication designs with potential heterogeneity. Such replication projects could report the new metric P_{orig} to convey consistency and could report the usual pooled estimate ($\widehat{\mu}$), heterogeneity estimate ($\widehat{V}$), plus $P_{>0}$ (and possibly also $P_{>q}$ and $P_{<q^*}$) to re-assess evidence strength for the scientific effect of interest. The proposed metrics account for all relevant sources of statistical uncertainty and can therefore yield different conclusions from existing metrics when the replications are

heterogeneous. These metrics can also help identify situations in which there is good statistical consistency, but weak evidence strength for scientifically meaningful effects (and vice versa). For example, a set of replications estimating a small average effect size might be statistically consistent with a low-powered original study that estimated a large effect size, yet may provide little evidence that the effects of interest are of scientifically meaningful size. In this case, P_{orig} would be fairly large, indicating good consistency, but $P_{>q}$ would be small, indicating a low proportion of scientifically meaningful effect sizes. Conversely, a set of replications estimating a moderate effect size may appear statistically inconsistent with an original study estimating a large effect size, but may nevertheless provide strong evidence for scientifically meaningful effect sizes.

The proposed analyses have limitations. As discussed, they assume the true effects are normally distributed; this assumption is already often used in pooled effect estimation and is often testable in practice. The metrics also rely on accurate statistical estimation of both the pooled effect size and its variance. When estimating these parameters via random-effects meta-analysis, there are many possible choices of heterogeneity estimator, and it is important to choose one that is known to perform well for the effect measure of choice, particularly when the number of replication studies is relatively small [116]. Additionally, we do not recommend using P_{orig} to conduct a dichotomous "hypothesis test" of consistency (by assessing whether $P_{orig} < 0.05$) between the original study and the replications; rather, P_{orig} is a continuous measure and is more informative when reported as such. Finally, we have assumed that the replications unbiasedly estimate true effects, which is often reasonable when the replications are preregistered and conducted by third-party investigators. In contrast, other forms of replications, such as multiple experiments reported in a published, non-registered paper may be subject to the same biases seen in published original research [38]. These metrics are mathematically very simple but are nevertheless, we believe, a useful supplement to current reporting practices to help quantitatively ground speculation about “replication success”.

In summary, the newly proposed metrics assess consistency of the original and replication studies and also assess evidence for effects of scientifically meaningful size while accounting for heterogeneity across the true effects. Such heterogeneity is fairly common in practice and can arise due to differing subject demographics or protocol variations. If reported in many-to-one replication projects, the proposed metrics could help directly and intuitively address the central objectives of replication research.

2.9 Reproducibility

All code required to reproduce the applied examples is publicly available in an RMarkdown preparation of this manuscript (https://osf.io/apnjk/).

3

New Metrics for Multiple Testing with Correlated Outcomes

3.1 Abstract

We propose new metrics comparing the observed number of hypothesis test rejections $(\widehat{\theta})$ at an unpenalized α-level to the distribution of rejections that would be expected if all tested null hypotheses held (the "global null"). Specifically, we propose reporting a "null interval" for the number of α-level rejections expected to occur in 95% of samples under the global null, the difference between $\widehat{\theta}$ and the upper limit of the null interval (the "excess hits"), and a one-sided joint test based on $\widehat{\theta}$ of the global null. For estimation, we describe resampling algorithms that asymptotically recover the sampling distribution under the global null. These methods accommodate arbitrarily correlated test statistics and do not require high-dimensional analyses. In a simulation study, we assess properties of the proposed metrics under varying correlation structures as well as the relative power of global tests constructed using existing FWER methods. We provide an R package, NRejections. Ultimately, existing procedures for multiple hypothesis testing typically penalize inference in each test, which is useful to temper interpretation of individual findings; yet on their own, these procedures do not fully characterize global evidence strength across the multiple tests. Our new metrics help remedy this limitation.

3.2 Introduction

In studies testing multiple hypotheses, the problem of inflated Type I error rates is usually handled, if at all, through procedures that preserve familywise error rate (FWER) or false discovery rate (FDR) by penalizing individual p-values or critical values. These procedures can be valuable for individually correcting inference for each hypothesis test. However, as standalone reporting methods, they may provide incomplete insight into the overall strength of evidence across tests. For example, if individual hypothesis tests of the associations between a single exposure of interest and 40 outcome measures result in a total of 10 rejections at an uncorrected $\alpha = 0.05$ and result in 1 rejection at a Bonferroni-corrected $\alpha \approx 0.001$, how strong is the overall evidence supporting associations between the exposure and the outcomes, considered jointly? Given only the information typically reported in corrected or uncorrected multiple tests, such questions can be hard to answer.

Intuitive speculation about overall evidence strength becomes especially challenging when the hypothesis tests are correlated, which is typically the case when related research questions are considered [9] or in "outcome-wide" analyses that assess associations between a single exposure and a number of outcomes [113]. Indeed, as we will illustrate, the results of a given set of individual tests (whether multiplicity-corrected or not) may be strongly suggestive of at least some genuine effects if the tests are independent, but may be entirely consistent with chance (in a manner we will formalize) if the tests are correlated. In practice, the correlation structure of the tests is usually unknown, further impeding intuitive assessment.

We therefore aim to supplement existing multiple-testing procedures (e.g., [30, 48, 83, 85, 121]) with simple metrics that directly characterize overall evidence strength while accommodating arbitrarily correlated test statistics. These metrics focus on the total number of hypothesis test rejections at an arbitrary α level (such as the usual, uncorrected $\alpha = 0.05$). First, we propose reporting a null interval representing a plausible range for the total number of rejections in 95% of samples that would occur if all null hypotheses were true (a scenario we call the "global null"), along with the difference between the number of observed rejections and the upper interval limit (the excess hits). For example, if we reject 10 of 40 hypotheses at $\alpha = 0.05$, we might be tempted to conclude intuitively that this is "more" than the expected $0.05 \times 40 = 2$ rejections. However, if the null interval is $[0, 11]$, accounting for correlation between the tests, this would suggest little evidence overall for genuine associations across the 40 tests. In contrast, if we instead reject 14 tests under the same correlation structure, the null interval indicates that we have observed 3 excess hits beyond the number that would be expected in 95% of samples generated under the global null, which is suggestive of strong overall evidence that at least some of the tested associations are present. Additionally, we propose using the number of rejections to conduct a one-sided global test of the global null, whose p-value represents the probability, in samples generated under the global null, of observing at least as many α-level rejections as were actually observed.

Although standard methods for FWER control are not explicitly designed to characterize overall evidence strength, they could in principle be repurposed into a global test. That is, rejection of at least one test with inference corrected to preserve a familywise $\alpha = 0.05$ implies rejection of the global null at $\alpha = 0.05$. Several existing methods strongly control FWER in hypothesis tests with unknown correlation structure and could therefore be suitable for a global test. The most widespread approaches are the classical one-step Bonferroni correction [30] and its uniformly more powerful successor, the step-down Holm method [48], both of which can be computed in closed form. By avoiding specifying or estimating the correlation structure, these naïve methods accommodate even worst-case correlation structures but can yield conservative inference. Other closed-form methods achieve better power by assuming independence (e.g., various procedures based on [94]'s inequality) or known logical dependencies between tests (e.g., [90]) but can produce anticonservative inference if these assumptions are violated [91]. We focus here on modern methods, detailed in Section 3.3, that avoid such assumptions by empirically estimating the correlation structure via resampling [83, 85, 121]. Related methods control FDR rather than FWER (e.g., [85, 109, 110]), but because FDR control does not appear to have a direct relationship with the types of global test or null interval discussed in the present paper, we do not further consider these methods. Alternative approaches are designed for a large number of hypothesis tests, as in high-dimensional genetic studies (e.g., [40, 62, 102]); however, because correlated hypothesis tests can be particularly problematic in traditional low-dimensional settings [19], we aim to provide methods that apply regardless of the number of tests.

In this paper, we first derive assumptions for the asymptotic validity of a resampling-based null interval, the corresponding excess hits, and a global test of the number of rejections, and we describe specific resampling algorithms fulfilling these assumptions. Second, we conduct a simulation study in which we (1) compare the null interval to the observed number of rejections for varying effect sizes; and (2) assess the relative power of global tests conducted using the number of rejections or derived from existing FWER-control methods, as discussed above. To our knowledge, prior simulation studies of existing FWER methods have not reported on their performance as global tests [85]. We illustrate application of our proposed metrics through an applied example.

3.3 Existing resampling-based multiplicity corrections

Table 3.3.1 summarizes existing methods that strongly control FWER for arbitrarily correlated tests. [121] proposes resampling algorithms that resemble the standard bootstrap, but that modify the data in order to enforce the global null (an approach similar to what we will adopt). For example, suppose we conduct one-sample t-tests on the potentially correlated variables $(Y_1, \cdots, Y_W)$ with the global null stating that $E[Y_w] = 0 \; \forall \; w \in \{1, \cdots, W\}$. To resample under the global null, the observed data could first be centered by their respective sample means, then resampled with replacement [121]. Thus, in the resampled datasets, the global null holds regardless of the true parameters $(E[Y_1], \cdots, E[Y_W])$ underlying the original sample. Then, [121]'s one-step "minP" method and uniformly more powerful step-down variant (here termed "Wstep") adjust the observed p-values using quantiles of the distribution of p-values calculated in the resamples.

Other FWER methods use parametric resampling approaches that do not enforce the global null in the resampled data, but rather that generate datasets resembling the original data [83, 85]. Essentially invoking the duality of hypothesis tests and confidence intervals, the resampled test statistics are then centered by their estimated values in the observed data in order to recover the null distribution; other related methods showed less favorable performance in prior simulations, so we do not further consider them here [85, 109].

The latter class of resampling approaches obviates a key assumption used by both minP and Wstep in order to simplify computation. This disputed "subset pivotality" assumption that has been discussed at length elsewhere (e.g., [83, 120–122]). To summarize, strong FWER-control methods that empirically estimate the correlation structure must control FWER not only when the global null holds, but also for any configuration of true and false null hypotheses. Although it might appear that resamples would therefore need to be generated under every such configuration, [121]'s methods circumvent this problem, requiring only one set of resamples under the global null, by invoking subset pivotality. This assumption states that for any subset K of hypotheses being tested, if all null hypotheses in K hold, then the distribution of the maximum test statistic in K is the same regardless of the truth or falsehood of all hypotheses not in K. (See [120] for a rigorous definition.)

Subset pivotality can fail, for example, when testing pairwise correlations ([83]'s Example 4.1) of three variables, X, Y, and Z. In this setting, the joint distribution of the statistics $\widehat{\rho}_{XY}$ and $\widehat{\rho}_{XZ}$ when a particular subset K of the null hypotheses hold, namely $\rho_{XY} = \rho_{XZ} = 0$, depends on ρ_{YZ} and hence on the truth or falsehood of a hypothesis not in K [83]. Thus, under [121]'s resampling approach, $\widehat{\rho}^{(j)}_{XY}$ and $\widehat{\rho}^{(j)}_{XZ}$ would be correctly centered at 0, but they would be independent because the global null is enforced. In contrast, under [85]'s resampling approach, $\widehat{\rho}^{(j)}_{XY}$ and $\widehat{\rho}^{(j)}_{XZ}$ would likewise be centered at 0 because they would have been centered by the sample estimates $\widehat{\rho}_{XY}$ and $\widehat{\rho}_{XZ}$, but they would also be correlated to an extent determined by ρ_{YZ}. In turn, the distribution of the maximum test statistic depends on the joint distribution of $\left(\widehat{\rho}_{XY}, \widehat{\rho}_{XZ}\right)$. Importantly, subset pivotality will not be required for our proposed methods: unlike FWER-control methods, our proposed methods concern only the global null, and thus even when subset pivotality does not hold, it is sufficient to estimate via resampling the single sampling distribution of the test statistics under the global null. To build upon these existing methods by directly characterizing global evidence strength, we now develop theory underlying our proposed metrics.

Table 3.3.1: Selected existing methods for strong FWER control with correlated hypothesis tests

Method	Type	Means of handling correlation
Bonferroni	1-step	Conservatively making no assumptions
Holm	Step-down	Conservatively making no assumptions
minP	1-step	Resampling under global null to estimate correlation structure
Wstep	Step-down	——————〃——————
Romano	Step-down	Resampling without restrictions to estimate correlation structure

3.4 Setting and notation

Suppose that K random variables are measured on N subjects, with the resulting matrix denoted $Z \in \mathbb{R}^{N \times K}$. Let Z_{nk} denote, for the n^{th}

subject, the k^{th} random variable. Consider a resampling algorithm that generates, for iterate j, a dataset $\mathrm{Z}^{(j)} \in \mathbb{R}^{N \times K}$ containing the random vector $\left(\mathrm{Z}^{(j)}_{n1}, \cdots, \mathrm{Z}^{(j)}_{nK}\right)$ for each subject n. There are a total of B resampled datasets. We use the superscript "$^{(j)}$" to denote random variables, distributions, and statistics in resampled dataset j. Further suppose that we conduct W tests of point null hypotheses, each at level α. Denote the w^{th} null hypothesis as H_{0w}. Let $c_{w,\alpha}$ be the critical value for the test statistic, T_w, of the w^{th} test. The W-vector of test statistics is $T = (T_1, \cdots, T_w)$. We define the "global null" as the case in which all W null hypotheses hold and use the superscript "$^{\circ}$" generally to denote distributions, data, or statistics generated under the global null.

Define the statistic corresponding to the observed number of α-level rejections as $\widehat{\theta} = \sum_{w=1}^{W} 1\{T_w > c_{w,\alpha}\}$. Its counterpart in a sample generated under the global null is $\widehat{\theta}^{\circ}$ and in resample j is $\widehat{\theta}^{(j)}$. Using F to denote cumulative distribution functions (CDFs), respectively define the true CDF of the number of rejections under the global null, its counterpart in the resamples, and its empirical estimator in the resamples as:

$$
\begin{aligned}
F_{\widehat{\theta}^{\circ}}(r) &= P\left(\widehat{\theta}^{\circ} \leq r\right) \\
F_{\widehat{\theta}^{(j)}}(r) &= P\left(\widehat{\theta}^{(j)} \leq r\right) \\
\widehat{F}_{\widehat{\theta}^{(j)}}(r) &= \frac{1}{B} \sum_{j^*=1}^{B} 1\left\{\widehat{\theta}^{(j^*)} \leq r\right\}
\end{aligned}
$$

We denote almost sure convergence, convergence in probability, convergence in distribution, and ordinary limits respectively as "$\xrightarrow[N \to \infty]{A.S.}$", "$\xrightarrow[N \to \infty]{P}$", "$\xrightarrow[N \to \infty]{D}$", and "$\xrightarrow[N \to \infty]{}$".

3.5 Main results

We now develop theory allowing us to approximate the sampling distribution of $F_{\widehat{\theta}^{\circ}}$ through resampling. Specifically, we show that under a certain class of resampling algorithms defined below, the empirical sampling distribution of the number of rejections in the resamples converges to the true distribution of the number of rejections in samples generated under the global null. We chose to characterize the sampling distribution empirically rather than theoretically because it does not appear to have a tractable closed form without imposing assumptions on the correlation structure of the tests and potentially requiring asymptotics on the number of hypothesis tests. (Despite the intractable sampling distribution, it is straightforward to derive at least the exact variance of $\widehat{\theta}^{\circ}$ if the pairwise correlations between the p-values are known (Appendix Section C.1).) Because simulation error associated with using a finite number of resamples to approximate the CDF of the resampled data can be made arbitrarily small by taking $B \to \infty$, we follow convention (e.g., [39]) in ignoring this source of error and considering only asymptotics on N.

3.5.1 An assumption on the resampling algorithm

To establish the main convergence result, we will use the following key assumption stating that, regardless of whether the observed sample was generated under the global null or under an alternative, the resampling algorithm must generate a sampling distribution for $T^{(j)}$ that converges to the sampling distribution of T° (that is, in samples generated under the global null). We will later discuss resampling algorithms that satisfy this assumption.

Assumption 3.5.3. The resampling algorithm used to generate $Z^{(j)}$

must ensure that $T^{(j)} \xrightarrow[N\to\infty]{D} T^{\circ}$, or equivalently $F_{T^{(j)}} \xrightarrow[N\to\infty]{} F_{T^{\circ}}$.

Typically, resampling algorithms fulfilling this assumption will need to preserve the correlation structure of all variables in the dataset, except where the global null dictates otherwise. If not, the distribution of the test statistics will usually not be preserved. Additionally, just as the original data are assumed to respect the parametric assumptions of all W hypothesis tests, the resampled data must be generated in a manner that also respects this parametric structure. Otherwise, hypothesis tests conducted on the resampled data may not preserve their nominal α-levels, which again affects the distribution of the test statistics.

Remark 3.5.1. For Assumption 3.5.3 to hold, it is sufficient for T to be a continuous function of Z and for $Z^{(j)} \xrightarrow[N\to\infty]{D} Z^{\circ}$. Note that this condition is not necessary; for example, [121] proposes several algorithms that induce the global null by centering the data themselves by sample estimates, rather than by centering the test statistics as in Algorithm 3.5.1 below. In such cases, Assumption 3.5.3 may hold without $Z^{(j)} \xrightarrow[N\to\infty]{D} Z^{\circ}$.

3.5.2 Valid residual resampling for OLS

We now consider the design of valid resampling algorithms for the familiar setting of ordinary least squares (OLS) multiple regression. Specializing the notation in Section 3.4, let M_{ij} denote the $(i,j)^{th}$ element of a matrix, M, and V_i denote the i^{th} element of a vector, V. Assume that each of W outcome variables, $(Y_1, \cdots, Y_W)$, is regressed on the same design matrix, $X \in \mathbb{R}^{N \times p}$, comprising an intercept term denoted X_1 (such that the residuals have mean 0), a single exposure of interest (taken without loss of generality to be X_2), and the adjusted covariates $(X_3, \cdots, X_C)$. Assume all covariates besides the intercept are mean-centered. Thus, the dataset Z contains a random vector $(1, X_{n2}, \cdots, X_{nC}, Y_{n1}, \cdots, Y_{nW})$ for each subject n.

Let $\varepsilon_w = (\varepsilon_{1w}, \cdots, \varepsilon_{Nw})$ denote the N-vector of true errors for the w^{th} regression such that $\varepsilon_{nw} \sim N(0, \sigma_w^2)$. Let $\widehat{\varepsilon_w}$ be its estimated counterpart (the residuals). Let $\sigma_w^2 = E[\varepsilon_{nw}^2 | X]$ as usual, and assume $\sigma_w^2 < \infty$. Letting β_w denote the coefficient of interest for X_2 in the w^{th} regression model and α_{jw} denote the nuisance coefficient for the j^{th} adjusted covariate or intercept in the w^{th} model, the W models are:

$$Y_{n1} = \alpha_{11} + \beta_1 X_{n2} + \sum_{j=3}^{C} \alpha_{j1} X_{nj} + \varepsilon_{n1}$$

$$\vdots$$

$$Y_{nW} = \alpha_{1W} + \beta_W X_{n2} + \sum_{j=3}^{C} \alpha_{jW} X_{nj} + \varepsilon_{nW} \tag{3.1}$$

Using superscripts to denote lengths and subscripts to denote indices, let $\beta^W = \left(\beta_1, \cdots, \beta_W\right)$ be a vector containing, for each of W regression models, the single coefficient of interest, and let $\widehat{\beta}^W$ and $\widehat{\beta}^{W(j)}$ denote its sample estimate in the original dataset and in resampled dataset j, respectively. Suppose without loss of generality that the null hypotheses of interest are $H_{0w} : \beta_w = 0$.

Letting hats denote the usual OLS estimates obtained from the original sample, the usual test statistics in the original sample are $T = \left(\frac{\widehat{\beta}_1}{\widehat{\sigma}_1 (X'X)^{-1/2}}, \cdots, \frac{\widehat{\beta}_W}{\widehat{\sigma}_W (X'X)^{-1/2}}\right)$; their unobservable counterparts centered to reflect the global null are $T^\circ = \left(\frac{\widehat{\beta}_1 - \beta_1}{\widehat{\sigma}_1 (X'X)^{-1/2}}, \cdots, \frac{\widehat{\beta}_W - \beta_W}{\widehat{\sigma}_W (X'X)^{-1/2}}\right)$.

Algorithm 3.5.1 (A valid resampling algorithm for OLS). A parametric resampling algorithm satisfying Assumption 3.5.3 (generalized from [39]) is to first fix the covariates $(X_1, \cdots, X_C)$ for all observations while setting the resampled "outcomes" equal to the fitted values plus a vector of residuals resampled with replacement.

