Published in final edited form as:
Epidemiology. 2020 May ; 31(3): 356–358. doi:10.1097/EDE.0000000000001180.

Robust Metrics and Sensitivity Analyses for Meta-Analyses of Heterogeneous Effects

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

[1]Quantitative Sciences Unit, Stanford University, Palo Alto, CA, USA

[2]Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA, USA

Abstract

We recently suggested new statistical metrics for routine reporting in random-effects meta-analyses to convey evidence strength for scientifically meaningful effects under effect heterogeneity. First, given a chosen threshold of meaningful effect size, we suggested reporting the estimated proportion of true effect sizes above this threshold. Second, we suggested reporting the proportion of effect sizes below a second, possibly symmetric, threshold in the opposite direction from the estimated mean. Our previous methods applied when the true effects are approximately normal, when the number of studies is relatively large, and when the proportion is between approximately 0.15 and 0.85. Here, we additionally describe robust methods for point estimation and inference that perform well under considerably more general conditions, as we validate in an extensive simulation study. The methods are implemented in the R package MetaUtility (function prop_stronger). We describe application of the robust methods to conducting sensitivity analyses for unmeasured confounding in meta-analyses.

Keywords

meta-analysis; effect sizes; heterogeneity; nonparametric; bootstrapping; confounding

Introduction

We recently suggested new statistical metrics for routine reporting in random-effects meta-analyses to convey evidence strength for scientifically meaningful effects under effect heterogeneity[1]. First, given a chosen threshold of meaningful effect size (q), we suggested reporting the estimated proportion of true effect sizes above this threshold ($\widehat{P}_{>q}$). Second, we suggested reporting the proportion of effect sizes below a second, possibly symmetric, threshold in the opposite direction from the estimated mean. These metrics can help identify if: (1) there are few effects of scientifically meaningful size despite a "statistically significant" pooled point estimate; (2) there are some large effects despite an apparently null point estimate; or (3) strong effects in the direction opposite of the pooled estimate also regularly occur. Additionally, these metrics can sometimes adjudicate apparent "conflicts" between meta-analyses[2] and can convey evidence strength in multisite replication projects[3].

Conflicts of interest: The authors declare that they have no conflicts of interest.

We had proposed parametric estimation methods with asymptotic inference based on the delta method[1]; those methods applied when the true effects are approximately normal, when the number of studies is relatively large, and when $\widehat{P}_{>q}$ is between approximately 0.15 and 0.85. Here, we additionally describe robust methods for point estimation and inference that perform well under more general conditions.

Methods

Let θ_i, $\hat{\theta}_i$, and $\hat{\sigma}_i$ respectively denote the true effect size, the point estimate, and the estimated standard error of the i^{th} study. If the parameter θ_i for each study were known, then a simple nonparametric estimate $\widehat{P}_{>q}$ would simply be the sample proportion of θ_i greater than q. Given that the θ_i are in fact unknown, it would seem intuitive to instead compute the sample proportion of point estimates $\hat{\theta}_i$ greater than q, but this approach is incorrect because dispersion in the $\hat{\theta}_i$ reflects not only true effect heterogeneity, but also statistical error due to finite sample sizes in the meta-analyzed studies. Thus, the $\hat{\theta}_i$ are overdispersed compared to the θ_i and would not themselves yield an unbiased estimate $\widehat{P}_{>q}$.

We therefore suggest computing the sample proportion using recently proposed “calibrated” estimates that have been appropriately shrunk to correct the overdispersion[4]. Let $\hat{\mu}$ and $\hat{\tau}^2$ represent classical Dersimonian-Laird[5] meta-analytic estimates of the mean and variance of the true effects; these moment-based estimates do not require parametric assumptions. Wang et al. (2019)[4] defined the calibrated point estimate for the i^{th} study as:

$$\tilde{\theta}_i = \hat{\mu} + \sqrt{\hat{\tau}^2/\left(\hat{\tau}^2 + \hat{\sigma}_i^2\right)}\left(\hat{\theta}_i - \hat{\mu}\right)$$

and showed that $\mathrm{Var}\left(\tilde{\theta}_i\right) = \hat{\tau}^2$, as desired. Intuitively, the calibrated estimate $\tilde{\theta}_i$ shrinks the point estimate $\hat{\theta}_i$ toward the estimated meta-analytic mean $\hat{\mu}$ with a degree of shrinkage that is inversely proportional to the study’s precision: relatively imprecise estimates $\hat{\theta}_i$ (i.e., those with large $\hat{\sigma}_i$) receive strong shrinkage toward $\hat{\mu}$, while relatively precise estimates receive less shrinkage and remain closer to their original values. Wang et al. (2019)[4] demonstrated that the calibrated estimates can be used to construct approximately unbiased prediction intervals for small meta-analyses and for non-normal true effect distributions. For our purposes, we propose estimating the proportion of scientifically meaningful effect sizes as the sample proportion of calibrated estimates above q, i.e., $\widehat{P}_{>q} = \widehat{P}\left(\tilde{\theta}_i > q\right)$. For inference, one can bootstrap pairs of $\left(\hat{\theta}_i,\ \hat{\sigma}_i\right)$ by drawing with replacement from the original sample and estimating in turn $\hat{\mu}$ and $\hat{\tau}^2$, $\tilde{\theta}_i$ for each study, and finally $\widehat{P}_{>q}$. A bias-corrected and accelerated (BCa) confidence interval[6;7] can then be constructed from the bootstrapped values of $\widehat{P}_{>q}$. (Naturally, analogous methods can be used to estimate the proportion of effects below another threshold.)