That is, letting n' denote an observation sampled with replacement, the resampled variables for observation n are:

$$X_{n1}^{(j)} := X_{n1}$$
$$\vdots$$
$$X_{nC}^{(j)} := X_{nC}$$

Then each test statistic in the resamples is computed using $\widetilde{H}_{0w} : \beta_w = \widehat{\beta}_w$ in order to recover the null sampling distribution [43]. That is:

$$T^{(j)} = \left(\frac{\widehat{\beta}_1^{(j)} - \widehat{\beta}_1}{\widehat{\sigma}_1^{(j)} (X'X)^{-1/2}}, \cdots, \frac{\widehat{\beta}_W^{(j)} - \widehat{\beta}_W}{\widehat{\sigma}_W^{(j)} (X'X)^{-1/2}} \right) \tag{3.2}$$

We show later that this resampling algorithm fulfills Assumption 3.5.3 because the distribution of each $\widehat{\beta}_w - \beta_w = (X'X)^{-1}X'\varepsilon_w$ (in the original sample) depends only on the true error distribution and not on β_w, so the resampling algorithm need only recover the true error distribution to provide valid inference under the global null.

Various other approaches that may appear valid in fact violate the assumption. For example, we could fix the design matrix X but resample with replacement the outcome vectors, $(Y_{n'1}, \cdots, Y_{n'W})$, rather than the residuals:

$$X_{n1}^{(j)} := X_{n1}$$
$$\vdots$$
$$X_{nC}^{(j)} := X_{nC}$$

Although this approach indeed enforces the global null and preserves the correlation between the outcomes, it fails to preserve the correlations between the outcomes and the adjusted covariates and thus does not recover the distribution of T°.

A second incorrect alternative would be to bootstrap parametrically from Equation (3.1) while enforcing the global null by constraining each $\beta_w^{(j)} := 0$:

$$
\begin{aligned}
X_{n1}^{(j)} &:= X_{n1} \\
&\vdots \\
X_{nC}^{(j)} &:= X_{nC}
\end{aligned}
$$

where $\varepsilon_{nw}^{(j)} \sim_{i.i.d.} N(0, \widehat{\sigma}^2_{\varepsilon_{nw}})$. However, this sequential algorithm fails to entirely preserve the correlations among the outcomes if there are unmeasured variables, beyond the adjusted covariates, that contribute to these correlations. In turn, the distribution of T° is not recovered. A final incorrect alternative would be a generic bootstrap hypothesis test performed by resampling with replacement entire rows of data and then centering the test statistics as in Equation (3.2). However, this algorithm incorrectly treats the design matrix as random rather than fixed, which would be appropriate for correlation models but not the intended regression models [39]. Additionally, this algorithm can produce data violating the assumptions of standard OLS inference, even when the original data fulfill the assumptions. Suppose, for example, that the design matrix contains only an intercept and a binary exposure of interest, $X_2 \in \{0, 1\}$, and that, for some outcome Y_{w^*}, we have $\beta_{w^*} \neq 0$ (i.e., the alternative hypothesis holds). Then, of course, $\widehat{\varepsilon}_{nw^*}$ may be normal, allowing valid OLS inference, despite that

Y_{w^*} itself may be bimodal with peaks at $E[Y_{w^*} \mid X_2 = 0]$ and $E[Y_{w^*} \mid X_2 = 1]$. This generic resampling algorithm retains the bimodality of Y_{w^*} while breaking the association between Y_{w^*} and X_1; thus, the resampled residuals $\widehat{\varepsilon}_{nw^*}^{(j)}$ will be bimodal rather than normal [121], and standard inference may fail.

Justification of Algorithm 3.5.1

In Theorem 3.5.3 below, we show that Algorithm 3.5.1 satisfies Assumption 3.5.3. The development of the proof is structured as follows. We make a regularity assumption (Assumption 3.5.4) and define how we will metrize convergence of the resampled test statistics (Definition 3.5.1). We bound the distance metric for certain types of random vectors (Lemma 3.5.1), in turn allowing us to bound the distance between the estimated sampling distribution in the resamples and the true sampling distribution to which the former should converge (Theorem 3.5.1). Using the latter bound, a triangle inequality argument, and convergence results regarding each term of the triangle inequality (Lemmas 3.5.2 and 3.5.3), we show the needed convergence result for the coefficient estimates (Theorem 3.5.2) and finally for the test statistics (Theorem 3.5.3).

First, assume the following regularity condition on the design matrix, which will later be relevant for the convergence of the coefficient estimates:

Assumption 3.5.4. Suppose without loss of generality that the regression covariate of interest is X_2. Correspondingly, let $B \in \mathbb{R}^{N \times 1}$ be the transposed second row of $(X'X)^{-1}X'$, or equivalently the first column of $X(X'X)^{-1}$. (More generally, if the covariate of interest is the i^{th} variable in the design matrix, then B is defined as the i^{th} row or column.) Assume that for some constant $k > 0$:

$$N \cdot B'B \xrightarrow[N\to\infty]{P} k$$
$$\Leftrightarrow N \sum_{n=1}^{N} [(X'X)^{-1}X']_{2n}^{2} \xrightarrow[N\to\infty]{P} k$$

Remark 3.5.2 (Sufficient conditions). Let $\mathcal{I}_{ij}$ denote an entry of the expected Fisher information matrix for an individual observation in the w^{th} regression. Then Assumption 3.5.4 holds if, for all w:

(A1) $\mathcal{I}_{ii} > 0 \; \forall \; i \in \{1, \cdots, p\}$

(A2) $E[X_{ni}X_{nj}] < \infty \; \forall \; i, j \in \{1, \cdots, p\}$

(A3) $\sigma_w^2 > 0$

(A1) states that the true standard errors of all p regression coefficients are finite. (A2) holds if the covariates are non-collinear and have finite expectations. (A3) states that the model does not fit perfectly.

Proof. Let $\widehat{\alpha_{iw}}$ be the i^{th} coefficient estimate in the w^{th} regression, such that $\widehat{\alpha_{2w}} = \widehat{\beta_w}$, the estimate of interest. Thus, let $\widehat{\alpha_w} = [\widehat{\alpha_{1w}}, \widehat{\beta_w}, \widehat{\alpha_{3w}}, \cdots, \widehat{\alpha_{pw}}]'$ be the p-vector of estimates in the w^{th} regression. Denote a pairwise covariance $\text{Cov}_{ij} = \text{Cov}\left(\widehat{\alpha_{iw}}, \widehat{\alpha_{jw}}\right)$, and similarly denote a pairwise correlation as ρ_{ij}. Then the estimated covariance of $\widehat{\beta_w}$ with $\widehat{\alpha_{iw}}$ is:

$$\begin{aligned}\widehat{\text{Cov}_{2i}} &= \widehat{\rho_{2i}} \cdot \widehat{\text{SE}}\left(\widehat{\beta_w}\right) \cdot \widehat{\text{SE}}\left(\widehat{\alpha_{iw}}\right) \\ &= \widehat{\rho_{2i}} \cdot \frac{1}{\sqrt{N\widehat{\mathcal{I}}_{22}}} \cdot \frac{1}{\sqrt{N\widehat{\mathcal{I}}_{ii}}} \qquad (3.3)\end{aligned}$$

With the LHS of Assumption 3.5.4 in view, we have:

$$(X'X)^{-1} = \frac{1}{\widehat{\sigma}_w^2}\begin{bmatrix} \widehat{\mathrm{Cov}}_{11} & \cdots & \widehat{\mathrm{Cov}}_{1p} \\ \widehat{\mathrm{Cov}}_{21} & \cdots & \widehat{\mathrm{Cov}}_{2p} \\ \vdots & & \vdots \\ \widehat{\mathrm{Cov}}_{p1} & \cdots & \widehat{\mathrm{Cov}}_{pp} \end{bmatrix}$$

$$[(X'X)^{-1}X']_{2n} = \frac{1}{\widehat{\sigma}_w^2}\left[\widehat{\mathrm{Cov}}_{21} \cdots \widehat{\mathrm{Cov}}_{2p}\right]\left[X_{n1} \cdots X_{np}\right]'$$

$$\begin{aligned} N\sum_{n=1}^{N}[(X'X)^{-1}X']_{2n}^2 &= N\sum_{n=1}^{N}\left(\sum_{i=1}^{p}\frac{1}{\widehat{\sigma}_w^2}\widehat{\mathrm{Cov}}_{2i}X_{ni}\right)^2 \\ &= N\frac{1}{\widehat{\sigma}_w^4}\sum_{n=1}^{N}\left(\sum_{i=1}^{p}\sum_{j=1}^{p}\widehat{\mathrm{Cov}}_{2i}\widehat{\mathrm{Cov}}_{2j}X_{ni}X_{nj}\right) \\ &= N\frac{1}{\widehat{\sigma}_w^4}\sum_{i=1}^{p}\sum_{j=1}^{p}\widehat{\mathrm{Cov}}_{2i}\widehat{\mathrm{Cov}}_{2j}\sum_{n=1}^{N}X_{ni}X_{nj} \end{aligned}$$

Re-expressing the estimated covariances using Equation (3.3):

$$= \frac{1}{\widehat{\sigma}_w^4}\sum_{i=1}^{p}\sum_{j=1}^{p}\widehat{\rho}_{2i}\ \widehat{\rho}_{2j}\ \frac{1}{\widehat{\mathcal{I}}_{22}\sqrt{\widehat{\mathcal{I}}_{ii}\widehat{\mathcal{I}}_{jj}}}\frac{1}{N}\sum_{n=1}^{N}X_{ni}X_{nj}$$

$$\xrightarrow[N\to\infty]{P} \frac{1}{\sigma_w^4}\sum_{i=1}^{p}\sum_{j=1}^{p}\rho_{2i}\ \rho_{2j}\ \frac{1}{\mathcal{I}_{22}\sqrt{\mathcal{I}_{ii}\mathcal{I}_{jj}}}\ E[X_{ni}X_{nj}]$$

If (A1)–(A3) above are fulfilled, this is a finite constant, as required.

□

We will consider validity of the bootstrap in terms of convergence on the Mallows-Wasserstein metric, a conventional choice that is defined as follows [27, 39].

Definition 3.5.1. Let G_A and G_B be arbitrary marginal distribution functions for random vectors $A \in \mathbb{R}^W$ and $B \in \mathbb{R}^W$, respectively. Then a form of Mallows-Wasserstein distance between G_A and G_B is the infimum, taken over all possible joint distributions for (A, B) such that $A \sim G_A$ and $B \sim G_B$ marginally, of the expected L_2 distance between A and B:

$$d_2^W(G_A, G_B) := \inf_{\substack{A \sim G_A \\ B \sim G_B}} E[||A - B||^2]^{1/2}$$

We proceed to prove that the residual-resampling bootstrap is consistent with respect to the Mallows-Wasserstein metric in a development roughly following [39] and [8], who considered the asymptotic validity of residual resampling in recovering the sampling distribution of a p-vector of coefficient estimates from a single multiple linear regression model. Here, we extend this work to consider the sampling distribution of $\widehat{\beta}^W$. We first establish a lemma bounding the Mallows-Wasserstein distance between the distributions of two random vectors constructed as products of different random matrices with a single fixed vector.

Lemma 3.5.1. Let C^* and $D^* \in \mathbb{R}^{W \times N}$ be random matrices from a specific joint distribution, and let $B \in \mathbb{R}^{N \times 1}$ be a fixed vector. Let G_C and G_D be the resulting marginal distribution functions of the vectors C^*B and $D^*B \in \mathbb{R}^{W \times 1}$, respectively. Then:

$$d_2^W(G_C, G_D)^2 \leq \mathrm{tr}\Big\{BB' \cdot E\big[(C^* - D^*)'(C^* - D^*)\big]\Big\}$$

Proof.

$$
\begin{aligned}
d_2^W(G_C, G_D)^2 &\leq E\Big[\|C^*B - D^*B\|^2\Big] \\
&= E\Big[\mathrm{tr}\{\underbrace{(C^*B - D^*B)}_{W\times 1}\underbrace{(C^*B - D^*B)'}_{1\times W}\}\Big] \\
&= E\Big[\mathrm{tr}\{\underbrace{(C^* - D^*)}_{W\times N}\underbrace{BB'}_{N\times N}\underbrace{(C^* - D^*)'}_{N\times W}\}\Big] \\
&= E\Big[\mathrm{tr}\{\underbrace{BB'(C^* - D^*)'(C^* - D^*)}_{N\times N}\}\Big] \\
&= \mathrm{tr}\Big\{E[BB'(C^* - D^*)'(C^* - D^*)]\Big\} \\
&= \mathrm{tr}\Big\{BB' \cdot E[\underbrace{(C^* - D^*)'(C^* - D^*)}_{N\times N}]\Big\}
\end{aligned}
$$

The inequality arises because the left-hand side is the infimum of the expectation over all possible joint distributions with marginals G_C and G_D, whereas the right-hand side is the expectation for a particular such joint distribution (the one giving rise to C^* and D^*). □

The next theorem bounds the distance between the true sampling distribution of the estimated coefficients and the estimated sampling distribution in the resamples in terms of the distance between the sampling distribution of the true errors and the resampled residuals.

Theorem 3.5.1. Let F denote the distribution function of the true errors for the W regression models, $(\varepsilon_{n1}, \cdots, \varepsilon_{nW})$, and let $\widehat{F}_N$ denote the empirical distribution function of the residuals, which is used to approximate F in Algorithm 3.5.1. Let $\Psi(F)$ denote the distribution of the standardized coefficient estimates, $\sqrt{N}\left(\widehat{\beta}^W - \beta^W\right)$, that are constructed as a function of the true error distribution; $\Psi(F)$ therefore represents the true sampling distribution to which a valid bootstrapped sampling distribution must converge. In contrast, let

$\Upsilon(\widehat{F}_N)$ be the distribution of the standardized coefficient estimates in the resamples, $\sqrt{N}\left(\widehat{\beta}^{W(j)} - \widehat{\beta}^{W}\right)$, in which the empirical distribution of the residuals is used to approximate the true distribution.

As in Assumption 3.5.4, let $B \in \mathbb{R}^{N\times 1}$ be the transposed second row of $(X'X)^{-1}X'$. Then:

$$d_2^W\left(\Upsilon(F), \Upsilon(\widehat{F}_N)\right)^2 \leq N \cdot \mathrm{tr}\{BB'\} \cdot d_2^W\left(F, \widehat{F}_N\right)^2$$

Proof. Let $U'_w \in \mathbb{R}^{1\times N} = [U_{1w}, \cdots, U_{Nw}]$ such that $(U_{n1}, \cdots, U_{nW}) \sim F$ and:

$$C \in \mathbb{R}^{W\times N} = \begin{bmatrix} — & U'_1 & — \\ & \vdots & \\ — & U'_W & — \end{bmatrix} = \begin{bmatrix} U_{11} & \dots & U_{N1} \\ U_{12} & \dots & U_{N2} \\ \vdots & & \vdots \\ U_{1W} & \dots & U_{NW} \end{bmatrix}$$

In general for multiple regression, we have $\widehat{\beta} - \beta = (X'X)^{-1} X'\varepsilon$. Thus, we can express $\Upsilon(F)$ as the distribution of the W-vector:

$$\sqrt{N}\left(\widehat{\beta}^{W} - \beta^{W}\right) = \sqrt{N}\begin{bmatrix} \left[(X'X)^{-1}X'U_1\right]_2 \\ \vdots \\ \left[(X'X)^{-1}X'U_W\right]_2 \end{bmatrix} = \sqrt{N}\begin{bmatrix} U'_1B \\ \vdots \\ U'_WB \end{bmatrix} = \sqrt{N} \cdot CB$$

whose w^{th} element pertains to the regression coefficient for X_2 in the w^{th} regression. Let D be the counterpart of C with $\left(\widehat{U}_{n1}, \cdots, \widehat{U}_{nW}\right) \sim \widehat{F}_N$ in place of $(U_{n1}, \cdots, U_{nW})$.

In view of Lemma 3.5.1, note that the entries of the matrix

$(C-D)'(C-D) \in \mathbb{R}^{N\times N}$ are:

$$\begin{aligned}[(C-D)'(C-D)]_{kj} &= \sum_{w=1}^{W}[(C-D)']_{kw}[C-D]_{wj} \\ &= \sum_{w=1}^{W}[C-D]_{wk}[C-D]_{wj} \\ &= \sum_{w=1}^{W}\left(U_{kw}-\widehat{U}_{kw}\right)\left(U_{jw}-\widehat{U}_{jw}\right)\end{aligned}$$

We have $E\Big[\left(U_{kw}-\widehat{U}_{kw}\right)\left(U_{jw}-\widehat{U}_{jw}\right)\Big] = \mathrm{Cov}\left(U_{kw}-\widehat{U}_{kw}, U_{jw}-\widehat{U}_{jw}\right)$, but for all $k \neq j$, the covariance is 0 because the observations are independent. Thus, letting I^N denote the $N \times N$ identity matrix, we have that $E[(C-D)'(C-D)]$ is a diagonal matrix such that

$$E[(C-D)'(C-D)] = I^N \cdot E\Big[\sum_{w=1}^{W}\left(U_{jw}-\widehat{U}_{jw}\right)^2\Big] \tag{3.4}$$

In order to apply Lemma 3.5.1, we will now restrict attention to a special choice of C and D. First note that, by definition:

$$d_2^W\left(F,\widehat{F}_N\right)^2 = \inf_{\substack{(U_{j1},\cdots,U_{jW})\sim F \\ (\widehat{U}_{j1},\cdots,\widehat{U}_{jW})\sim \widehat{F}_N}} E\Big[\sum_{w=1}^{W}\left(U_{jw}-\widehat{U}_{jw}\right)^2\Big] \tag{3.5}$$

Now let $C^* \in \mathbb{R}^{W\times N}$ and $D^* \in \mathbb{R}^{W\times N}$ be a pair of random matrices constructed using $\left(U_{j1},\cdots,U_{jW}\right)$ and $\left(\widehat{U}_{j1},\cdots,\widehat{U}_{jW}\right)$ from the infimum-attaining joint distribution in Equation (3.5); that is, such that $E[(C^*-D^*)'(C^*-D^*)] = I^N \cdot d_2^W\left(F,\widehat{F}_N\right)^2$ per the representations in Equations (3.4) and (3.5). (Such a choice exists by [8]'s Lemma 8.1.) The result then follows immediately from applying Lemma 3.5.1, setting $G_C = \Psi(F)$, $G_D = \Psi\left(\widehat{F}_N\right)$, and B, C^*, and D^* as defined above and pulling the scalar $\sqrt{N}$ outside the squared distance. □

Next, to apply the bound in Theorem 3.5.1, we will first bound the term on the right-hand side using a triangle inequality, which applies because $d_2^W(\cdot,\cdot)$ is a metric [8]. To this end, let F_N denote the unobserved empirical distribution function of the true error vector, ε^W. Then we have the following triangle inequality:

$$d_2^W\left(\widehat{F}_N, F\right) \leq d_2^W\left(\widehat{F}_N, F_N\right) + d_2^W\left(F_N, F\right) \tag{3.6}$$

The first term on the RHS relates the empirical distribution of the residuals to the empirical distribution of the true errors (which are both discrete distributions taking N values); the second term relates the latter empirical distribution to the true error distribution (which is continuous). The next two lemmas bound the terms on the RHS of Equation (3.6); we will later use them to bound the LHS.