We also considered a simulation-based nonparametric method (here termed the “sign test method”) that was originally designed to estimate a given percentile of interest (e.g., the median) of a distribution of effect sizes and to construct a confidence interval[8]. The method involves first conducting nonparametric hypothesis tests that are similar to sign tests for each of many possible values for the percentile of interest, then inverting the rejection region to form a confidence interval. This method can be straightforwardly repurposed to provide an estimate and confidence interval for the proportion of effects above a threshold, $P_{>q}$, as we show in the eAppendix.

We assessed all methods' performance in an extensive simulation study of 480 scenarios with a range of true effect distributions (including normal, highly skewed, bimodal, and heavy-tailed distributions), meta-analyses with 5 to 50 studies of varying size, varying heterogeneity, and true proportions from 0.05 to 0.50. We included both realistic and extreme distributions of effect sizes in order to establish the boundary conditions under which the statistical methods would perform well. Some of the more extreme distributions may be unlikely to occur in practice, and conducting an aggregate meta-analysis when effects are clearly multimodal may be scientifically ill-advised in the first place. Details on the simulation study design appear in the eAppendix.

Results and conclusions

Based on the simulation results (eAppendix), we recommend reporting $\widehat{P}_{>q}$ and inference only for meta-analyses with at least 10 studies. In such meta-analyses, we recommend by default estimating $\widehat{P}_{>q}$ using the calibrated estimates; this method was the least biased for all distributions, though its root mean square error (RMSE) was sometimes higher than that of other methods. For inference, even when the effects appear normal, we recommend by default constructing the confidence interval by applying the bias-corrected and accelerated bootstrap to the calibrated estimates ("BCa-calibrated"); this method achieved nominal coverage in almost all scenarios and always achieved at least 90% coverage. The sign test method sometimes performed poorly when heterogeneity was low to moderate, and it offered few advantages over the BCa-calibrated method. We therefore do not recommend its use in practice to estimate or conduct inference for $P_{>q}$.

The BCa-calibrated method did sometimes lose considerable precision compared to the parametric method in certain scenarios in which the latter achieved approximately nominal coverage (e.g., see eFigure 7), so for large meta-analyses with apparently normal effects and estimating a proportion close to 0.50, one might reasonably choose to substitute the parametric interval for the default BCa-calibrated interval. For example, in relatively small meta-analyses, estimating the amount of heterogeneity can be inherently imprecise[9]. This uncertainty propagates to the confidence interval for $P_{>q}$ and, in small meta-analyses, may result in confidence intervals that span most or all of the possible range [0, 1]. Reporting confidence intervals in these settings may nevertheless be informative: a very wide confidence interval may instill appropriate circumspection about what can be learned regarding the distribution of true effects in a small meta-analysis, even if $\hat{\mu}$ itself may have a narrow confidence interval (for example, see the applied example in the eAppendix). A wide confidence interval may further suggest the value of performing a larger meta-analysis when more literature becomes available. As an additional limitation, the BCa-calibrated interval may sometimes fail to converge for small meta-analyses. (When the BCa-calibrated interval fails to converge, it may seem attractive to construct a simpler bootstrapped confidence interval using percentiles of the bootstrapped calibrated estimates. However, we recommend against this method; additional simulation results suggested that it performed quite poorly.)

The recommended methods are implemented in the function prop_stronger in the R package MetaUtility as of version 2.0.0. We illustrate this software and approach with an applied example in the eAppendix, where we also discuss extensions to sensitivity analysis for unmeasured confounding.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Source of funding: MBM and TVW were supported by NIH grant R01 CA222147. The funders had no role in the design, conduct, or reporting of this research.

References

[1]. Mathur Maya B and VanderWeele Tyler J. New metrics for meta-analyses of heterogeneous effects. Statistics in Medicine, 2018.

[2]. Mathur Maya B and VanderWeele Tyler J. Finding common ground in meta-analysis “wars” on violent video games. Perspectives on Psychological Science, page 1745691619850104, 2019.

[3]. Mathur Maya B and VanderWeele Tyler J. New statistical metrics for multisite replication projects. Under review. Preprint retrieved from https://osf.io/w89s5/.

[4]. Wang Chia-Chun and Lee Wen-Chung. A simple method to estimate prediction intervals and predictive distributions: Summarizing meta-analyses beyond means and confidence intervals. Research Synthesis Methods, 10(2):255–266, 2019. [PubMed: 30835918]

[5]. DerSimonian Rebecca and Laird Nan. Meta-analysis in clinical trials. Controlled Clinical Trials, 7(3):177–188, 1986. [PubMed: 3802833]

[6]. Efron Bradley. Better bootstrap confidence intervals. Journal of the American Statistical Association, 82(397):171–185, 1987.

[7]. Carpenter James and Bithell John. Bootstrap confidence intervals: When, which, what? A practical guide for medical statisticians. Statistics in Medicine, 19(9):1141–1164, 2000. [PubMed: 10797513]

[8]. Wang Rui, Tian Lu, Cai Tianxi, and Wei LJ. Nonparametric inference procedure for percentiles of the random effects distribution in meta-analysis. The Annals of Applied Statistics, 4(1):520, 2010. [PubMed: 25678939]

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