Lemma 3.5.2. For the expectation of the first term on the RHS of Equation (3.6):

$$E\left[d_2^W\left(\widehat{F}_N, F_N\right)\right] \xrightarrow[N\to\infty]{} 0$$

Proof. As in Definition 3.5.1, let $U \sim \widehat{F}_N$ and $V \sim F_N$ be arbitrary random variables in $\mathbb{R}^W$ that follow the empirical marginal distributions of the residuals and of the true errors. Denote their elements $(U_1, \cdots, U_W)$ and $(V_1, \cdots, V_W)$. Let (U^*, V^*) be the special choice of (U, V) that follow not only the marginal empirical distributions $\widehat{F}_N$ and F_N, but also the empirical joint distribution of the residuals and the true errors. Then:

$$\begin{aligned} d_2^W\left(\widehat{F}_N, F_N\right)^2 &:= \inf_{\substack{U\sim\widehat{F}_N \\ V\sim F_N}} E[||U-V||^2] \\ &\leq E\left[||U^*-V^*||^2\right] \end{aligned}$$

because (U^*, V^*) represents a choice of a single element from the set over which the infimum is taken. Expressing the RHS as the expectation of the joint ECDF:

$$= \frac{1}{N} \sum_{n=1}^{N} \underbrace{\sum_{w=1}^{W} \left(\widehat{\varepsilon}_{nw} - \varepsilon_{nw}\right)^2}_{\|\cdot\|^2 \text{ of a } W\text{-vector}}$$

$$= \frac{1}{N} \sum_{w=1}^{W} \underbrace{\sum_{n=1}^{N} \left(\widehat{\varepsilon}_{nw} - \varepsilon_{nw}\right)^2}_{\|\cdot\|^2 \text{ of an } N\text{-vector}}$$

$$= \frac{1}{N} \sum_{w=1}^{W} \|\widehat{\varepsilon}_{w} - \varepsilon_{w}\|^2$$

Taking expectations and using [39]'s Eq. (2.2), this implies:

$$E\left[d_2^W\left(\widehat{F}_N, F_N\right)^2\right] = \frac{p}{N} \sum_{w=1}^{W} \sigma_w^2$$

$$\xrightarrow[N\to\infty]{} 0$$

By Jensen's inequality:

$$E\left[d_2^W\left(\widehat{F}_N, F_N\right)\right] \xrightarrow[N\to\infty]{} 0$$

The interchange of summations in lines 3-4 is used to express W norms involving residuals from different regressions, summed over N observations, as N norms involving residuals of observations within a regression, summed over W regressions. The latter is more convenient because it allows application of existing theory for a single multiple regression model. □

Lemma 3.5.3. For the second term on the RHS of Equation (3.6):

$$d_2^W(F_N, F) \xrightarrow[N\to\infty]{P} 0$$

Proof. Letting P_N denote an empirical probability, F_N can be expressed as:

$$\begin{aligned} P_N(\varepsilon_{n1} \leq c_1, \cdots, \varepsilon_{nW} \leq c_W) &= \frac{1}{N}\sum_{n=1}^{N} 1\{\varepsilon_{n1} \leq c_1, \cdots, \varepsilon_{nW} \leq c_W\} \\ &\xrightarrow[N\to\infty]{A.S.} P(\varepsilon_{n1} \leq c_1, \cdots, \varepsilon_{nW} \leq c_W) \end{aligned}$$

with the last line following from the SLLN. Thus, $F_N \xrightarrow[N\to\infty]{A.S.} F$. Also by the SLLN, $\int \|x\|^p \, F_N(dx) \xrightarrow[N\to\infty]{A.S.} \int \|x\|^p \, F(dx)$ because the LHS is a sample average whereas the RHS is its true expectation. These two results immediately imply condition (a) of [8]'s Lemma 8.3, which yields $d_2^W(F_N, F)^2 \xrightarrow[N\to\infty]{P} 0$ and hence the desired result. □

Theorem 3.5.2. The residual bootstrap is weakly consistent under the Mallows-Wasserstein metric for the OLS coefficient estimates (Definition 29.2 of [27]); that is:

$$d_2^W\left(\Psi(F), \Psi(\widehat{F}_N)\right) \xrightarrow[N\to\infty]{P} 0$$

Proof. Combining Theorem 3.5.1 with the triangle inequality in Equation (3.6) and observing that $\text{tr}\{BB'\} = \sum_{n=1}^{N} B_N^2 \geq 0$ yields:

$$d_2^W\left(\Psi(F), \Psi(\widehat{F}_N)\right) \leq \sqrt{N \cdot \text{tr}\{BB'\}} \cdot \left(d_2^W\left(\widehat{F}_N, F_N\right) + d_2^W(F_N, F)\right)$$

The term $\sqrt{N \cdot \text{tr}\{BB'\}} \xrightarrow[N\to\infty]{P} k$ by Assumption 3.5.4 because BB' is scalar. By Markov's inequality, the convergence in mean of Lemma 3.5.2 implies that $d_2^W\left(\widehat{F}_N, F_N\right) \xrightarrow[N\to\infty]{P} 0$. Last, by Lemma 3.5.3,

$d_2^W(F_N, F) \xrightarrow[N\to\infty]{P} 0$, so the desired result holds. □

The next theorem uses the above result regarding convergence of the resampling-based coefficient estimates to establish convergence of the test statistics.

Theorem 3.5.3. Algorithm 3.5.1 fulfills Assumption 3.5.3; namely:

$$T^{(j)} \xrightarrow[N\to\infty]{D} T^{\circ}$$

Proof. By [8]'s Lemma 8.3, Theorem 3.5.2 implies that

$$\sqrt{N}\left(\widehat{\beta}^{W(j)} - \widehat{\beta}^{W}\right) \xrightarrow[N\to\infty]{D} \sqrt{N}\left(\widehat{\beta}^{W} - \beta^{W}\right)$$

By [39]'s Theorem 2.2, each $\widehat{\sigma}_w^{(j)} \xrightarrow[N\to\infty]{P} \sigma_w$. The desired result then follows from the multivariate Slutsky's Theorem. □

3.5.3 Valid inference on the number of rejections

We now present the main theorem establishing that resampling algorithms fulfilling Assumption 3.5.3, such as Algorithm 3.5.1 for OLS, also yield valid inference on the number of rejections.

Theorem 3.5.4. Under Assumption 3.5.3, $\widehat{\theta}^{(j)} \xrightarrow[N\to\infty]{D} \widehat{\theta}^{\circ}$.

Proof. Define the r-family of "rejection sets" as all possible configurations of the W test statistics that lead to r rejections:

$$\mathcal{A}_r = \left\{(A_1, \cdots, A_W) \in \mathbb{R}^W : (T_1 \in A_1, \cdots, T_W \in A_W) \implies \widehat{\theta} = r\right\}$$

Consider the limiting distribution of $\widehat{\theta}^{(j)}$:

$$\begin{aligned}\lim_{N\to\infty} P\left(\widehat{\theta}^{(j)} = r\right) &= \lim_{N\to\infty} P\left(\sum_{w=1}^{W} 1\left\{T_w^{(j)} > c_{w,a}\right\} = r\right)\\ &= \lim_{N\to\infty} \sum_{(A_1,\cdots,A_W)\in\mathcal{A}} P\left(T_1^{(j)} \in A_1, \cdots, T_W^{(j)} \in A_W\right)\\ &= \sum_{(A_1,\cdots,A_W)\in\mathcal{A}} P\left(T_1^{\circ} \in A_1, \cdots, T_W^{\circ} \in A_W\right)\\ &= P\left(\widehat{\theta}^{\circ} = r\right)\end{aligned}$$

where the second equality follows from Assumption 3.5.3. □

To summarize, Theorem 3.5.4 implies that valid inference, including the null interval and global test, can be conducted using the distribution of the number of rejections in resamples generated using an algorithm fulfilling Assumption 3.5.3.

3.6 Practical use and interpretation

In practice, to estimate the proposed metrics, one would first use a resampling algorithm fulfilling Assumption 3.5.3 to generate a large number of resamples under the global null (e.g., $B = 1,000$). Then, the lower and upper bounds of a 95% null interval can be defined as the 2.5th and 97.5th percentiles of $\left(\widehat{\theta}^{(1)}, \cdots, \widehat{\theta}^{(j)}\right)$, and the p-value for the global test is the empirical tail probability

$$P_N\left(\widehat{\theta}^{(j)} \geq \widehat{\theta}\right) = \sum_{j^*=1}^{B} 1\left\{\widehat{\theta}^{(j^*)} \geq \widehat{\theta}\right\}$$

We provide an R package, NRejections, to automate the resampling and estimation process for OLS models (see Appendix Section C.4).

The null interval can be interpreted as the plausible range of $\widehat{\theta}$ in samples generated under the global null. The excess hits, computed as the difference between $\widehat{\theta}$ and the upper limit of the null interval, can be interpreted as the number of rejections exceeding what would be expected in 95% of samples under the global null. Note, of course, that the excess hits is not equivalent to the number of "true" effects, a point we will reiterate in the Discussion. The p-value for the global test can be interpreted as the probability of observing at least $\widehat{\theta}$ rejections in samples generated under the global null. We further illustrate these interpretations in the following applied example.

3.7 Applied example

Existing epidemiologic analyses have investigated the extent to which an individual's experience of parental warmth during childhood is associated with the individual's later "flourishing" in mid-life [15]. Flourishing has been broadly conceived as a state of positive mental health comprising high emotional, psychological, and social well-being [54], and reductive analyses that individually assess its theorized components, such as perceived purpose in life and positive affect, may not fully capture potential impacts of the overall experience of flourishing [56, 113].

3.7.1 Methods

Closely reproducing [15]'s methods, we conducted longitudinal analyses of a subset of $N = 2,697$ subjects from the "Mid-life in the United States" (MIDUS) cohort study [13] of $7,108$ adults, recruited to include siblings and twin pairs, and for simplicity in these analyses, we randomly selected only one sibling from within each sibship. In an initial wave of data collection (1995-1996), subjects recalled the parental warmth that they experienced during childhood as an average of separate scales of maternal and paternal warmth. In a second wave

(2004-2006), the same subjects reported 13 continuous subscales of flourishing in emotional, psychological, and social domains [54].

We first reproduced [15]'s main analysis by assessing the association between a one-unit increase in standardized parental warmth (i.e., an increase of one standard deviation on the raw scale) with a standardized, continuous composite measure of flourishing ("overall flourishing"), which aggregated the 13 subscales per [54, 55]. We conducted similar analyses for the remaining 16 continuous outcome variables in [15]'s analyses, namely the 3 standardized composite scores for each domain (emotional, psychological, and social) treated separately and the 13 individual subscales. All of our analyses controlled for age, sex, race, nativity status, parents' nativity status, number of siblings, and other childhood family factors. We expected correlation among the resulting 17 test statistics both because of conceptual similarities between the subscale variables (e.g., social acceptance and social integration) and because of the composite and domain measures' direct arithmetic relationships with the subscales. Last, to characterize overall evidence strength across the 17 outcomes, we resampled per Algorithm 3.5.1 with $B = 5,000$ to estimate the proposed null interval and excess hits (with each test conducted at either $\alpha = 0.05$ or $\alpha = 0.01$) and to conduct the global test using the number of rejections in individual tests conducted at $\alpha = 0.05$. All data and code required to reproduce these analyses is publicly available and documented (https://osf.io/qj9wa/).

3.7.2 Results

Appendix Table C.2.1 displays demographics and childhood family characteristics in our sample, comprising all covariates adjusted in analysis. The 17 outcome measures had a median correlation magnitude of $|r| = 0.39$ (minimum = 0.12; maximum = 0.89; 25^{th}

percentile = 0.28; 75^{th} percentile = 0.55). The composite analysis estimated that, controlling for demographics and childhood family factors, individuals reporting an additional standard deviation (SD) of parental warmth in childhood experienced greater mid-life flourishing by, on average, $b = 0.22$ (95% CI: [0.18, 0.26]) SDs.

Of the 17 outcomes considered individually, all were "significantly" associated with parental warmth at $\alpha = 0.05$ (i.e., $\widehat{\theta} = 17$), and 15 were "significantly" associated at $\alpha = 0.01$ with a mean standardized effect size of $b = 0.14$. The directions of all effects suggested that increased parental warmth was associated with improved flourishing outcomes (Table 3.7.1). In contrast, if parental warmth were in fact unassociated with any of the outcomes, we would expect $17 \times 0.05 = 0.85$ rejections with a null interval of [0, 5] at $\alpha = 0.05$ (Figure 3.7.1). At $\alpha = 0.01$, we would expect 0.17 rejections with null interval [0, 2] at $\alpha = 0.01$. Thus, at $\alpha = 0.05$ and $\alpha = 0.01$ respectively, we observed $17 - 5 = 12$ and $15 - 2 = 13$ excess hits above what would be expected in 95% of samples under the global null. Indeed, a global test based on the number of rejections at $\alpha = 0.05$ suggested very strong evidence against the global null ($p = 0$ because every resampled dataset had < 17 rejections; Figure 3.7.1). (By comparison, simple inference based on the exact binomial distribution, assuming anticonservatively that the outcomes are independent, yields a too-narrow null interval at $\alpha = 0.05$ of $[0, 3]$ and a global p-value of $0.05^{17} = 7 \times 10^{-23}$.) Overall, our composite analyses strongly support small effects of parental warmth on composite flourishing, as reported by [15] (Table 3.7.1, first row); our novel analyses of $\widehat{\theta}$ additionally provide compelling global evidence for associations of parental warmth with flourishing across the 17 outcomes, accounting for their correlation structure.

Table 3.7.1: OLS estimate $(\widehat{\beta})$ characterizing association of a 1-SD increase in parental warmth with each of 17 standardized flourishing outcomes, adjusting for all covariates in Appendix Table C.2.1. Inference is not multiplicity-corrected.

Outcome	$\widehat{\beta}$ [95% CI]	p-value
Overall and domain composites		
Overall flourishing	0.22 [0.18, 0.26]	$< 2 \times 10^{-16}$
Emotional well-being	0.21 [0.17, 0.25]	$< 2 \times 10^{-16}$
Social well-being	0.13 [0.08, 0.17]	2×10^{-9}
Psychological well-being	0.20 [0.16, 0.24]	$< 2 \times 10^{-16}$
Emotional well-being subscales		
Positive affect	0.19 [0.15, 0.23]	$< 2 \times 10^{-16}$
Life satisfaction	0.19 [0.15, 0.23]	$< 2 \times 10^{-16}$
Social well-being subscales		
Meaningfulness of society	0.04 [0, 0.08]	0.048
Social integration	0.15 [0.11, 0.19]	5×10^{-13}
Social acceptance	0.09 [0.05, 0.13]	3×10^{-5}
Social contribution	0.09 [0.05, 0.13]	1×10^{-5}
Social actualization	0.06 [0.02, 0.11]	0.002
Psychological well-being subscales		
Autonomy	0.08 [0.04, 0.12]	3×10^{-4}
Environmental mastery	0.14 [0.09, 0.18]	6×10^{-11}
Personal growth	0.11 [0.07, 0.15]	4×10^{-7}
Positive relations	0.25 [0.21, 0.29]	$< 2 \times 10^{-16}$
Purpose in life	0.05 [0.01, 0.09]	0.018
Self-acceptance	0.22 [0.18, 0.26]	$< 2 \times 10^{-16}$

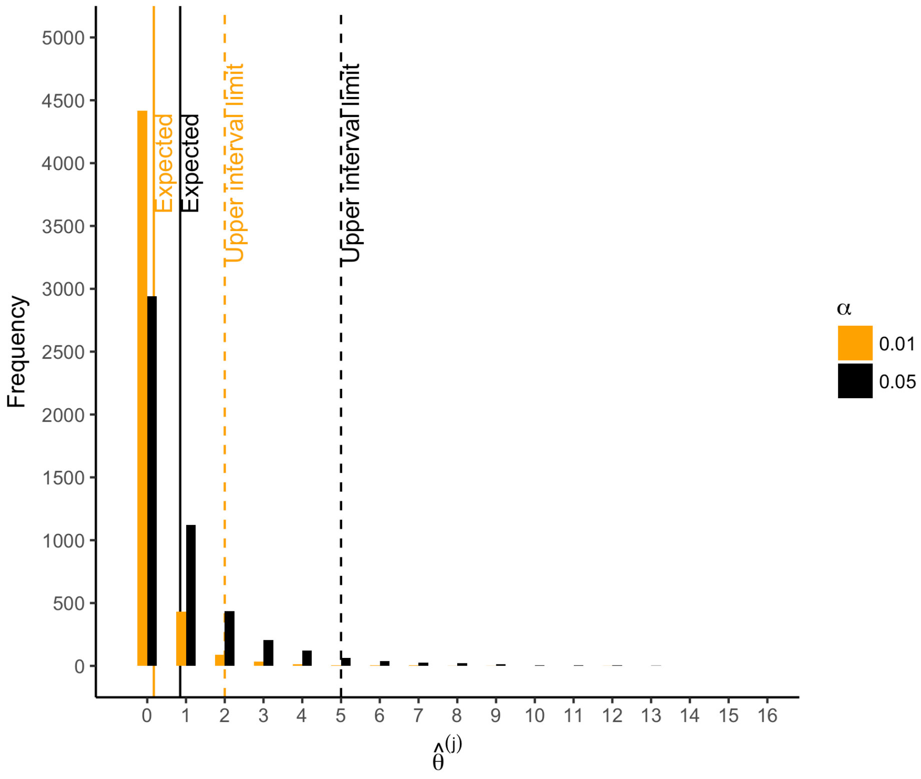

Figure 3.7.1: Number of rejections $(\widehat{\theta}^{(j)})$ for each of 5,000 resamples. Solid lines: $E[\widehat{\theta}^{\circ}] = \alpha \times 17$. Dashed lines: upper limit of 95% null interval.

3.8 Simulation study

We conducted a simulation study with two objectives. First, we aimed to visualize the null interval versus $\widehat{\theta}$ for varying effect sizes in an outcome-wide study and to characterize how its precision depends on the strength of correlation between the hypothesis tests and on the α level used for each test. Second, we aimed to assess the relative power of global tests conducted using the number of rejections with $\alpha = 0.05$ or $\alpha = 0.01$ for each individual test or derived from the five existing FWER-control methods listed in Table 3.3.1. All code required to reproduce the simulation study is publicly available (https://osf.io/qj9wa/).

3.8.1 Methods

We generated multivariate standard normal data, comprising 1 covariate (X) and 40 outcomes ($Y_1, \cdots, Y_{40}$) for a fixed $N = 1,000$ subjects. The correlation between each pair of outcomes was ρ_{YY}. The correlation between X and a proportion, q, of outcomes was ρ_{XY} (with remaining pairs having correlation 0). We manipulated scenario parameters in a full-factorial design (Table 3.8.1). Each of 500 simulations per scenario proceeded as follows:

1. We generated an observed dataset according to the scenario.

2. We regressed each outcome Y_w on X and conducted a t-test at level α on the coefficient for X. We computed $\widehat{\theta}$.

3. For each resampling iterate j (with $B = 1,000$), we resampled based on the algorithm in Algorithm 3.5.1. We conducted a t-test at level α on the coefficient for X and computed $\widehat{\theta}^{(j)}$.

4. We used the quantiles of $\left(\widehat{\theta}^{(1)}, \cdots, \widehat{\theta}^{(B)}\right)$ to construct the null

Table 3.8.1: Possible values of simulation parameters.

ρ_{XY}	ρ_{YY}	q	α
0.03	0	0	0.01
0.05	0.10	0.20	0.05
0.10	0.30	0.50	
0.15	0.60	1	

interval, compute the excess hits, and conduct our proposed joint test.

5. We used the t-statistics or p-values from the resamples to conduct joint tests based on the existing methods.

(We resampled per Algorithm 3.5.1 for all resampling-based methods. However, Section 4.2.2 of [121] suggests a different residual-resampling algorithm for OLS in which the resampled residuals alone are used as the resampled outcomes, such that $Y_{nw}^{(j)} := \widehat{Y}_{n'w} - Y_{n'w}$, where n' is a resampled observation. Thus, the global null is already enforced in the resampled data, and the test statistics do not require centering. Because the truth or falsehood of each null hypothesis changes the sampling distribution of the OLS coefficient estimates only by a location shift and the subset pivotality assumption described in Section 3.3 holds for OLS [121, Section 4.2.2], the difference between this algorithm and the one we used is immaterial, as confirmed by additional simulations that are not shown.)

3.8.2 Results

Figure 3.8.1 displays $\widehat{\theta}$ in samples generated under the global null (row 4, panel 1) or under varying alternatives, as well as mean limits of 95% null intervals. (For simplicity, Figure 3.8.1 does not show all scenarios, but rather excludes some smaller effect sizes. For comprehensive results, see Appendix Section C.3.1.) As expected for a resampling algorithm fulfilling Assumption 3.5.3, the null intervals appeared identical regardless of whether the data were generated under the global null. As the pairwise correlation strength between outcomes increased, the null intervals became substantially less precise. For example, with tests conducted at $\alpha = 0.05$, the mean upper limit of the null interval was more than twice as high for $\rho_{YY} = 0.60$ versus $\rho_{YY} = 0$ (i.e., 14.8 versus 5.0 rejections; see the leftmost and rightmost null intervals within each panel). Thus, with a true effect size of $\rho_{XY} = 0.05$ for all pairs (Figure 3.8.1, row 1, panel 3), the mean number of observed rejections at $\alpha = 0.05$ (i.e., 14.0) would be within the 95% null interval if the outcomes had correlation strength of $\rho_{YY} = 0.60$ (excess hits = $14.0 - 14.8 = -0.8$), but would be well outside the null interval, and thus provide stronger evidence for global association, if the outcomes were independent (excess hits = $14.0 - 5.0 = 9.0$).

Figure 3.8.2 shows the power of each global test. (Again, we show a subset of scenarios, excluding those in which all methods had nearly 100% power and excluding some intermediate correlation strengths. These scenarios differ from those in Figure 3.8.1. Comprehensive results appear in Appendix Section C.3.1.) As expected, when data were generated under the global null, all methods had approximately nominal or conservative false positive rates (Figure 3.8.2, row 3, panel 1). Our proposed global test achieved its best performance with weakly correlated or independent statistics (Figure 3.8.2, left sides of each panel) and when a moderate to high proportion of alternative hypotheses were true ($q > 0.20$). In contrast, its power suffered when few alternative hypotheses were true (e.g., $q = 0.20$; Figure 3.8.2, rows 1-2, panels 1), likely because in these scenarios, $\widehat{\theta}$ would often have been near its expectation under the global null. Simultaneously, the small number of p-values corresponding to the true alternative hypotheses may often have been quite small, improving the power of tests derived from FWER methods. Interestingly, in all scenarios we considered, [85]'s method uniformly outperformed methods other than the one we propose; it also outperformed ours with highly correlated test statistics, but not always with weakly correlated or independent statistics.

Beyond [85]'s method, the other existing methods, even the conservative naïve methods, performed comparably (within approximately 10 percentage points of power of one another for nearly all scenarios). Based on simple additional simulations (Appendix Figure C.3.2), we speculate that this somewhat counterintuitive finding arises because the methods appear to differ primarily in their degree of adjustment for those p-values that are >> 0.05, with the resampling-based methods typically yielding substantially smaller, but still "nonsignificant", adjusted values for these large p-values. In contrast, p-values near the 0.05 threshold – those that could potentially affect results of the global test – appear to receive only small and comparable adjustments across all methods. Thus, we speculate that it is rather unlikely that a sample would have all adjusted p-values above 0.05 under a naïve approach, but would have at least one p-value adjusted to below 0.05 under a resampling approach.

3.9 Discussion

This paper has characterized global evidence strength across arbitrarily correlated hypothesis tests without being restricted to the setting of high-dimensional analyses. Specifically, we proposed metrics that compare the observed number of test rejections, $\widehat{\theta}$, to its expected sampling distribution under the global null. $\widehat{\theta}$ is a simple summary measure that seems of natural interest; the proposed

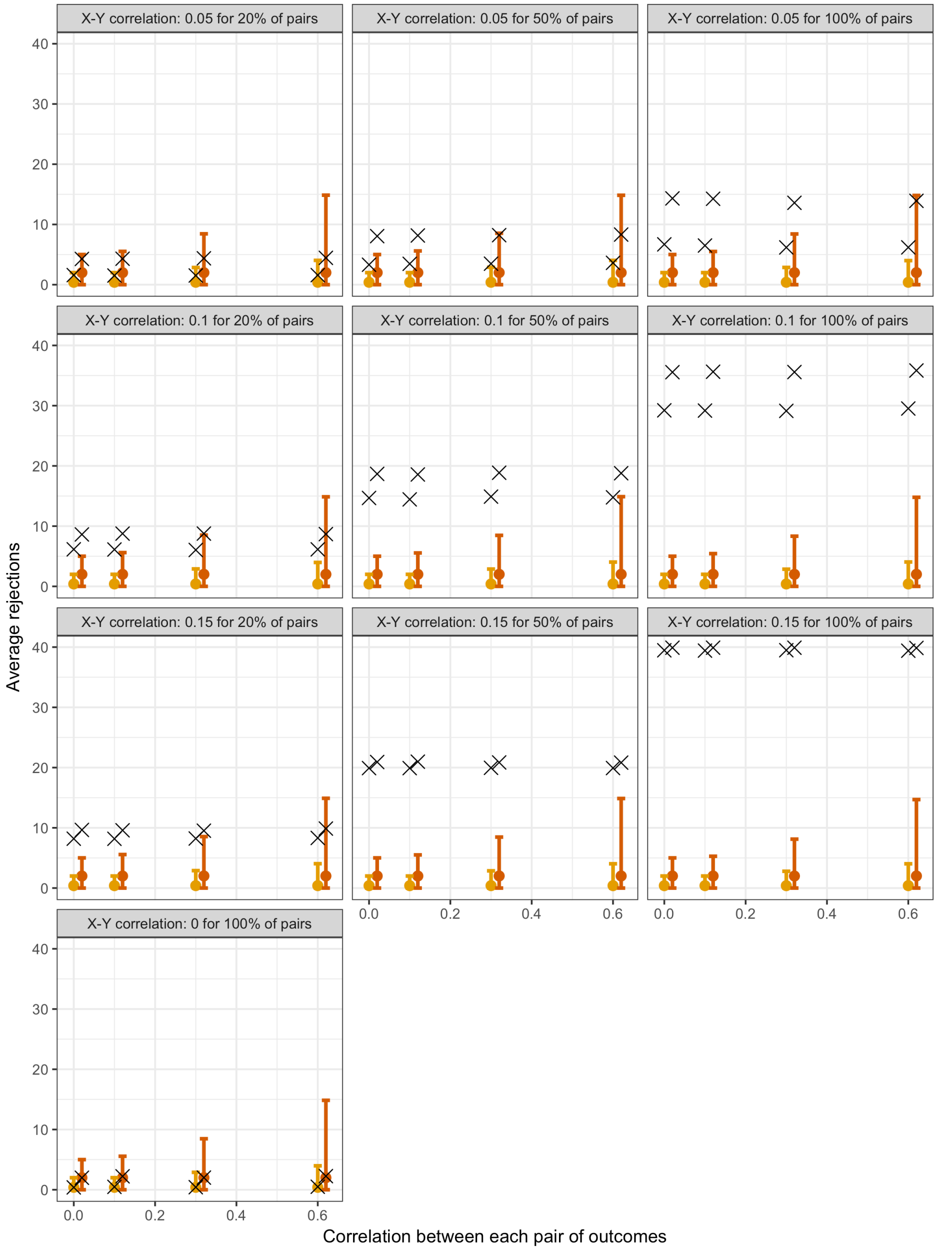

Figure 3.8.1: 95% null intervals versus mean rejections in observed datasets (×). Panels: Null and alternative data-generating mechanisms of original samples. Points and error bars: Mean $\widehat{\theta}^{(j)}$ and mean limits of null intervals with tests at $\alpha = 0.01$ (yellow) or at $\alpha = 0.05$ (red).

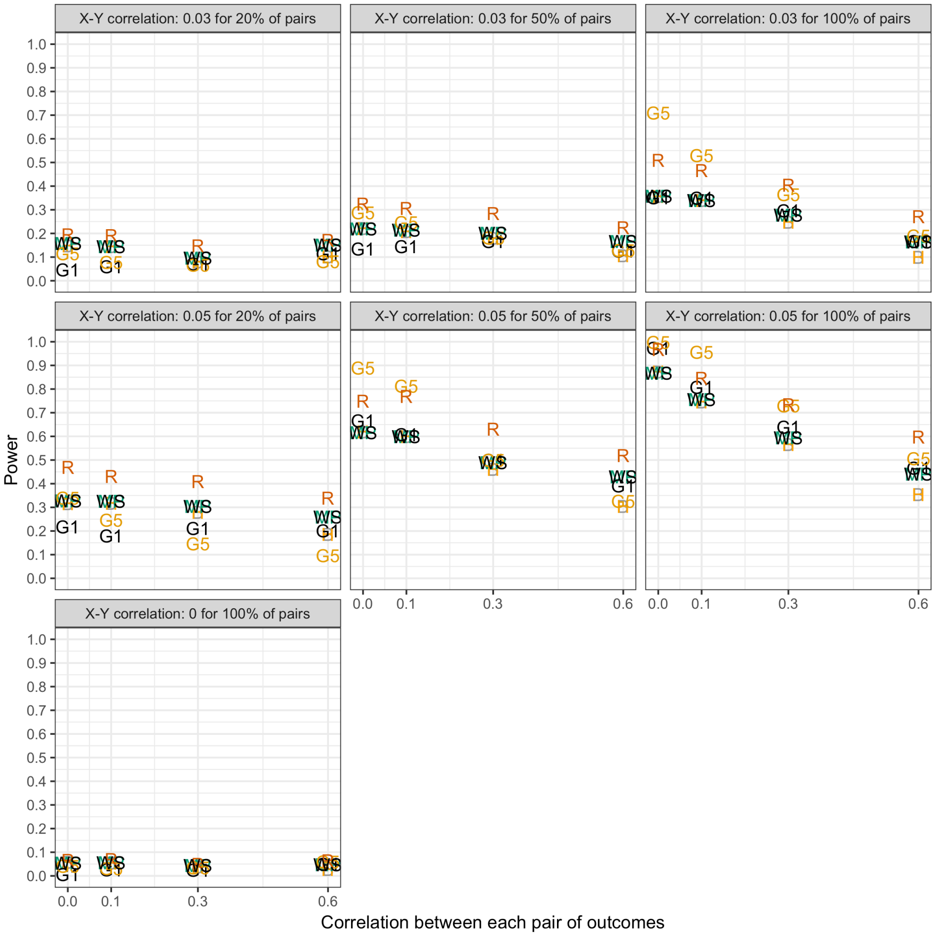

Figure 3.8.2: Power of global tests based on existing FWER-control procedures and on the number of rejections. B=Bonferroni, H=Holm, MP=minP, WS=Wstep, R=Romano, G1=number of rejections at $\alpha = 0.01$, G5=number of rejections at $\alpha = 0.05$.

metrics help to rigorously ground intuition regarding its behavior when tests are correlated. First, we proposed reporting a null interval for the number of α-level rejections expected in 95% of samples generated under the global null along with the number of excess hits observed above the upper interval limit. Second, we proposed reporting a one-sided test of the global null whose p-value represents the probability of observing at least $\widehat{\theta}$ rejections in samples generated under the global null. For OLS models, these metrics can be easily estimated via resampling using our R package, NRejections.

Existing methods that control FWER for arbitrarily correlated tests can also be used to conduct such a global test, so we conducted a simulation study assessing their relative power. To our knowledge, this is the first direct comparison of these methods as global tests, rather than as FWER-control procedures. All methods showed nominal or conservative false positive rates, as expected theoretically. Our method performed well when tests were independent or weakly correlated and when a moderate to high proportion of alternative hypotheses were true; therefore, it may be most suitable for studies in which the uncorrected p-values are relatively similar to one another, rather than for studies in which a small number of uncorrected p-values are much smaller than the others. The global test based on [85]'s method performed very well overall and, in the OLS scenarios we considered, appeared to uniformly outperform existing methods other than sometimes our own, concerning which [85]'s method was often more powerful, though ours sometimes performed better with weakly correlated or independent tests.

We speculate that the often superior power of [85]'s method as a global test, despite its additional need to strongly control FWER, reflects the loss of information inherent in dichotomizing p-values at α to compute the number of rejections. A more powerful global test might be based, for example, on departures of the observed joint

ECDF of the p-values, treated as continuous, from their CDF under the global null, as estimated via resampling methods such as those outlined in this paper and in [121]. However, even in contexts in which a global test derived from [85]'s method provides better power, the null interval and excess hits may still be of interest. More broadly, we view $\widehat{\theta}$ and the proposed metrics as useful summaries of global evidence strength that do lose some information in the process of summarization. As such, they supplement, rather than replace, reporting individual, continuous p-values with and without standard multiplicity corrections.

Our consideration of existing methods has focused on repurposing those that adjust individual p-values or critical values. Other existing methods, like our proposed metrics, do directly characterize overall evidence strength and merit some discussion. For example, global inference on regression coefficients for different outcomes can be conducted using multivariate regression [52] or [123]'s "seemingly unrelated regressions" generalization. However, these approaches only modestly improve efficiency compared to that achieved in W separate OLS models, and when the design matrix is shared across models, coefficient estimates are identical to those in OLS models [75]. Another approach to global inference is to meta-analyze the effect sizes from each analysis [26, 41, 91]. Compared to direct analysis of the raw data, meta-analysis is likely to be inefficient. Last, one could conduct global inference on a reduced number of outcomes by constructing composite measures (as in the applied example) or applying statistical dimension reduction, such as principal components analysis, though some information is lost.

When interpreting our proposed metrics, it is important to note that they characterize the sampling distribution under, specifically, the global null. Thus, rejecting the global test at $\alpha = 0.05$ indicates that there is no more than a 5% probability of observing at least $\widehat{\theta}$

rejections in samples generated under the global null. The excess hits must therefore not be misinterpreted as the number of true associations (that is, the number of null hypotheses that are false). In practice, statements to this effect can be made using procedures that strongly control FWER. By construction, these procedures ensure that for a familywise $\alpha = 0.05$, in 95% of samples generated under any configuration of null and alternative hypotheses, each rejected test will represent a true positive. Therefore, the number of rejections based on inference adjusted to strongly control FWER can be interpreted as the number of true associations, such that this statement will be incorrect in $\leq 5\%$ of samples regardless of their underlying distribution.

An additional contribution of this paper is the theoretical justification of residual resampling for OLS models in the context of multiple testing, informed by [39]'s work for a single regression model and [121]'s related algorithms. Indeed, a central challenge for resampling-based methods for multiple testing in general is the design of valid resampling procedures. The present theory supports using residual resampling under the global null for OLS in the context of our methods, of FWER control [83, 85, 121], of FDP control [122], and of corrections for "data snooping" [84]. We focused on OLS-based hypothesis tests because of their generality and ability to subsume many common tests. However, for certain other tests, such as those based on GLMs with non-identity link functions, validly resampling under the global null appears to be an open problem, although algorithms have been developed outside the multiple testing context for confidence intervals (e.g., [71]) and, under additional assumptions, for permutation hypothesis tests [81]. Other estimators, such as those using propensity score matching, pose challenges for resampling because the estimators lack certain smoothness properties; these challenges arise even without the need to enforce the global null [1]. Algorithms fulfilling Assumption 3.5.3 for such estimators could potentially use subsampling to relax some of the smoothness assumptions of with-replacement resampling [80].

Correlated test statistics can naturally arise not only when testing multiple associations between exposures and outcomes, but also when multiple hypothesis tests are used to investigate the same question, as in "data snooping" [84]. For example, investigators often fit several regression models to investigate the same association of interest, adjusting for different sets of covariates or using different subsets of the data. Situating these "researcher degrees of freedom" [95] within a formal multiple testing context [84], rather than merely reporting a single result chosen post hoc, could help reduce unnecessary false positives in the literature and may additionally foster a more balanced overall view of the evidence. Our proposed metrics provide one approach to summarizing evidence in such settings; for example, the p-value from the global test could help characterize evidence supporting a true effect in at least one of the multiple model specifications.

In summary, the number of rejections across correlated hypothesis tests can be a useful summary measure of overall evidence strength when reported with metrics such as a null interval, the number of excess hits, and a test of the global null. Reporting these metrics alongside p-values with and without standard multiplicity corrections may provide a richer view of global evidence strength than corrected inference alone.

3.10 Reproducibility

All code required to reproduce the applied example and simulation study is publicly available (https://osf.io/qj9wa/).

Appendices

A

Sensitivity Analysis for Unmeasured Confounding in Meta-Analyses

A.1 Derivation of main results

A.1.1 $\widehat{p}(q)$f

Causative case

Under the model described in the main text, we have [29]:

$$\begin{aligned}
M^t + B^* &= M^c \\
\mu^t &= E[M^c - B^*] = \mu^c - \mu_{B^*} \\
\text{Var}\left(M^t + B^*\right) &= \text{Var}\left(M^c\right) \\
V^t + \sigma^2_{B^*} &= V^c && \text{(independence)} \\
V^t &= V^c - \sigma^2_{B^*}
\end{aligned}$$

Then, $M^t = M^c - B^*$ is the difference of correlated normal random variables, so is itself normal. By Slutsky's Theorem, replace parameters with consistent estimators:

$$P\left(M^t > q\right) \approx 1 - \Phi\left(\frac{q + \mu_{B^*} - \widehat{y}^c_R}{\sqrt{\tau^2_c - \sigma^2_{B^*}}}\right), \; \tau^2_c > \sigma^2_{B^*}$$

Preventive case

The apparently preventive case is nearly identical.

A.1.2 Standard error for $\widehat{p}(q)$

We first establish a general result (Theorem A.1.1) regarding the independence of $\widehat{y}_R$ and τ^2 for many choices of estimators τ^2. Lemmas A.1.1 and A.1.2 help establish the theorem, and we provide proofs of these lemmas for completeness.

Lemma A.1.1. Let $\widehat{y}_R$ be the Dersimonian-Laird estimator of the pooled effect, where within-study variances σ^2_i are considered fixed and known:

$$\widehat{y_R} = \frac{\sum_i w_i y_i}{\sum_i w_i} = \frac{\sum_i \frac{1}{V+\sigma_i^2} y_i}{\sum_i \frac{1}{V+\sigma_i^2}}$$

Then $\widehat{y_R}$ is a complete and sufficient statistic for μ.

Proof. Fix V and consider the marginal individual and joint distributions of the y_i under the random-effects model:

$$\begin{aligned}
f_{Y_i}(y_i) &= N(y_i \mid \mu, V + \sigma_i^2) \qquad\qquad \text{(independence)} \\
f_{\mathbf{Y}}(\mathbf{y}) &= \left(\frac{1}{\sqrt{2\pi}}\right)^k \prod_{i=1}^{k} \frac{1}{\sqrt{V+\sigma_i^2}} \exp\left\{ -\frac{1}{2} \frac{(y_i-\mu)^2}{V+\sigma_i^2} \right\} \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^k \prod_{i=1}^{k} \frac{1}{\sqrt{V+\sigma_i^2}} \exp\left\{ -\frac{1}{2} \left(\frac{y_i^2}{V+\sigma_i^2} - 2\frac{y_i\mu}{V+\sigma_i^2} + \frac{\mu^2}{V+\sigma_i^2} \right) \right\} \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^k \frac{1}{\prod_i \sqrt{V+\sigma_i^2}} \exp\left\{ -\frac{1}{2} \left(\sum_{i=1}^{k} \frac{y_i^2}{V+\sigma_i^2} - 2\sum_{i=1}^{k} \frac{y_i\mu}{V+\sigma_i^2} + \sum_{i=1}^{k} \frac{\mu^2}{V+\sigma_i^2} \right) \right\} \\
&= \underbrace{\left(\frac{1}{\sqrt{2\pi}}\right)^k \frac{1}{\prod_i \sqrt{V+\sigma_i^2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{k} \frac{y_i^2}{V+\sigma_i^2} \right\}}_{h(\mathbf{y})} \cdot \underbrace{\exp\left\{ \sum_{i=1}^{k} \frac{\mu^2}{V+\sigma_i^2} \right\}}_{c(\mu)} \\
&\quad \cdot \exp\left\{ \underbrace{-2\mu}_{w(\mu)} \underbrace{\sum_{i=1}^{k} \frac{y_i}{V+\sigma_i^2}}_{t(\mathbf{y})} \right\}
\end{aligned}$$

This is a 1-parameter exponential family with support $\mathbf{y} \in \mathbb{R}^k$, and canonical parameter $w(\mu) = -2\mu,\ \mu \in \mathbb{R}$ forms an open set on the real line. Thus, a complete and sufficient statistic for μ is:

$$t(\mathbf{y}) = \sum_{i=1}^{k} \frac{y_i}{V + \sigma_i^2} = \widehat{y}_R \sum_{i=1}^{k} \frac{1}{V + \sigma_i^2}$$

Since $\widehat{y}_R$ is a function only of $t(\mathbf{y})$ and fixed quantities, it too is complete and sufficient. Since the proof holds for a fixed, arbitrary V, it must hold for all V. □

Lemma A.1.2. Let $\{a_i : i = 1, \cdots, k\}$ be an arbitrary set of positive weights that are independent of μ, and let $\widehat{\mu}_a = (\sum_i a_i y_i) / \sum_i a_i$ be a general weighted estimator of μ. Then, asymptotically, $\left(y_i - \widehat{\mu}_a\right)^2$ is ancillary for μ.

Proof. Since $y_i \sim N(\mu, V + \sigma_i^2)$:

$$\begin{aligned} \frac{(y_i - \mu)^2}{V + \sigma_i^2} &\sim \chi_1^2 \\ (y_i - \mu)^2 &\sim \text{Gamma}\left(\frac{1}{2}, \frac{2}{V + \sigma_i^2}\right) \\ \left(y_i - \widehat{\mu}_a\right)^2 &\approx \text{Gamma}\left(\frac{1}{2}, \frac{2}{V + \sigma_i^2}\right) \end{aligned}$$

where the last line follows asymptotically because $\widehat{\mu}_a$ is consistent for μ. Thus, the asymptotic distribution of $\left(y_i - \widehat{\mu}_a\right)^2$ is independent of μ, as required. □

Fact A.1.1. Let $a_i = w_{i,FE} = 1/\sigma_i^2$ be the standard weights for a fixed-effects model (i.e., assuming no heterogeneity), such that $\widehat{\mu}_a = \widehat{\mu}_{FE}$ is the fixed-effects pooled estimate. Define τ_{DL}^2 per DerSimonian and Laird [28]:

$$\tau_{DL}^2 = \max\left(0 \ , \ \frac{Q - (k-1)}{\sum_i w_{i,FE} - \frac{\sum_i w_{i,FE}^2}{\sum_i w_{i,FE}}}\right)$$

where $Q = \sum_i w_{i,FE}\left(y_i - \widehat{\mu}_{FE}\right)^2$. Then τ^2_{DL} is asymptotically independent of $\widehat{y}_R$.

Proof. τ^2_{DL} is a function only of Q (which is asymptotically ancillary for μ by Lemma A.1.2) and quantities that do not depend on μ. Therefore, τ^2_{DL} is also ancillary for μ. On the other hand, $\widehat{y}_R$ is complete and sufficient for μ. The result then follows directly from Basu's Theorem. □

Many other common τ^2 estimators (including, not exhaustively, the maximum likelihood and restricted maximum likelihood estimators and those proposed by Paule and Mandel [78], Sidik and Jonkman [93], Hartung and Makambi [44], and Hedges and Olkin [46]) retain this property with similar proofs.

Causative case

We now derive an asymptotic confidence interval for $\widehat{p}(q)$ for an apparently causative relative risk via the delta method. We assume use of the standard Dersimonian-Laird estimator, $\widehat{y}^c_R$, and an arbitrary estimator τ^2_c such that, asymptotically:

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

(Asymptotic normality is theoretically justified for the maximum likelihood and restricted maximum likelihood estimators τ^2_c and, in simulations, also appears to hold for those proposed by DerSimonian and Laird [28], Paule and Mandel [78], Sidik and Jonkman [93], and

Hedges and Olkin [46].) Apply the delta method:

$$
\begin{aligned}
h\left(x_1, x_2\right) &= \widehat{p}(q) = 1 - \Phi\left(\frac{q + \mu_{B^*} - x_1}{\sqrt{x_2 - \sigma^2_{B^*}}}\right) \\
\nabla = \begin{bmatrix} \frac{\partial h}{\partial x_1} \\ \frac{\partial h}{\partial x_2} \end{bmatrix} &= \begin{bmatrix} \frac{1}{\sqrt{x_2 - \sigma^2_{B^*}}} \cdot \varphi\left(\frac{q + \mu_{B^*} - x_1}{\sqrt{x_2 - \sigma^2_{B^*}}}\right) \\ \frac{1}{2}\left(x_2 - \sigma^2_{B^*}\right)^{-3/2} \cdot \left(q + \mu_{B^*} - x_1\right) \cdot \varphi\left(\frac{q + \mu_{B^*} - x_1}{\sqrt{x_2 - \sigma^2_{B^*}}}\right) \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{\sqrt{x_2 - \sigma^2_{B^*}}} \cdot \varphi\left(\frac{q + \mu_{B^*} - x_1}{\sqrt{x_2 - \sigma^2_{B^*}}}\right) \\ \frac{q + \mu_{B^*} - x_1}{2\left(x_2 - \sigma^2_{B^*}\right)^{3/2}} \cdot \varphi\left(\frac{q + \mu_{B^*} - x_1}{\sqrt{x_2 - \sigma^2_{B^*}}}\right) \end{bmatrix}
\end{aligned}
$$

$$
\sqrt{k}\left[h\left(\widehat{y^c_R}, \tau^2\right) - h\left(M^c, V\right)\right] \to N\left(0, \nabla'\Sigma\nabla|_{M^c,V}\right)
$$

$$
\begin{aligned}
\nabla'\Sigma\nabla &= \nabla_1\left(\nabla_1\Sigma_{11} + \nabla_2\Sigma_{21}\right) + \nabla_2\left(\nabla_1\Sigma_{12} + \nabla_2\Sigma_{22}\right) \\
&= \frac{\partial h}{\partial x_1}\left(\frac{\partial h}{\partial x_1}\mathrm{Var}\left(\widehat{y^c_R}\right) + \frac{\partial h}{\partial x_2}\mathrm{Cov}\left(\widehat{y^c_R}, \tau^2_c\right)\right) \\
&+ \frac{\partial h}{\partial x_2}\left(\frac{\partial h}{\partial x_1}\mathrm{Cov}\left(\widehat{y^c_R}, \tau^2_c\right) + \frac{\partial h}{\partial x_2}\mathrm{Var}\left(\tau^2_c\right)\right)
\end{aligned}
$$

Denote consistent estimators with hats and apply Slutsky's

Theorem:

$$\widehat{\mathrm{Var}}\left(\widehat{p}(q)\right) = \nabla' \Sigma \nabla|_{M^c, V^c}$$

$$\approx \frac{\widehat{\mathrm{Var}}\left(\widehat{y}^c_R\right)}{\tau^2_c - \sigma^2_{B^*}} \cdot \left[\varphi\left(\frac{q + \mu_{B^*} - \widehat{y}^c_R}{\sqrt{\tau^2_c - \sigma^2_{B^*}}}\right)\right]^2 +$$

$$\left(\frac{1}{\sqrt{\tau^2_c - \sigma^2_{B^*}}}\right) \frac{q + \mu_{B^*} - \widehat{y}^c_R}{2\left(\tau^2_c - \sigma^2_{B^*}\right)^{3/2}} \cdot \widehat{\mathrm{Cov}}\left(\widehat{y}^c_R, \tau^2_c\right) \cdot \left[\varphi\left(\frac{q + \mu_{B^*} - \widehat{y}^c_R}{\sqrt{\tau^2 - \sigma^2_{B^*}}}\right)\right]^2 +$$

$$\frac{\widehat{\mathrm{Var}}\left(\tau^2_c\right)\left(q + \mu_{B^*} - \widehat{y}^c_R\right)^2}{4\left(\tau^2_c - \sigma^2_{B^*}\right)^3} \cdot \left[\varphi\left(\frac{q + \mu_{B^*} - \widehat{y}^c_R}{\sqrt{\tau^2_c - \sigma^2_{B^*}}}\right)\right]^2$$

$$= \left[\frac{\widehat{\mathrm{Var}}\left(\widehat{y}^c_R\right)}{\tau^2_c - \sigma^2_{B^*}} + \frac{\left(q + \mu_{B^*} - \widehat{y}^c_R\right)\widehat{\mathrm{Cov}}\left(\widehat{y}^c_R, \tau^2_c\right)}{\left(\tau^2_c - \sigma^2_{B^*}\right)^2} + \frac{\widehat{\mathrm{Var}}\left(\tau^2_c\right)\left(q + \mu_{B^*} - \widehat{y}^c_R\right)^2}{4\left(\tau^2_c - \sigma^2_{B^*}\right)^3}\right]$$

$$\cdot \left[\varphi\left(\frac{q + \mu_{B^*} - \widehat{y}^c_R}{\sqrt{\tau^2_c - \sigma^2_{B^*}}}\right)\right]^2$$

$$\widehat{\mathrm{SE}}\left(\widehat{p}(q)\right) \approx \sqrt{\frac{\widehat{\mathrm{Var}}\left(\widehat{y}^c_R\right)}{\tau^2_c - \sigma^2_{B^*}} + \frac{\left(q + \mu_{B^*} - \widehat{y}^c_R\right)\widehat{\mathrm{Cov}}\left(\widehat{y}^c_R, \tau^2_c\right)}{\left(\tau^2_c - \sigma^2_{B^*}\right)^2} + \frac{\widehat{\mathrm{Var}}\left(\tau^2_c\right)\left(q + \mu_{B^*} - \widehat{y}^c_R\right)^2}{4\left(\tau^2_c - \sigma^2_{B^*}\right)^3}}$$

$$\cdot \varphi\left(\frac{q + \mu_{B^*} - \widehat{y}^c_R}{\sqrt{\tau^2_c - \sigma^2_{B^*}}}\right)$$

For choices of estimators τ^2_c that are asymptotically independent of $\widehat{y}^c_R$, this reduces to:

$$\widehat{\mathrm{SE}}\left(\widehat{p}(q)\right) \approx \sqrt{\frac{\widehat{Var}\left(\widehat{y}^c_R\right)}{\tau^2_c - \sigma^2_{B^*}} + \frac{\widehat{Var}\left(\tau^2_c\right)\left(q + \mu_{B^*} - \widehat{y}^c_R\right)^2}{4\left(\tau^2_c - \sigma^2_{B^*}\right)^3}} \cdot \varphi\left(\frac{q + \mu_{B^*} - \widehat{y}^c_R}{\sqrt{\tau^2_c - \sigma^2_{B^*}}}\right)$$

Preventive case

For an apparently preventive relative risk, there is simply a sign change in the numerators:

$$\widehat{\mathrm{SE}}\left(\widehat{p}(q)\right) \approx \sqrt{\frac{\widehat{\mathrm{Var}}\left(\widehat{y}_R^c\right)}{\tau_c^2 - \sigma_{B^*}^2} + \frac{\widehat{\mathrm{Var}}\left(\tau_c^2\right)\left(q - \mu_{B^*} - \widehat{y}_R^c\right)^2}{4\left(\tau_c^2 - \sigma_{B^*}^2\right)^3}} \cdot \varphi\left(\frac{q - \mu_{B^*} - \widehat{y}_R^c}{\sqrt{\tau_c^2 - \sigma_{B^*}^2}}\right)$$

A.1.3 $\widehat{T}(r, q)$

Causative case

Simply solve $\widehat{p}(q)$ for μ_{B^*}, setting the latter equal to $\log \widehat{T}(r, q)$ and setting $\sigma_{B^*}^2 = 0$:

$$\begin{aligned} r &= 1 - \Phi\left(\frac{q + \log \widehat{T}(r, q) - \widehat{y}_R^c}{\sqrt{\tau_c^2}}\right) \\ \Phi^{-1}(1 - r) &= \frac{q + \log \widehat{T}(r, q) - \widehat{y}_R^c}{\sqrt{\tau_c^2}} \\ \widehat{T}(r, q) &= \exp\left\{\Phi^{-1}(1 - r)\sqrt{\tau_c^2} - q + \widehat{y}_R^c\right\} \end{aligned}$$

Preventive case

$$\begin{aligned} r &= \Phi\left(\frac{q - \log \widehat{T}(r, q) - \widehat{y}_R^c}{\sqrt{\tau_c^2}}\right) \\ \Phi^{-1}(r) &= \frac{q - \log \widehat{T}(r, q) - \widehat{y}_R^c}{\sqrt{\tau_c^2}} \\ \widehat{T}(r, q) &= \exp\left\{q - \widehat{y}_R^c - \Phi^{-1}(r)\sqrt{\tau_c^2}\right\} \end{aligned}$$

A.1.4 Standard error for $\widehat{T}(r, q)$

Causative case

Apply the delta method:

$$
\begin{aligned}
h\left(x_1, x_2\right) &= \widehat{T}(r, q) = \exp\left\{x_2^{1/2}\left(\Phi^{-1}(1-r)\right) - q + x_1\right\} \\
\nabla = \begin{bmatrix} \frac{\partial h}{\partial x_1} \\ \frac{\partial h}{\partial x_2} \end{bmatrix} &= \begin{bmatrix} \exp\left\{x_2^{1/2}\left(\Phi^{-1}(1-r)\right) - q + x_1\right\} \\ \exp\left\{x_2^{1/2}\left(\Phi^{-1}(1-r)\right) - q + x_1\right\} \cdot \Phi^{-1}(1-r) \cdot \frac{1}{2} x_2^{-1/2} \end{bmatrix} \\
\widehat{\mathrm{Var}}\left(\widehat{T}(r, q)\right) &= \nabla' \Sigma \nabla |_{M^c, V^c} \\
&\approx \left(\exp\left\{\sqrt{(\tau_c^2)}\left(\Phi^{-1}(1-r)\right) - q + \widehat{y}_R^c\right\}\right)^2 \\
&\left(\widehat{\mathrm{Var}}\left(\widehat{y}_R^c\right) + \frac{\left(2\widehat{\mathrm{Cov}}\left(\widehat{y}_R^c, \tau_c^2\right) + \widehat{\mathrm{Var}}\left(\tau_c^2\right)\right)\left(\Phi^{-1}(1-r)\right)^2}{4\tau_c^2}\right) \\
\widehat{\mathrm{SE}}\left(\widehat{T}(r, q)\right) &= \exp\left\{\sqrt{\tau_c^2}\left(\Phi^{-1}(1-r)\right) - q + \widehat{y}_R^c\right\} \\
&\sqrt{\widehat{\mathrm{Var}}\left(\widehat{y}_R^c\right) + \frac{\left(2\widehat{\mathrm{Cov}}\left(\widehat{y}_R^c, \tau_c^2\right) + \widehat{\mathrm{Var}}\left(\tau_c^2\right)\right)\left(\Phi^{-1}(1-r)\right)^2}{4\tau_c^2}}
\end{aligned}
$$

For estimators such that $\widehat{y}_R^c$ is asymptotically independent of τ_c^2:

$$
\widehat{\mathrm{SE}}\left(\widehat{T}(r, q)\right) = \exp\left\{\sqrt{\tau_c^2}\left(\Phi^{-1}(1-r)\right) - q + \widehat{y}_R^c\right\} \sqrt{\widehat{\mathrm{Var}}\left(\widehat{y}_R^c\right) + \frac{\widehat{\mathrm{Var}}\left(\tau_c^2\right)\left(\Phi^{-1}(1-r)\right)^2}{4\tau_c^2}} \tag{A.1}
$$

Preventive case

For the apparently preventive case under asymptotic independence, there is a sign change, and the cumulative distribution function is evaluated at r instead of $1-r$:

$$\widehat{\mathrm{SE}}\left(\widehat{T}(r,q)\right) = \exp\left\{q - \widehat{y}^{c}_{R} - \sqrt{\tau^{2}_{c}}\left(\Phi^{-1}(r)\right)\right\}\sqrt{\widehat{\mathrm{Var}}\left(\widehat{y}^{c}_{R}\right) + \frac{\widehat{\mathrm{Var}}\left(\tau^{2}_{c}\right)\left(\Phi^{-1}(r)\right)^{2}}{4\tau^{2}_{c}}} \tag{A.2}$$

A.1.5 $\widehat{G}(r,q)$

Set $B^{*} = \log B^{+}$ and $\widehat{G}(r,q) = RR_{XU} = RR_{UY}$:

$$\begin{aligned} B^{*} &= \log\left(\frac{\widehat{G}(r,q)^{2}}{2\widehat{G}(r,q)-1}\right) \\ 0 &= \widehat{G}(r,q)^{2} - 2\exp\left(B^{*}\right)\widehat{G}(r,q) + \exp\left(B^{*}\right) \end{aligned}$$

Apply the quadratic formula:

$$\widehat{G}(r,q) = \exp\left(B^{*}\right) + \sqrt{\left(\exp\left(B^{*}\right)\right)^{2} - \exp\left(B^{*}\right)}$$

A.1.6 Standard error for $\widehat{G}(r,q)$

Apply the delta method to transform $\widehat{T}(r,q)$ into $\widehat{G}(r,q)$:

$$\begin{aligned} h(x) &= x + \sqrt{x^{2}-x} \\ \frac{dh}{dx} &= 1 + \frac{2x-1}{2\sqrt{x^{2}-x}} \end{aligned}$$

$$\begin{aligned} \widehat{\mathrm{Var}}\left(\widehat{G}(r,q)\right) &= \left(\frac{dh}{dx}\right)^{2}\mathrm{Var}(x)\Big|_{\widehat{T}(r,q)} \\ &= \left(1 + \frac{2\widehat{T}(r,q)-1}{2\sqrt{\widehat{T}(r,q)^{2}-\widehat{T}(r,q)}}\right)^{2}\mathrm{Var}\left(\widehat{T}(r,q)\right) \end{aligned}$$

Causative case

Plug in variance estimator (A.1):

$$\widehat{\text{SE}}\left(\widehat{G}(r,q)\right) = \left(1 + \frac{2\widehat{T}(r,q) - 1}{2\sqrt{\widehat{T}(r,q)^2 - \widehat{T}(r,q)}}\right) \cdot \exp\{\sqrt{\tau_c^2}\,(\Phi^{-1}(1-r)) - q + \widehat{y}_R^c\} \cdot \sqrt{\widehat{\text{Var}}\left(\widehat{y}_R^c\right) + \frac{\widehat{\text{Var}}\,(\tau_c^2)\,(\Phi^{-1}(1-r))^2}{4\tau_c^2}}$$

Preventive case

Plug in variance estimator (A.2):

$$\widehat{\text{SE}}\left(\widehat{G}(r,q)\right) = \left(1 + \frac{2\widehat{T}(r,q) - 1}{2\sqrt{\widehat{T}(r,q)^2 - \widehat{T}(r,q)}}\right) \cdot \exp\{\sqrt{\tau_c^2}\,(\Phi^{-1}(r)) - q - \widehat{y}_R^c\} \cdot \sqrt{\widehat{\text{Var}}\left(\widehat{y}_R^c\right) + \frac{\widehat{\text{Var}}\,(\tau_c^2)\,(\Phi^{-1}(r))^2}{4\tau_c^2}}$$

A.2 Fidelity of homogeneous-bias approximation

Table 1 in the main text provides upper or lower bounds on $\widehat{p}(q)$ that arise from assuming homogeneous bias (i.e., $\sigma^2_{B^*} = 0$). Here, we consider how closely these bounds approximate $\widehat{p}(q)$. Define $\delta = \frac{q + \mu_{B^*} - \widehat{y}_R^c}{\tau_c}$ for the apparently causative case and $\delta = \frac{q - \mu_{B^*} - \widehat{y}_R^c}{\tau_c}$ for the apparently preventive case. This quantity represents the difference between the threshold q and the bias-corrected mean estimate $\widehat{y}_R^t$ (i.e., $\widehat{y}_R^c - \mu_{B^*}$ for the causative case and $\widehat{y}_R^c + \mu_{B^*}$ for the preventive case), standardized by τ_c, the standard deviation of the confounded effect distribution. Let $w = \tau_c^2 / \sigma_{B^*}^2 > 1$, so that $1/w$ represents the proportion of variance in the confounded effects that is due to variability across studies in unmeasured confounding bias rather than to genuine effect heterogeneity. Let $\widetilde{p}(q)$ be the estimator $\widehat{p}(q)$ computed with $\sigma_{B^*}^2 = 0$. Then, for the apparently causative case, the ratio relating the homogeneous-bias approximation to the unbiased estimate is:

$$\begin{aligned}\frac{\widetilde{p}(q)}{\widehat{p}(q)} &= \frac{1 - \Phi\left(\delta\right)}{1 - \Phi\left(\delta \frac{1}{\sqrt{1-\frac{1}{w}}}\right)} \\ &= \frac{\Phi\left(-\delta\right)}{\Phi\left(-\delta \frac{1}{\sqrt{1-\frac{1}{w}}}\right)}\end{aligned}$$

The absolute difference is:

$$|\widetilde{p}(q) - \widehat{p}(q)| = \left|\Phi\left(-\delta\right) - \Phi\left(-\delta \frac{1}{\sqrt{1-\frac{1}{w}}}\right)\right|$$

The apparently preventive case is symmetrical because, whereas $\delta > 0$ for an upper bound in the causative case, $\delta < 0$ for an upper bound in the preventive case (see Table 1 in the main text), and in the above expression, $-\delta$ is also replaced with δ for the apparently preventive case (see Section 4.1 in the main text). A comparable symmetry argument holds for lower bounds. Table S1 displays $\frac{\widetilde{p}(q)}{\widehat{p}(q)}$ as a function of $|\delta|$ and w and illustrates that, on the ratio scale, the homogeneous-bias approximation holds most closely for small $|\delta|$ and large w; that is, when q is chosen to be relatively close to the bias-corrected mean estimate and when $\sigma_{B^*}^2$ is small compared to τ_c^2. Table S2 displays $|\widetilde{p}(q) - \widehat{p}(q)|$ and illustrates that the large ratios in the lower left of Table S1 correspond to cases in which $\widehat{p}(q)$ and $\widetilde{p}(q)$ are both very small, such that a large ratio corresponds to a small

Table A.2.1: Ratio of homogeneous-bias approximation with $\sigma^2_{B^*} = 0$ to the unbiased estimate, $\widehat{p}(q)$.

	$w = 1.5$	2	4	6	8	10
$\lvert\delta\rvert = 0.25$	1.21	1.11	1.04	1.02	1.02	1.01
0.5	1.60	1.29	1.09	1.06	1.04	1.03
1	3.81	2.02	1.28	1.16	1.11	1.09
1.5	14.25	3.94	1.60	1.33	1.23	1.17
2	85.53	9.73	2.17	1.60	1.40	1.30
2.5	833.38	30.52	3.19	2.01	1.65	1.48

Table A.2.2: Absolute difference of homogeneous-bias approximation with $\sigma^2_{B^*} = 0$ and the unbiased estimate, $\widehat{p}(q)$.

	$w = 1.5$	2	4	6	8	10
$\lvert\delta\rvert = 0.25$	0.07	0.04	0.01	0.01	0.01	0.01
0.5	0.12	0.07	0.03	0.02	0.01	0.01
1	0.12	0.08	0.03	0.02	0.02	0.01
1.5	0.06	0.05	0.03	0.02	0.01	0.01
2	0.02	0.02	0.01	0.01	0.01	0.01
2.5	0.01	0.01	0.00	0.00	0.00	0.00

absolute difference.

A.3 Sufficient conditions for approximate normality of bias factor

Lemma A.3.1. Let X and Y be iid $N(\mu, \sigma^2)$ with $\mu > 0$ and $\sigma^2 << \mu$. Then:

$$\log\left(e^X + e^Y - 1\right) \approx N\left(\log\left(2e^\mu - 1\right), \frac{2e^{2\mu}}{\left(2e^\mu - 1\right)^2}\,\sigma^2\right)$$

Proof. Let $h(X, Y) = \log\left(e^X + e^Y - 1\right)$. Then, apply a first-order Taylor expansion around μ, dropping higher-order terms because $\sigma^2 << \mu$:

$$\begin{aligned}
\frac{\partial h}{\partial X} &= \frac{e^X}{(e^X + e^Y - 1)} \\
\frac{\partial h}{\partial Y} &= \frac{e^Y}{(e^X + e^Y - 1)} \\
h(X,Y) &\approx \log(2e^\mu - 1) + \frac{e^\mu}{2e^\mu - 1}(X - \mu) + \frac{e^\mu}{2e^\mu - 1}(Y - \mu) \\
&= \left[\log(2e^\mu - 1) - \frac{2\mu e^\mu}{2e^\mu - 1} \right] + \frac{e^\mu}{2e^\mu - 1} X + \frac{e^\mu}{2e^\mu - 1} Y \\
E\Big[h(X,Y)\Big] &\approx \left[\log(2e^\mu - 1) - \frac{2\mu e^\mu}{2e^\mu - 1} \right] + \frac{e^\mu}{2e^\mu - 1} E[X] + \frac{e^\mu}{2e^\mu - 1} E[Y] \\
&= \log(2e^\mu - 1) \\
\text{Var}\,(h(X,Y)) &\approx \frac{2e^{2\mu}}{(2e^\mu - 1)^2}\,\sigma^2
\end{aligned}$$

The result then follows from the fact that $h(X,Y)$ is approximately a linear combination of Normal random variables. □

Fact A.3.1. Suppose $\log \text{RR}_{XU}$ and $\log \text{RR}_{UY}$ are iid $N\left(\mu_U, \sigma^2_U\right)$. Then $\log B^+$ is approximately normal.

Proof. We have $\log B^+ = \log(RR_{XU}) + \log(RR_{UY}) - \log(RR_{XU} + RR_{UY} - 1)$; the result follows immediately from invoking Lemma A.3.1 for the last term. □

A.4 Introduction to the package EValue

Here we briefly summarize the functions contained in the package EValue; details and examples are available in the standard R documentation.

The function confounded_meta computes point estimates, standard errors, and confidence interval bounds for (1) the proportion of studies with true effect sizes above q (or below q for an apparently preventive

$\widehat{y}_R^c$) as a function of the bias parameters; (2) the minimum bias factor on the relative risk scale ($\widehat{T}(r, q)$) required to reduce to less than r the proportion of studies with true effect sizes more extreme than q; and (3) the counterpart to (2) in which bias is parameterized as the minimum relative risk for both confounding associations ($\widehat{G}(r, q)$).

The function sens_table produces several types of tables (returned as dataframes) at the user's specification. The prop option yields a table showing the proportion of true effect sizes more extreme than q across a grid of bias parameters μ_{B^*} and σ_{B^*}. Alternatively, the Tmin and Gmin options yield tables showing the minimum bias factor (as in Table 2) or confounding strength required to reduce to less than r the proportion of true effects more extreme than q (across a grid of r and q).

The function sens_plot produces two types of plots. With the line option, the plot shows the bias factor on the relative risk scale (with pointwise 95% confidence band) versus the proportion of studies with true relative risks more extreme than q (as in Figure 1). The plot includes a secondary, rescaled X-axis showing the minimum strength of confounding to produce the given bias factor. With the dist option, the plot overlays the estimated densities of the confounded effects and of the true effects for a user-provided range of μ_{B^*} and scalar σ_{B^*}.

The function scrape_meta is designed to facilitate sensitivity analyses of existing meta-analyses. Given relative risks and upper bounds of 95% confidence intervals from a forest plot or summary table, the function returns a dataframe ready for meta-analysis (e.g., via the metafor package) with the log-RRs and their variances. Optionally, the user may indicate studies for which the point estimates represent odds ratios of a common outcome rather than relative risks; for such studies, the function first applies a square-root transformation to convert the odds ratio to an approximate risk ratio [112].

A.5 Code to reproduce applied example

The below code reproduces the applied example in Section 8. Extended code is also maintained at https://osf.io/2r3gm/.

```
# was run on R 3.3.3
# get data from Trock et al.’s Table 1
RRs = c(0.4, 1.8, 0.78, 0.96, 0.9, 1.4, 0.66, 0.76, 0.47,
         0.5, 2.0, 1.07, 0.66, 1.00, 0.83, 0.61, 1.0, 0.46,
         0.47, 1.16 )
UBs = c(0.8, 3.6, 1.0, 1.31, 1.3, 3.0, 0.88, 1.18, 1.33,
         1.1, 4.3, 1.47, 1.02, 1.30, 1.51, 0.97, 1.3, 0.84,
         0.74, 1.39 )

# compute point estimates and within-study variances
library(EValue)  # version 1.1.0
d = scrape_meta( type = ”RR”, est = RRs, hi = UBs )

# meta-analyze
library(metafor)  # version 2.0-0
m = rma.uni(yi=d$yi, vi=d$vyi, method=”PM”, measure=”RR”, test=”knha”)
yr = as.numeric(m$b)  # returned estimate is on log scale
vyr = as.numeric(m$vb)  # this is the KNHA-adjusted SE^2
t2 = m$tau2
vt2 = m$se.tau2^2

# reproduce Figure 1
library(ggplot2)
sens_plot( type=”line”,
                q=log(0.9),
                Bmin=log(1),
                Bmax=log(2),
                sigB=0.1,
                yr=yr,
                vyr=vyr,
                t2=t2,
                vt2=vt2,
                breaks.x1=seq(1, 2, .25) )

# now for just one choice of sensitivity parameters
# represents a single cross-section of the plot (at muB = log(1.25))
confounded_meta( q = log(.90),
             muB = log(1.25),
             sigB = 0.10,
             yr=yr,
             vyr = vyr,
             t2 = t2,
             vt2 = vt2,
             CI.level = 0.95)

# reproduce Tmin in Table 2
sens_table( meas=”Tmin”,
               q=c( log(0.70), log(0.80), log(0.90) ),
                    r=seq(0.1, 0.5, 0.1),
                    yr=yr,
                    t2=t2 )

# reproduce Gmin in Table 2
sens_table( meas=”Gmin”,
                q=c( log(0.70), log(0.80), log(0.90) ),
                r=seq(0.1, 0.5, 0.1),
                yr=yr,
                t2=t2 )
```

B

New Statistical Metrics for Multisite Replication Projects

B.1 Agreement in "statistical significance"

Suppose the original study tested the null hypothesis $H_\circ : \theta = \theta_N$, where θ is an unknown population parameter. Consider for now a single replication study, and let $\widehat{\theta}_{orig}$ and $\widehat{\theta}_{rep}$ be estimates of θ from the original and replication study, respectively. Assume that under both the null and alternative hypotheses, $\widehat{\theta}_{orig}$ and $\widehat{\theta}_{rep}$ are approximately and independently normal with a common mean but potentially different standard errors:

$$\begin{aligned} \widehat{\theta}_{orig} &\sim N\left(\theta, SE^2_{orig}\right) \amalg \widehat{\theta}_{rep} \sim N\left(\theta, SE^2_{rep}\right) \\ &\Rightarrow \widehat{\theta}_{rep} - \widehat{\theta}_{orig} \sim N\left(0, SE^2_{rep} + SE^2_{orig}\right) \end{aligned} \tag{B.1}$$

where $\amalg$ denotes statistical independence. (Critically, this setup does not allow for heterogeneity in that it assumes that the replication and original studies measure the same true effect, θ. A later section in the Appendix demonstrates the impact of this stringent assumption.) Considering first the case in which the original estimate is above the null (i.e., $\widehat{\theta}_{orig} - \theta_N > 0$), we can derive the probability of a "significant" replication estimate that is also above the null ($\widehat{\theta}_{rep} - \theta_N > 0$) given the original estimate and standard error ($\widehat{\theta}_{rep}$ and SE_{rep}). Let $c_\alpha = \Phi^{-1}(1 - \alpha)$ be the critical value of the normalized test statistic (e.g., 1.96 for $\alpha = 0.05$). Standardize $\widehat{\theta}_{rep}$ to construct the usual standard-normal test statistic and express the desired probability as:

$$
\begin{aligned}
P\left(\frac{\widehat{\theta}_{rep} - \theta_N}{\widehat{SE}_{rep}} > c_a \;\middle|\; \widehat{\theta}_{orig}, SE_{orig}\right) &= P\left(\widehat{\theta}_{rep} > c_a\widehat{SE}_{rep} + \theta_N \;\middle|\; \widehat{\theta}_{orig}, SE_{orig}\right) \\
&= P\left(\underbrace{\frac{\widehat{\theta}_{rep} - \widehat{\theta}_{orig}}{\sqrt{SE^2_{orig} + SE^2_{rep}}}}_{N(0,1)} > \frac{c_a\widehat{SE}_{rep} + \theta_N - \widehat{\theta}_{orig}}{\sqrt{SE^2_{orig} + SE^2_{rep}}}\right) \\
&\qquad \text{(re-standardize using Eq. B.1)} \\
&= 1 - \Phi\left(\frac{c_a\widehat{SE}_{rep} + \theta_N - \widehat{\theta}_{orig}}{\sqrt{SE^2_{orig} + SE^2_{rep}}}\right) \\
&\approx 1 - \Phi\left(\frac{c_a\widehat{SE}_{rep} + \theta_N - \widehat{\theta}_{orig}}{\sqrt{\widehat{SE}^2_{orig} + \widehat{SE}^2_{rep}}}\right) \qquad \text{(B.2)}
\end{aligned}
$$

where the last expression follows approximately by substituting estimated standard errors for their true counterparts. Similarly, considering the case in which the original estimate is below the null ($\widehat{\theta}_{orig} - \theta_N < 0$), the probability of a “significant” replication estimate that is also below the null is:

$$
\begin{aligned}
P\left(\frac{\widehat{\theta}_{rep} - \theta_N}{\widehat{SE}_{rep}} < -c_a \,\middle|\, \widehat{\theta}_{orig}, SE_{orig}\right) &= P\left(\widehat{\theta}_{rep} < -c_a \widehat{SE}_{rep} + \theta_N \,\middle|\, \widehat{\theta}_{orig}, SE_{orig}\right) \\
&= P\left(\underbrace{\frac{\widehat{\theta}_{rep} - \widehat{\theta}_{orig}}{\sqrt{SE^2_{orig} + SE^2_{rep}}}}_{N(0,1)} < \frac{-c_a \widehat{SE}_{rep} + \theta_N - \widehat{\theta}_{orig}}{\sqrt{SE^2_{orig} + SE^2_{rep}}}\right) \\
&\qquad \text{(re-standardize using Eq. B.1)} \\
&\approx \Phi\left(\frac{-c_a \widehat{SE}_{rep} + \theta_N - \widehat{\theta}_{orig}}{\sqrt{\widehat{SE}^2_{orig} + \widehat{SE}^2_{rep}}}\right) \qquad \text{(B.3)}
\end{aligned}
$$

When there are multiple replications (in either a many-to-one or one-to-one design), one can simply apply either Equation B.2 or B.3 to each replication study depending on the sign of the relevant original estimate.

B.2 Estimating the true effect distribution

We assume that the replication studies estimate (with statistical error) potentially different true effect sizes that follow a normal distribution. The distribution of true effects is distinct from the observed distribution of replication estimates; the latter is more variable due to uncertainty reflecting finite sample sizes in the replication studies. The proposed analyses therefore begin by using the replication studies to estimating the mean and variance of the distribution of true effects using one of two straightforward modeling approaches (though these are not exhaustive possibilities). Both approaches begin with shared assumptions. Let $\widehat{\theta}_i$ denote the effect estimate in the i^{th} replication such that $\widehat{\theta}_i = \mu + \gamma_i + \varepsilon_i$, where $\gamma_i \sim N(0, V)$ denotes deviations of site-specific true effects from the grand mean (μ) and $\varepsilon_i \sim N(0, SE_i^2)$ denotes statistical error due to finite sample sizes in the replication studies. Assume that γ_i and ε_i are independent. In other words, the true effect in replication site i is $\mu + \gamma_i$, which is normal with mean μ and variance V. Its estimate, incorporating additional error due to ε_i, is $\widehat{\theta_i}$ and is marginally normal with mean μ and variance $V + SE_i^2$.

To estimate μ and V, one option is compute an effect estimate within each site (for example, using the same model as in the original study) and then to conduct a random-effects meta-analysis on these site-level summary measures. Such analyses are already commonplace in many-to-one designs. One can then use the meta-analytic pooled estimate as $\widehat{\mu}$ and the heterogeneity estimate (usually denoted τ^2) as $\widehat{V}$. A second option, which avoids aggregating data by site prior to analysis, is to fit a mixed model to the observation-level data with independent, identically normal random intercepts and slopes by site; this is a form of "individual participant data meta-analysis" (G. B. Stewart et al., 2012). For example, suppose the original study used ordinary least squares regression to estimate the effect (β_1) of a binary experimental manipulation X on a continuous dependent variable Y with the usual specification $Y_j = \beta_0 + \beta_1 X_j + \varepsilon_j$ for subjects $j = 1, \cdots, n$ and with the error terms ε_j assumed independent and identically ("iid") normal. Then, for the replications, one possible mixed model specification is:

$$
\begin{aligned}
&Y_{ij} = \alpha_0 + \zeta_{0i} + \alpha_1 X_{ij} + \zeta_{1i} X_{ij} + \varepsilon_{ij}^* \\
&\zeta_{0i} \sim_{iid} N\left(0, \sigma_{\zeta_0}^2\right) \amalg \ \zeta_{1i} \sim_{iid} N\left(0, \sigma_{\zeta_1}^2\right) \amalg \ \varepsilon_{ij}^* \sim_{iid} N\left(0, \sigma_{\varepsilon^*}^2\right)
\end{aligned}
$$

where i indexes sites. Then, we can estimate μ (the average true effect size across all sites) using the usual maximum likelihood or restricted maximum likelihood estimate, $\widehat{\alpha_1}$. We can estimate V (the variance of the true effect sizes across all sites) with $\widehat{\sigma}^2_{\zeta_0}$. Depending on the experimental design, of course, a different mixed model specification may be warranted (for example, with additional random terms by subject) as long as it retains the normal assumption on the effect sizes across sites and yields unbiased, approximately normally distributed, and approximately independent estimates of μ and V. Specifications that do not pre-aggregate data within sites may often be more statistically efficient that the meta-analytic approach, but the meta-analysis method may sometimes provide more flexibility because it models the effect sizes rather than the dependent variable itself. Lastly, a third possible modeling approach could simply ignore site and fit the same analysis model as was used in the original study, but we do not recommend this approach because clustering within sites will likely violate statistical assumptions regarding conditionally independent residuals, such specifications preclude estimation of V, and they can lead to bias due to Simpson's Paradox (RÃŒcker & Schumacher, 2008).

B.3 Derivation of P_{orig}

Given the estimates $\widehat{\mu}$ and $\widehat{V}$ from the above development, we can derive the probability that, if the original study and replications come from the same, potentially heterogeneous distribution of true effects, the original study would estimate an effect size as extreme or more extreme than its actual estimate. As above, let $\widehat{\theta}_{orig}$ be the effect estimate in the original study and SE_{orig} its standard error. Letting $\widehat{\theta}^*$ be a random variable denoting the effect estimate in an arbitrary study with the same standard error as the original, we first consider the distribution of $\widehat{\theta}^* - \widehat{\mu}$. Assume that $\widehat{\mu} \sim N(\mu, SE^2(\widehat{\mu}))$; that is, the estimate is approximately unbiased and normal. (This holds for both the meta-analysis and the mixed model approaches above under standard assumptions). Since $\widehat{\theta}^* \sim N\left(\mu, V + SE^2_{orig}\right)$ independently of $\widehat{\mu}$, we can derive the first proposed metric as follows:

$$\begin{aligned}
\widehat{\theta}^* - \widehat{\mu} &\sim N\left(0, V + SE^2_{orig} + SE^2\left(\widehat{\mu}\right)\right) \\
\widehat{\theta}^* &\sim N\left(\widehat{\mu}, V + SE^2_{orig} + SE^2\left(\widehat{\mu}\right)\right)
\end{aligned} \tag{B.4}$$

$$P\left(|\widehat{\theta}^* - \widehat{\mu}| \geq |\widehat{\theta}_{orig} - \widehat{\mu}|\right) = P\left(\widehat{\theta}^* - \widehat{\mu} \geq |\widehat{\theta}_{orig} - \widehat{\mu}|\right) + P\left(\widehat{\theta}^* - \widehat{\mu} \leq -|\widehat{\theta}_{orig} - \widehat{\mu}|\right) \tag{B.5}$$

$$\begin{aligned}
&= P\left(\underbrace{\frac{\widehat{\theta}^* - \widehat{\mu}}{\sqrt{V + SE^2_{orig} + SE^2\left(\widehat{\mu}\right)}}}_{\sim N(0,1)} \geq \frac{|\widehat{\theta}_{orig} - \widehat{\mu}|}{\sqrt{V + SE^2_{orig} + SE^2\left(\widehat{\mu}\right)}}\right) + \\
&\quad P\left(\underbrace{\frac{\widehat{\theta}^* - \widehat{\mu}}{\sqrt{V + SE^2_{orig} + SE^2\left(\widehat{\mu}\right)}}}_{\sim N(0,1)} \leq \frac{-|\widehat{\theta}_{orig} - \widehat{\mu}|}{\sqrt{V + SE^2_{orig} + SE^2\left(\widehat{\mu}\right)}}\right) \qquad \text{(standardize)} \\
&= 1 - \Phi\left(\frac{|\widehat{\theta}_{orig} - \widehat{\mu}|}{\sqrt{V + SE^2_{orig} + SE^2\left(\widehat{\mu}\right)}}\right) + \Phi\left(\frac{-|\widehat{\theta}_{orig} - \widehat{\mu}|}{\sqrt{V + SE^2_{orig} + SE^2\left(\widehat{\mu}\right)}}\right) \\
&= 2 \times \left(1 - \Phi\left(\frac{|\widehat{\theta}_{orig} - \widehat{\mu}|}{\sqrt{V + SE^2_{orig} + SE^2\left(\widehat{\mu}\right)}}\right)\right) \qquad \text{(B.6)}
\end{aligned}$$

We arrive at the approximation in the main text (i.e., P_{orig}) by substituting estimates of SE^2_{orig} and $SE\left(\widehat{\mu}\right)$ for the true parameters.

We now show that P_{orig} subsumes Patil et al. (2016)'s prediction interval in the sense that if we assume a single replication study and no heterogeneity, and if we dichotomize P_{orig} at $\alpha = 0.05$, we mathematically recover the prediction interval. In Equation B.6, set the left-hand side equal to 0.05 (for a 95% prediction interval) and $V = 0$ (for no heterogeneity). Let $\theta^*_{0.05}$ be a value for the replication effect estimate that marks the lower or upper boundary of the 95% prediction interval. Since the prediction interval concerns a single replication study, set $\widehat{\mu} = \theta^*_{0.05}$ and $SE^2\left(\widehat{\mu}\right) = SE^2_{rep}$. Thus, we can solve for the boundary values of the prediction interval, i.e., the pair of replication estimates that are sufficiently extreme to make the probability on the left-hand side equal to 0.05:

$$
\begin{aligned}
0.05 &= 2 \times \left(1 - \Phi \left(\frac{|\widehat{\theta}_{orig} - \theta^*_{0.05}|}{\sqrt{SE^2_{orig} + SE^2_{rep}}} \right) \right) \\
|\widehat{\theta}_{orig} - \theta^*_{0.05}| &= \Phi^{-1}\left(0.975\right) \sqrt{SE^2_{orig} + SE^2_{rep}} && \text{(solve algebraically)} \\
\widehat{\theta}_{orig} - \theta^*_{0.05} &= \pm\Phi^{-1}\left(0.975\right) \sqrt{SE^2_{orig} + SE^2_{rep}} \\
\theta^*_{0.05} &= \widehat{\theta}_{orig} \pm \Phi^{-1}\left(0.975\right) \sqrt{SE^2_{orig} + SE^2_{rep}}
\end{aligned}
$$

which is exactly Patil et al. (2016)'s prediction interval.

B.4 Derivation of $P_{>0}$, $P_{>q}$, and $P_{<q^*}$

Since we consider drawing a true effect size (θ) from the distribution generating the replications, we have $\theta \sim N(\mu, V)$ by assumption. The expressions in the main text then follow immediately from properties of the normal distribution; the standard error can be derived using the delta method and is a special case of work in Mathur & VanderWeele (2017b), which focused instead on meta-analyses of observational data.

B.4.1 Impact of ignoring heterogeneity in existing metrics

We now show that ignoring heterogeneity when estimating the expected “significance agreement” always underestimates consistency when there truly is heterogeneity. We begin by generalizing Equation B.1 (which ignores heterogeneity) to accommodate heterogeneity via the same framework developed in the section “Estimating the true effect distribution”:

$$\begin{aligned}\widehat{\theta}_{orig} \sim N\left(\theta, V + SE^2_{orig}\right) \perp\!\!\!\perp \widehat{\theta}_{rep} \sim N\left(\theta, V + SE^2_{rep}\right)\\ \Rightarrow \widehat{\theta}_{rep} - \widehat{\theta}_{orig} \sim N\left(0, 2V + SE^2_{rep} + SE^2_{orig}\right)\end{aligned}$$

For an original estimate above the null, we can compute the probability of “significance agreement” allowing for heterogeneity as:

$$\begin{aligned}P\left(\frac{\widehat{\theta}_{rep} - \theta_N}{\widehat{SE}_{rep}} > c_a \,\middle|\, \widehat{\theta}_{orig}, SE_{orig}\right) &= P\left(\widehat{\theta}_{rep} > c_a\widehat{SE}_{rep} + \theta_N \,\middle|\, \widehat{\theta}_{orig}, SE_{orig}\right)\\ &= P\left(\underbrace{\frac{\widehat{\theta}_{rep} - \widehat{\theta}_{orig}}{\sqrt{2V + SE^2_{orig} + SE^2_{rep}}}}_{N(0,1)} > \frac{c_a\widehat{SE}_{rep} + \theta_N - \widehat{\theta}_{orig}}{\sqrt{2V + SE^2_{orig} + SE^2_{rep}}}\right)\\ &\approx 1 - \Phi\left(\frac{c_a\widehat{SE}_{rep} + \theta_N - \widehat{\theta}_{orig}}{\sqrt{2V + \widehat{SE}^2_{orig} + \widehat{SE}^2_{rep}}}\right)\end{aligned}$$

The only difference between this expression and Equation B.2 (which had assumed no heterogeneity) is the $2V$ term in the denominator. Since the presence of heterogeneity implies that $2V > 0$, this probability is always larger than the probability in Equation B.2, proving our claim. The case in which the original estimate is below the null is nearly identical, so is omitted.

Similarly, when there is heterogeneity, the prediction interval is too narrow (and thus, it is less likely that the replication will fall inside the prediction interval). Using the previous result showing equivalence of P_{orig} with the prediction interval when there is no heterogeneity, we can set $V = 0$ in Equation B.6 to yield the p-value counterpart to the prediction interval. Since Equation B.6 is strictly increasing in V, constraining $V = 0$ in this expression yields a lower p-value than allowing $V > 0$. Thus, if there is heterogeneity, the p-value counterpart to the prediction interval is an underestimate. By the duality of p-values and intervals, the prediction interval is therefore too narrow when $V > 0$.

B.5 Methods for choosing an effect size threshold

Much existing work, spanning a variety of disciplinary perspectives, has discussed how to choose thresholds for scientifically meaningful effect sizes. [25] provides an excellent review and examples of numerous methods in the context of health outcomes. In particular, they discuss a variety of "anchoring-based" methods in which an effect size threshold is chosen by relating the outcome measure to external benchmarks bearing immediate scientific or policy relevance. Within psychology, this approach may be particularly relevant for applied or interventional studies; for example, when investigating effects of educational interventions, a minimum effect size threshold could be determined in relation to differences in the outcome (academic achievement) between naturally-occurring subject groups (such as children attending low- versus high-performing schools or children of different ages) [47]. Numerous other types of external "anchoring" criteria have also been used in the health outcomes literature [25].

When the aggregate public impact of an outcome (such as juvenile delinquency) is the primary concern, investigators could draw upon the extensive literature on cost-effectiveness decision rules in selecting an effect size threshold. For example, much existing work has discussed or empirically quantified the cost threshold at which societies or individuals are willing to pay for a specific improvement in physical or mental health, such as an addition of one quality-adjusted life-year (e.g., [12, 35]). Such findings could be used to "convert" hypothetical statistical effect sizes for a given outcome to a concrete financial scale, such as dollars. A minimum effect size threshold could then be defined in relation to the utility, expressed in dollars, of the intervention or exposure of interest.

In contrast, in disciplines such as clinical psychology, the original study may investigate an effect in which individuals' subjective experience of distress or pain is the primary concern (instead of, or in addition to, aggregate public impact). In this case, it may be useful to set the threshold as the minimum effect size that is subjectively perceptible [51, 59, 73, 82]. A systematic review considered 62 studies that attempted to estimate such thresholds for a wide variety of health outcomes, for example by relating patients' subjective self-assessments to objective measurements of health condition severity [73]. This review found that $d = 0.50$ was a surprisingly consistent minimally detectable effect size for health outcomes, perhaps reflecting fundamental mechanisms of human sensory discrimination or constraints on categorical discrimination due to working memory capacity. For ease of comparison to other statistical measures of effect size, the threshold $d = 0.50$ is approximately equivalent (under some distributional assumptions) to an odds ratio of 2.5 or to a risk ratio of 1.6 [17, 112]. However, it is important to note that an intervention that has only small effects on the individual level, even ones that are not subjectively perceptible, may still have very substantial impacts on a population level; thus, as described above, much lower thresholds might often be considered.

While the above considerations may work well for applied or interventional psychology, many replication efforts to date have focused on classic experimental psychology, conducted using stylized tasks (such as a Stroop task or [69]'s hypothetical hiring task) in order to examine basic mechanisms of, for example, cognition or perception. Although some of the above considerations are harder to apply in these classic experimental contexts, external benchmarks could still be determined using effect sizes on similar experimental tasks, preferably those estimated by meta-analyses of existing literature. For example, a meta-analysis of the enormous literature on intergroup contact and prejudice estimated a pooled effect size of $r = -0.21$ among all study designs and $r = -0.33$ among experimental studies [79]. We might treat experimental intergroup-contact interventions as a "gold standard" representing the effect sizes on prejudice that are achievable through purposefully designed interventions. In contrast, the proposed moral credentialing effect is not a designed intervention on prejudice but rather a specific, potentially more subtle, cognitive mechanism of prejudice. Thus, to select an effect size threshold for moral credentialing, we might somewhat reduce the magnitude of the gold-standard interventions to, for example, $|r| = 0.20$ or $|r| = 0.10$. (Additionally, the latter threshold is often considered a standard benchmark for a "small" effect size [21].)

B.6 Software

The R package "Replicate" contains the following functions; details are available in the standard R documentation.

- prob_signif_agree computes the theoretical probability that a given replication study would agree in "statistical significance" and effect direction with the original study, if the true effect is indeed the same in the two studies.

- pred_int computes prediction interval limits and indicators for whether each replication estimate is within its corresponding prediction interval.

- p_orig computes P_{orig}, i.e., the probability of observing an original estimate as extreme as that actually observed (compared to the replication studies) if the original is indeed consistent with the estimated distribution of the replication studies.

- stronger_than estimates $P_{>q}$ or $P_{<q^*}$, i.e., the probability of a true effect above or below a user-specified threshold of scientific importance using estimates of the true effect distribution (based on the replications).

C

New Metrics for Multiple Testing with Correlated Outcomes

C.1 Exact variance under global null

Let p_w° be the p-value in the w^{th} test under the global null, treated as a random variable. Then we have:

$$\begin{aligned}
\mathrm{Var}\left(\widehat{\theta}^{\circ}\right) &= \mathrm{Var}\left(\sum_{w=1}^{W} 1\{p_w^{\circ} < \alpha\}\right) \\
&= \sum_{w=1}^{W} \mathrm{Var}\left(1\{p_w^{\circ} < \alpha\}\right) + 2 \sum_{1 \leq i < j \leq W} \mathrm{Cov}\left(1\{p_i^{\circ} < \alpha\}, 1\{p_j^{\circ} < \alpha\}\right) \\
&= W\alpha\left(1 - \alpha\right) + 2 \sum_{1 \leq i < j \leq W} E[1\{p_i^{\circ} < \alpha, p_j^{\circ} < \alpha\}] - E[1\{p_i^{\circ} < \alpha\}]E[1\{p_j^{\circ} < \alpha\}] \\
&= W\alpha\left(1 - \alpha\right) + 2 \sum_{1 \leq i < j \leq W} \Big[\underbrace{P\left(p_i^{\circ} < \alpha, p_j^{\circ} < \alpha\right)}_{=\alpha^2 \text{ under independence}} - \alpha^2 \Big]
\end{aligned}$$

C.2 Applied example

Table C.2.1: Demographic and childhood family characteristics of $2,697$ analyzed subjects. ^{a}: By subject's adolescence, subject's family had ever been on welfare. ^{b}: Ranged from 1 ("a lot better off" than others) to 7 ("a lot worse off" than others). ^{c}: By age 16, subject had ever lived with an alcoholic.

Characteristic	Mean (SD) or %
Age	46.89 (12.35)
Female	53.7%
Race	
White	93.3%
Black	3.6%
Other	3.2%
Born in US	95.8%
Mother born in US	90.5%
Father born in US	90.2%
Lived with biological parents	81.1%
Number of siblings	2.92 (1.57)
Highest parental education	
Less than high school	25.8%
High school	36.0%
Some college	15.8%
College degree or more	22.5%
Childhood welfare^a	5.6%
Subjective SES^b	4.07 (1.29)
Residential area	
Rural	23.1%
Small town	25.6%
Medium town	12.1%
Suburbs	16.8%

Continued on next page

Table C.2.1 (Continued)

Characteristic	Mean (SD) or %
City	18.3%
Moved around	4.1%
Residentially stable	74.1%
Mother smoked	32.6%
Father smoked	62.0%
Lived with alcoholics[c]	20.9%
Importance of religion	
Very important	43.5%
Somewhat important	35.7%
Not very important	16.0%
Not at all important	4.7%

C.3 Extended simulation results

C.3.1 Applied example

The following figures show all results presented in the main text as well as additional scenarios.

C.3.2 Comparison of p-values adjusted by existing methods

We performed a rudimentary visual comparison of p-value adjustments produced by one naïve method (Holm) and one resampling-based method (Wstep). We generated a single dataset as in the simulation study with 1 covariate, 100 outcomes, $N = 1,000$, $\rho_{XY} = 0.08$ for all outcomes, and $\rho_{YY} = 0.25$. We chose these parameters to yield a large number of adjusted p-values < 0.05 for illustrative purposes. Figure C.3.3 plots the 100 p-values adjusted using the Holm and Wstep methods (obtained by resampling as in the applied

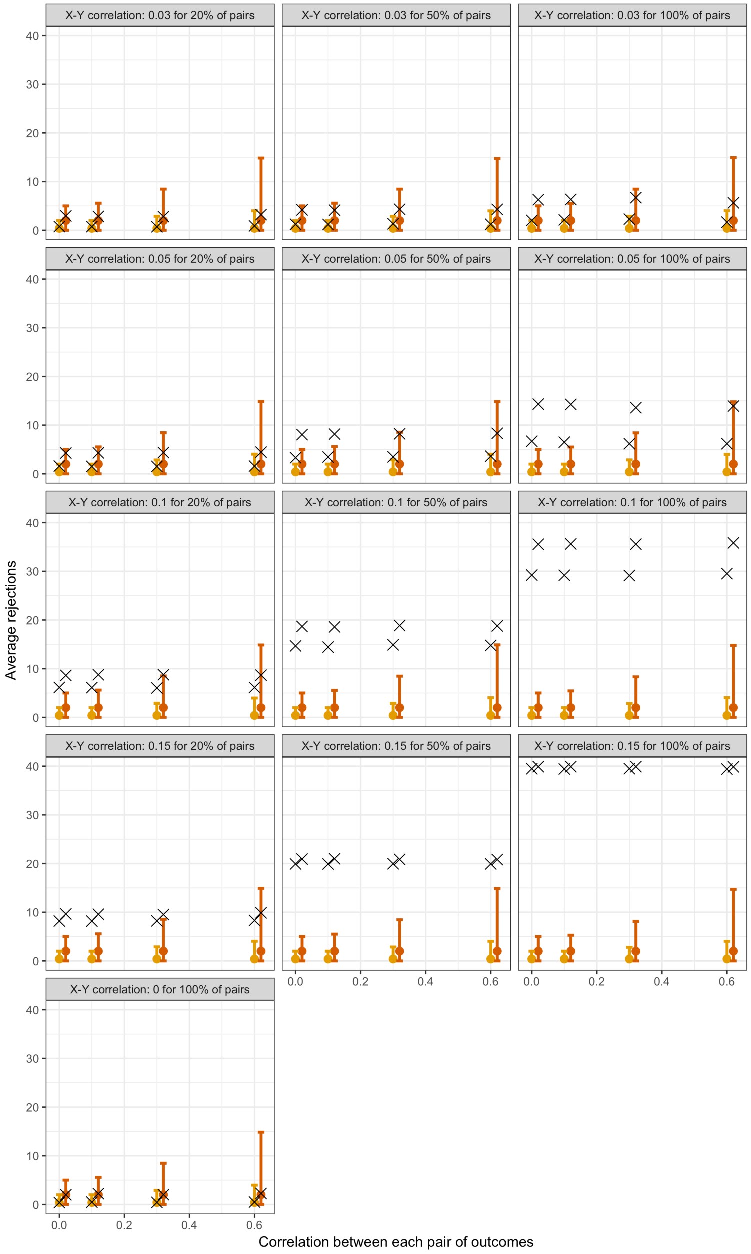

Figure C.3.1: 95% null intervals versus mean rejections in observed datasets (×). Panels: Null and alternative data-generating mechanisms of original samples. Points and error bars: Mean $\widehat{\theta}^{(j)}$ and mean limits of null intervals with tests at $\alpha = 0.01$ (yellow) or at $\alpha = 0.05$ (red).

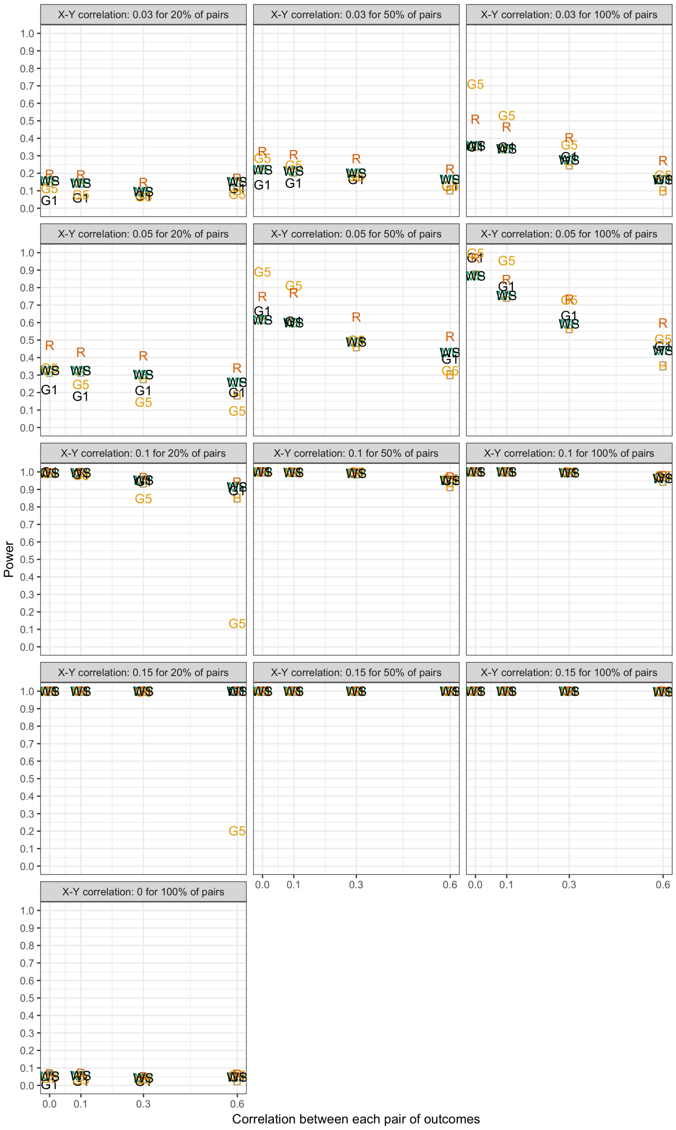

Figure C.3.2: Power of global tests based on existing FWER-control procedures and on the number of rejections. B=Bonferroni, H=Holm, MP=minP, WS=Wstep, R=Romano, G1=number of rejections at $\alpha = 0.01$, G5=number of rejections at $\alpha = 0.05$.

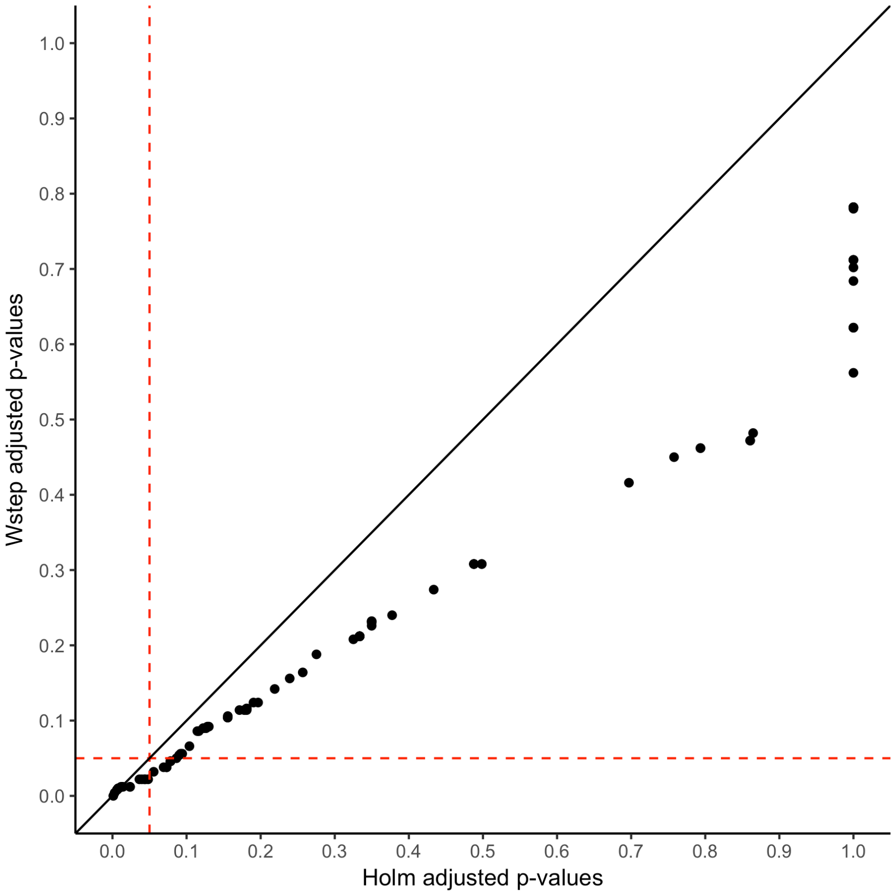

Figure C.3.3: p-values in a single simulated dataset adjusted by the Holm method versus the Wstep method. Red dashed lines: $\alpha = 0.05$ threshold.

example with $B = 500$ resamples) and suggests that in this simple simulation, the methods differ little in their adjustments to p-values near $\alpha = 0.05$; rather, the differences appear to emerge primarily for $p >> 0.05$. We obtained qualitatively similar results when comparing other pairs of existing methods (not shown).

C.4 Introduction to the package NRejections

Here we briefly describe the R package NRejections; note that additional functions, details, and additional examples are available in the standard R documentation. For OLS models as described in Section 3.5.2, the null interval, excess hits, and global test can be conducted by calling a single wrapper function, corr_tests. This function first fits the W models in the original dataset, adjusting for any user-specified covariates. Then, resamples are generated via Algorithm 3.5.1 and used to estimate and return our proposed metrics, along with estimates and inference from the original sample. Optionally, the global test can additionally be conducted using any combination of methods in Table 3.3.1. Below is a minimal example.

```
# this was run on R version 3.3.3
# and NRejections version 1.0.0

library(NRejections)

# simulate data with 40 outcomes and 1 covariate of interest,
#  similarly to simulation study
# 80% of the 40 associations are non-null (correlation strength of 0.08);
#  and the others are null
cor = make_corr_mat( nX = 1,
  nY = 40,
  rho.XX = 0,
  rho.YY = 0.15,
  rho.XY = 0.08,
  prop.corr = .8 )

d = sim_data( n = 1000, cor = cor )
```

```
# may take 5-10 min to run on 8-core personal computer
res = corr_tests( d,
                  X = ”X1”,
                  Ys = names(d)[ grep( ”Y”, names(d) ) ],
                  B = 1000,
                  method = ”nreject” )

# main results
res$null.int
res$excess.hits
res$global.test
```

As described in the Discussion, Algorithm 3.5.1 is more broadly applicable to multiple-testing procedures outside the scope of this paper. For these general applications, the user could first obtain residuals and point estimates from the original dataset using the function dataset_result and pass these to resid_resample, which returns matrices of *p*-values and test statistics from the resamples. See ?resid_resample for examples.

References

[1] Alberto Abadie and Guido W Imbens. On the failure of the bootstrap for matching estimators. Econometrica, 76(6): 1537–1557, 2008.

[2] VK Alogna, Matthew K Attaya, Philip Aucoin, Š Bahník, Stacy Birch, Angela R Birt, Brian H Bornstein, Samantha Bouwmeester, Maria A Brandimonte, Charity Brown, et al. Registered Replication Report: Schooler and Engstler-Schooler (1990). Perspectives on Psychological Science, 9(5):556–578, 2014.

[3] Samantha F Anderson and Scott E Maxwell. There's more than one way to conduct a replication study: Beyond statistical significance. Psychological Methods, 21(1):1, 2016.

[4] Isaiah Andrews and Maximilian Kasy. Identification of and correction for publication bias. Technical report, National Bureau of Economic Research, 2017.

[5] Association for Psychological Science. Ongoing replication projects. https://www.psychologicalscience.org/publications/replication/ongoing-projects., 2018. Accessed: 2018-04-18.

[6] Dagfinn Aune, Doris SM Chan, Rosa Lau, Rui Vieira, Darren C Greenwood, Ellen Kampman, and Teresa Norat. Dietary fibre, whole grains, and risk of colorectal cancer: systematic review and dose-response meta-analysis of prospective studies. BMJ, 343:d6617, 2011.

[7] Roy F Baumeister and Kathleen D Vohs. Misguided effort with elusive implications. Perspectives on Psychological Science, 11 (4):574–575, 2016.

[8] Peter J Bickel and David A Freedman. Some asymptotic theory for the bootstrap. The Annals of Statistics, pages 1196–1217, 1981.

[9] Richard E Blakesley, Sati Mazumdar, Mary Amanda Dew, Patricia R Houck, Gong Tang, Charles F Reynolds III, and Meryl A Butters. Comparisons of methods for multiple hypothesis testing in neuropsychological research. Neuropsychology, 23(2):255, 2009.

[10] Michael Borenstein, Larry V Hedges, Julian Higgins, and Hannah R Rothstein. Introduction to meta-analysis. Wiley Online Library, 2009.

[11] Samantha Bouwmeester, Peter PJL Verkoeijen, Balazs Aczel, Fernando Barbosa, Laurent Bègue, Pablo Brañas-Garza, Thorsten GH Chmura, Gert Cornelissen, Felix S Døssing, Antonio M Espín, et al. Registered Replication Report: Rand, Greene, and Nowak (2012). Perspectives on Psychological Science, 12(3):527–542, 2017.

[12] R Scott Braithwaite, David O Meltzer, Joseph T King Jr, Douglas Leslie, and Mark S Roberts. What does the value of modern medicine say about the $50,000 per quality-adjusted life-year decision rule? Medical Care, 46(4):349–356, 2008.

[13] Orville Gilbert Brim, Carol D Ryff, and Ronald C Kessler. The MIDUS National Survey: an overview. University of Chicago Press, 2004.

[14] Colin F Camerer, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, et al. Evaluating replicability of laboratory experiments in economics. Science, 351(6280):1433–1436, 2016.

[15] Ying Chen, Laura D Kubzansky, and Tyler J VanderWeele. Parental warmth and flourishing in mid-life. Under review, page XXX, 2018.

[16] Irene Cheung, Lorne Campbell, Etienne P LeBel, RA Ackerman, B Aykutoğlu, Š Bahník, JD Bowen, CA Bredow, C Bromberg, PA Capraricllo, et al. Registered Replication Report: Study 1 from Finkel, Rusbult, Kumashiro, & Hannon (2002). Perspectives on Psychological Science, 11(5):750–764, 2016.

[17] Susan Chinn. A simple method for converting an odds ratio to effect size for use in meta-analysis. Statistics in Medicine, 19 (22):3127–3131, 2000.

[18] Mei Chung, Jiantao Ma, Kamal Patel, Samantha Berger, Joseph Lau, and Alice H Lichtenstein. Fructose, high-fructose corn syrup, sucrose, and nonalcoholic fatty liver disease or indexes of liver health: a systematic review and meta-analysis. The American Journal of Clinical Nutrition, 100(3):833–849, 2014.

[19] Sandy Clarke, Peter Hall, et al. Robustness of multiple testing procedures against dependence. The Annals of Statistics, 37(1): 332–358, 2009.

[20] Jacob Cohen. Statistical power analysis for the behavioral sciences, 1977.

[21] Jacob Cohen. A power primer. Psychological Bulletin, 112(1): 155, 1992.

[22] Jerome Cornfield, William Haenszel, E Cuyler Hammond, Abraham M Lilienfeld, Michael B Shimkin, and Ernst L Wynder. Smoking and lung cancer: recent evidence and a discussion of some questions. Journal of the National Cancer Institute, 22:173–203, 1959.

[23] Cova, F., et al. Estimating the reproducibility of experimental philosophy. https://osf.io/7f5hs/., 2018. Accessed: 2018-04-18.

[24] Christian S Crandall and Jeffrey W Sherman. On the scientific superiority of conceptual replications for scientific progress. Journal of Experimental Social Psychology, 66:93–99, 2016.

[25] Ross D Crosby, Ronette L Kolotkin, and G Rhys Williams. Defining clinically meaningful change in health-related quality of life. Journal of clinical epidemiology, 56(5):395–407, 2003.

[26] Geoff Cumming. The new statistics: Why and how. Psychological Science, 25(1):7–29, 2014.

[27] A DasGupta. Asymptotic theory of statistics and probability, 2008.

[28] Rebecca DerSimonian and Nan Laird. Meta-analysis in clinical trials. Controlled Clinical Trials, 7(3):177–188, 1986.

[29] Peng Ding and Tyler J VanderWeele. Sensitivity analysis without assumptions. Epidemiology, 27(3):368, 2016.

[30] Olive Jean Dunn. Multiple comparisons among means. Journal of the American Statistical Association, 56(293):52–64, 1961.

[31] Charles R Ebersole, Olivia E Atherton, Aimee L Belanger, Hayley M Skulborstad, Jill M Allen, Jonathan B Banks, Erica Baranski, Michael J Bernstein, Diane BV Bonfiglio, Leanne Boucher, et al. Many Labs 3: Evaluating participant pool quality across the academic semester via replication. Journal of Experimental Social Psychology, 67:68–82, 2016.

[32] Ebersole, C.R., et al. Many Labs 5: Can conducting formal peer review in advance improve reproducibility? https://osf.io/7a6rd/., 2018. Accessed: 2018-04-18.

[33] A Eerland, Andrew M Sherrill, Joseph P Magliano, Rolf A Zwaan, JD Arnal, Philip Aucoin, SA Berger, AR Birt, Nicole Capezza, Marianna Carlucci, et al. Registered replication report: Hart & albarracín (2011). Perspectives on Psychological Science, 11(1):158–171, 2016.

[34] Matthias Egger, Martin Schneider, and G Davey Smith. Spurious precision? Meta-analysis of observational studies. BMJ, 316(7125):140, 1998.

[35] Hans-Georg Eichler, Sheldon X Kong, William C Gerth, Panagiotis Mavros, and Bengt Jönsson. Use of cost-effectiveness analysis in health-care resource allocation decision-making: how are cost-effectiveness thresholds expected to emerge? Value in Health, 7(5):518–528, 2004.

[36] Alexander Etz and Joachim Vandekerckhove. A Bayesian perspective on the Reproducibility Project: Psychology. PLoS One, 11(2):e0149794, 2016.

[37] W Dana Flanders and Mum J Khoury. Indirect assessment of confounding: graphic description and limits on effect of adjusting for covariates. Epidemiology, 1(3):239–246, 1990.

[38] Gregory Francis. The psychology of replication and replication in psychology. Perspectives on Psychological Science, 7(6): 585–594, 2012.

[39] David A Freedman. Bootstrapping regression models. The Annals of Statistics, pages 1218–1228, 1981.

[40] Chloé Friguet, Maela Kloareg, and David Causeur. A factor model approach to multiple testing under dependence. Journal of the American Statistical Association, 104(488):1406–1415, 2009.

[41] Jin X Goh, Judith A Hall, and Robert Rosenthal. Mini meta-analysis of your own studies: Some arguments on why and a primer on how. Social and Personality Psychology Compass, 10(10):535–549, 2016.

[42] Martin S Hagger, Nikos LD Chatzisarantis, Hugo Alberts, Calvin O Anggono, Cedric Batailler, Angela R Birt, Ralf Brand, Mark J Brandt, Gene Brewer, Sabrina Bruyneel, et al. A multilab preregistered replication of the ego-depletion effect. Perspectives on Psychological Science, 11(4):546–573, 2016.

[43] Peter Hall and Susan R Wilson. Two guidelines for bootstrap hypothesis testing. Biometrics, pages 757–762, 1991.

[44] J Hartung and KH Makambi. Positive estimation of the between-study variance in meta-analysis. South African Statistical Journal, 36(1):55–76, 2002.

[45] Joachim Hartung and Guido Knapp. On tests of the overall treatment effect in meta-analysis with normally distributed responses. Statistics in Medicine, 20(12):1771–1782, 2001.

[46] Lawrence Hedges and Ingram Olkin. Statistical methods for meta-analysis. Academic Press, 1985.

[47] Carolyn J Hill, Howard S Bloom, Alison Rebeck Black, and Mark W Lipsey. Empirical benchmarks for interpreting effect sizes in research. Child Development Perspectives, 2(3):172–177, 2008.

[48] Sture Holm. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, pages 65–70, 1979.

[49] Guildo W Imbens. Sensitivity to exogeneity assumptions in program evaluation. American Economic Review, 93(2): 126–132, 2003.

[50] Joanna IntHout, John PA Ioannidis, and George F Borm. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Medical Research Methodology, 14(1):1, 2014.

[51] Roman Jaeschke, Joel Singer, and Gordon H Guyatt. Measurement of health status: ascertaining the minimal clinically important difference. Controlled Clinical Trials, 10(4): 407–415, 1989.

[52] Richard A Johnson and Dean Wichern. Multivariate analysis. Wiley Online Library, 2002.

[53] D.A. Kenny and C.M. Judd. The unappreciated heterogeneity of effect sizes: implications for power, planning of research, and replication. https://osf.io/s9gfm/., 2018. Accessed: 2018-04-18.

[54] Corey LM Keyes. The mental health continuum: From languishing to flourishing in life. Journal of Health and Social Behavior, pages 207–222, 2002.

[55] Corey LM Keyes. Promoting and protecting mental health as flourishing: A complementary strategy for improving national mental health. American Psychologist, 62(2):95, 2007.

[56] Corey LM Keyes and Eduardo J Simoes. To flourish or not: Positive mental health and all-cause mortality. American Journal of Public Health, 102(11):2164–2172, 2012.

[57] Richard A Klein, Kate A Ratliff, Michelangelo Vianello, Reginald B Adams Jr, Štěpán Bahník, Michael J Bernstein, Konrad Bocian, Mark J Brandt, Beach Brooks, Claudia Chloe Brumbaugh, et al. Investigating variation in replicability. Social Psychology, 2014.

[58] Klein, R.A., et al. Many Labs 2: investigating variation in replicability across sample and setting. https://osf.io/8cd4r/., 2018. Accessed: 2018-04-18.

[59] D Lakens, M Scheel, and PM Isager. Equivalence testing for psychological research: a tutorial. 2017. Preprint retrieved from https://psyarxiv.com/v3zkt/.

[60] Daniël Lakens. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Frontiers in Psychology, 4, 2013.

[61] Daniël Lakens. Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. Social Psychological and Personality Science, 8(4):355–362, 2017.

[62] Jeffrey T Leek and John D Storey. A general framework for multiple testing dependence. Proceedings of the National Academy of Sciences, 105(48):18718–18723, 2008.

[63] Danyu Y Lin, Bruce M Psaty, and Richard A Kronmal. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. Biometrics, pages 948–963, 1998.

[64] John G Lynch, Eric T Bradlow, Joel C Huber, and Donald R Lehmann. Reflections on the Replication Corner: In praise of conceptual replications. International Journal of Research in Marketing, 32(4):333–342, 2015.

[65] MB Mathur and TJ VanderWeele. Sensitivity analysis for unmeasured confounding in meta-analyses. 2017. Preprint retrieved from https://osf.io/wdmht/.

[66] MB Mathur and TJ VanderWeele. Package "Replicate", 2017. https://cran.r-project.org/web/packages/Replicate/Replicate.pdf.

[67] Lawrence C. McCandless. Meta-analysis of observational studies with unmeasured confounders. The International Journal of Biostatistics, 8(2):368, 2012.

[68] Benoît Monin. Be careful what you wish for: Commentary on Ebersole et al. (2016). Journal of Experimental Social Psychology, 67:95–96, 2016.

[69] Benoît Monin and Dale T Miller. Moral credentials and the expression of prejudice. Journal of Personality and Social Psychology, 81(1):33, 2001.

[70] Benoît Monin, Daniel M Oppenheimer, Melissa J Ferguson, Travis J Carter, Ran R Hassin, Richard J Crisp, Eleanor Miles, Shenel Husnu, Norbert Schwarz, Fritz Strack, et al. Commentaries and rejoinder on Klein et al.(2014). Social Psychology, 2014.

[71] Lawrence H Moulton and Scott L Zeger. Bootstrapping generalized linear models. Computational statistics & data analysis, 11:53–63, 1991.

[72] Elizabeth Mullen and Benoît Monin. Consistency versus licensing effects of past moral behavior. Annual Review of Psychology, 67, 2016.

[73] Geoffrey R Norman, Jeff A Sloan, and Kathleen W Wyrwich. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Medical Care, 41(5):582–592, 2003.

[74] Brian A Nosek, Jeffrey R Spies, and Matt Motyl. Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. Perspectives on Psychological Science, 7(6):615–631, 2012.

[75] Rosa Oliveira and Armando Teixeira-Pinto. Analyzing multiple outcomes: Is it really worth the use of multivariate linear regression? Journal of Biometrics & Biostatistics, 6, 2015.

[76] Open Science Collaboration. Estimating the reproducibility of psychological science. Science, 349(6251):aac4716, 2015.

[77] Prasad Patil, Roger D Peng, and Jeffrey T Leek. What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. Perspectives on Psychological Science, 11(4):539–544, 2016.

[78] Robert C Paule and John Mandel. Consensus values and weighting factors. Journal of Research of the National Bureau of Standards, 87(5):377–385, 1982.

[79] Thomas F Pettigrew and Linda R Tropp. How does intergroup contact reduce prejudice? meta-analytic tests of three mediators. European Journal of Social Psychology, 38(6): 922–934, 2008.

[80] Dimitris N Politis and Joseph P Romano. Large-sample confidence regions based on subsamples under minimal assumptions. The Annals of Statistics, pages 2031–2050, 1994.

[81] Douglas M Potter. A permutation test for inference in logistic regression with small-and moderate-sized data sets. Statistics in Medicine, 24(5):693–708, 2005.

[82] Donald A Redelmeier, Gordon H Guyatt, and Roger S Goldstein. Assessing the minimal important difference in symptoms: a comparison of two techniques. Journal of Clinical Epidemiology, 49(11):1215–1219, 1996.

[83] Joseph P Romano and Michael Wolf. Exact and approximate stepdown methods for multiple hypothesis testing. Journal of the American Statistical Association, 100(469):94–108, 2005.

[84] Joseph P Romano and Michael Wolf. Stepwise multiple testing as formalized data snooping. Econometrica, 73(4):1237–1282, 2005.

[85] Joseph P Romano and Michael Wolf. Control of generalized error rates in multiple testing. The Annals of Statistics, pages 1378–1408, 2007.

[86] Kenneth J Rothman, Sander Greenland, and Timothy L Lash. Modern Epidemiology. Lippincott Williams & Wilkins, 2008.

[87] James J Schlesselman. Assessing effects of confounding variables. American Journal of Epidemiology, 108(1):3–8, 1978.

[88] M Schweinsberg and EL Uhlmann. The Pipeline Project 2. https://osf.io/skq2b/., 2018. Accessed: 2018-04-18.

[89] M Schweinsberg, Nikhil Madan, Michelangelo Vianello, S Amy Sommer, Jennifer Jordan, Warren Tierney, Eli Awtrey, Luke Lei Zhu, Daniel Diermeier, Justin E Heinze, et al. The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. Journal of Experimental Social Psychology, 66:55–67, 2016.

[90] Juliet Popper Shaffer. Modified sequentially rejective multiple test procedures. Journal of the American Statistical Association, 81(395):826–831, 1986.

[91] Juliet Popper Shaffer. Multiple hypothesis testing. Annual Review of Psychology, 46(1):561–584, 1995.

[92] Ian Shrier, Jean-François Boivin, Russell J Steele, Robert W Platt, Andrea Furlan, Ritsuko Kakuma, James Brophy, and Michel Rossignol. Should meta-analyses of interventions include observational studies in addition to randomized controlled trials? A critical examination of underlying principles. American Journal of Epidemiology, 166(10):1203–1209, 2007.

[93] Kurex Sidik and Jeffrey N Jonkman. Simple heterogeneity variance estimation for meta-analysis. Journal of the Royal Statistical Society: Series C (Applied Statistics), 54(2):367–384, 2005.

[94] R John Simes. An improved Bonferroni procedure for multiple tests of significance. Biometrika, 73(3):751–754, 1986.

[95] Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science, 22(11):1359–1366, 2011.

[96] Daniel J Simons. The value of direct replication. Perspectives on Psychological Science, 9(1):76–80, 2014.

[97] Daniel J Simons, Alex O Holcombe, and Barbara A Spellman. An introduction to Registered Replication Reports at Perspectives on Psychological Science. Perspectives on Psychological Science, 9(5):552–555, 2014.

[98] Uri Simonsohn. Small telescopes: Detectability and the evaluation of replication results. Psychological Science, 26(5): 559–569, 2015.

[99] Patty W Siri-Tarino, Qi Sun, Frank B Hu, and Ronald M Krauss. Meta-analysis of prospective cohort studies evaluating the association of saturated fat with cardiovascular disease. The American Journal of Clinical Nutrition, pages ajcn–27725, 2010.

[100] Jeremiah Stamler. Diet-heart: a problematic revisit. The American Journal of Clinical Nutrition, 91(3):497–499, 2010.

[101] Gavin B Stewart, Douglas G Altman, Lisa M Askie, Lelia Duley, Mark C Simmonds, and Lesley A Stewart. Statistical analysis of individual participant data meta-analyses: a comparison of methods and recommendations for practice. PloS One, 7(10):e46042, 2012.

[102] Wenguang Sun and Tony Cai. Large-scale multiple testing under dependence. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 71(2):393–424, 2009.

[103] Alex J Sutton, Keith R Abrams, David R Jones, David R Jones, Trevor A Sheldon, and Fujian Song. Methods for meta-analysis in medical research. 2000.

[104] Jin-Ling Tang. Weighting bias in meta-analysis of binary outcomes. Journal of Clinical Epidemiology, 53(11):1130–1136, 2000.

[105] Kristian Thorlund, Georgina Imberger, Michael Walsh, Rong Chu, Christian Gluud, Jørn Wetterslev, Gordon Guyatt, Philip J Devereaux, and Lehana Thabane. The number of patients and events required to limit the risk of overestimation of intervention effects in meta-analysis: a simulation study. PLoS One, 6(10):e25491, 2011.

[106] Bruce J Trock, Leena Hilakivi-Clarke, and Robert Clarke. Meta-analysis of soy intake and breast cancer risk. Journal of the National Cancer Institute, 98(7):459–471, 2006.

[107] Rebecca M Turner, David J Spiegelhalter, Gordon Smith, and Simon G Thompson. Bias modelling in evidence synthesis. Journal of the Royal Statistical Society: Series A (Statistics in Society), 172(1):21–47, 2009.

[108] Jeffrey C Valentine and Simon G Thompson. Issues relating to confounding and meta-analysis when including non-randomized studies in systematic reviews on the effects of interventions. Research Synthesis Methods, 4(1):26–35, 2013.

[109] Mark J van der Laan, Sandrine Dudoit, and Katherine S Pollard. Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. Statistical Applications in Genetics and Molecular Biology, 3(1):1–25, 2004.

[110] Mark J van der Laan, Merrill D Birkner, and Alan E Hubbard. Empirical Bayes and resampling based multiple testing procedure controlling tail probability of the proportion of false positives. Statistical Applications in Genetics and Molecular Biology, 4(1), 2005.

[111] Tyler VanderWeele and Peng Ding. Sensitivity analysis in observational research: introducing the E-value. Annals of Internal Medicine, pages doi: 10.7326/M16–2607, 2017.

[112] Tyler J VanderWeele. On a square-root transformation of the odds ratio for a common outcome. Epidemiology, 28(6):e58–e60, 2017.

[113] Tyler J VanderWeele. Outcome-wide epidemiology. Epidemiology, 28(3):399–402, 2017.

[114] Tyler J VanderWeele and Onyebuchi A Arah. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. Epidemiology, 22(1): 42–52, 2011.

[115] Josine Verhagen and Eric-Jan Wagenmakers. Bayesian tests to quantify the result of a replication attempt. Journal of Experimental Psychology: General, 143(4):1457, 2014.

[116] Areti Angeliki Veroniki, Dan Jackson, Wolfgang Viechtbauer, Ralf Bender, Jack Bowden, Guido Knapp, Oliver Kuss, Julian Higgins, Dean Langan, and Georgia Salanti. Methods to estimate the between-study variance and its uncertainty in meta-analysis. Research Synthesis Methods, 2015.

[117] Wolfgang Viechtbauer et al. Conducting meta-analyses in R with the metafor package. Journal of Statistical Software, 36(3): 1–48, 2010.

[118] E-J Wagenmakers, Titia Beek, Laura Dijkhoff, Quentin F Gronau, A Acosta, RB Adams Jr, DN Albohn, ES Allard, SD Benning, E-M Blouin-Hudon, et al. Registered Replication Report: Strack, Martin, & Stepper (1988). Perspectives on Psychological Science, 11(6):917–928, 2016.

[119] NJ Welton, AE Ades, JB Carlin, DG Altman, and JAC Sterne. Models for potentially biased evidence in meta-analysis using empirically based priors. Journal of the Royal Statistical Society: Series A (Statistics in Society), 172(1):119–136, 2009.

[120] Peter H Westfall and James F Troendle. Multiple testing with minimal assumptions. Biometrical Journal, 50(5):745–755, 2008.

[121] Peter H Westfall and S Stanley Young. Resampling-based multiple testing: Examples and methods for p-value adjustment. Taylor & Francis Group, 1993.

[122] Daniel Yekutieli and Yoav Benjamini. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. Journal of Statistical Planning and Inference, 82 (1-2):171–196, 1999.

[123] Arnold Zellner. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. Journal of the American Statistical Association, 57(298):348–368, 1962.