



OPEN ACCESS



# Assessing robustness to worst case publication bias using a simple subset meta-analysis

Maya B Mathur<sup>1</sup>

<sup>1</sup>Quantitative Sciences Unit and Department of Pediatrics, Stanford University, Palo Alto, CA 94304, USA

Correspondence to: mmathur@stanford.edu (ORCID 0000-0001-6698-2607)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2024;384:e076851 <http://dx.doi.org/10.1136/bmj-2023-076851>

Accepted: 15 January 2024

This article discusses a simple method, known as a meta-analysis of non-affirmative studies, to assess how robust a meta-analysis is to publication bias that favors affirmative studies (studies with significant P values and point estimates in the desired direction) over non-affirmative studies (studies with non-significant P values or point estimates in the undesired direction). This method is a standard meta-analysis that includes only non-affirmative studies. The resulting meta-analytical estimate corrects for worst case publication bias, a hypothetical scenario in which affirmative studies are almost infinitely more likely to be published than non-affirmative studies. If this estimate remains in the same direction as the uncorrected estimate and is of clinically meaningful size, this suggests that the meta-analysis conclusions would not be overturned by any amount of publication bias favoring affirmative studies. Meta-analysis of non-affirmative studies complements an uncorrected meta-analysis and other publication bias analyses by accommodating small meta-analyses, non-normal effects, heterogeneous effects across studies,

and additional forms of selective reporting (in particular, P-hacking).

Meta-analytical evidence contributes substantially to clinical guidelines and policy, but the results of meta-analyses can be severely compromised by publication bias. Such bias can occur, for example, if studies that support a given hypothesis are more likely to be published than studies with null or negative results.<sup>1</sup> Assessing the robustness of results to potential publication bias is therefore a key component of performing a meta-analysis.<sup>1</sup> However, a review of recent meta-analyses in high impact medical journals found that 55% did not assess publication bias at all.<sup>2</sup> Of meta-analyses that did assess publication bias, most (85%) exclusively considered whether there was asymmetry in funnel plots, which plot the point estimates versus standard errors from studies (fig 1).<sup>2</sup> Asymmetry in funnel plots is often assessed visually, or by using statistical methods such as Egger's regression,<sup>4</sup> trim and fill,<sup>5</sup> precision effect test (PET),<sup>6</sup> or precision effect test and precision effect estimate with standard errors (PET-PEESE).<sup>6</sup> These methods assess whether there are small-study effects, which occur when small studies tend to have larger point estimates than large studies.<sup>4 5</sup> As the methods' originators and others have noted, small-study effects might reflect not only publication bias but also genuine scientific differences between smaller and larger studies.<sup>2 4 7 8</sup> For example, the most effective interventions might be used in smaller studies if such interventions are more expensive to implement.<sup>2</sup>

Despite longstanding directives about interpreting funnel plots and related methods, medical researchers do routinely use them with the intention of assessing publication bias.<sup>2 9</sup> Interpreting funnel plot asymmetry in this manner requires making implicit assumptions about how effect sizes are distributed across studies and about the mechanism of publication bias. In papers that use funnel plots methods, these assumptions are rarely made explicit. Furthermore, in many meta-analyses, the assumptions might not capture the way publication bias actually operates. The key assumptions are that the publication process favors large point estimates rather than significant P values, and that it does not affect the largest studies. Also, when interpreted as tests of publication bias, some methods based on funnel plots can perform inadequately for small meta-analyses or when effects differ across studies (heterogeneity).<sup>2 7 8</sup> Last, these methods can perform poorly if the estimates included in the meta-analysis are statistically dependent, which can occur if, for example, the meta-analyst includes multiple estimates from the same study.<sup>10</sup> I concur

## SUMMARY POINTS

Meta-analytical evidence contributes substantially to clinical guidelines and policy, but the results of meta-analyses can be severely compromised by publication bias

Meta-analysis of non-affirmative studies is a simple method to conservatively assess the robustness of meta-analysis results to publication bias

To perform this method, a standard meta-analysis is conducted of only studies that are non-significant or have estimates in the undesired direction

As a conservative sensitivity analysis, this approach complements existing methods by accommodating additional forms of selective reporting, small meta-analyses, and heterogeneous effects that might not be normal

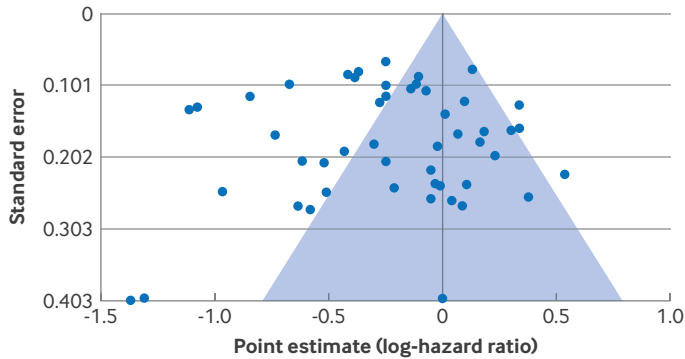


Fig 1 | Basic funnel plot for a meta-analysis by Li et al<sup>3</sup>

with others that funnel plot methods can be useful to investigate small-study effects in general, but that the methods should not be interpreted as specific assessments of publication bias.<sup>7</sup>

Other methods are specifically designed to investigate or to correct for publication bias. For example, selection models involve assuming a certain model of how publication bias operates and then fitting a model that estimates the usual meta-analytical quantities of interest (the mean and heterogeneity of population effect sizes) as well as the strength of the publication bias itself. Numerous selection models exist, but perhaps the most widely used is the two-step selection model.<sup>11</sup> This model assumes that studies with significant P values and positive estimates are more likely to be published than studies with non-significant P values or negative estimates. If the publication process is thought to favor negative-signed estimates rather than positive-signed estimates, then the same definitions apply but with “positive” and “negative” reversed.

Other selection models make different assumptions about the mechanism of publication bias<sup>12,13</sup>; some, for example, assume that a study’s probability of being published is a continuous function of the study’s one-tailed P value.<sup>13</sup> Selection models are statistically well justified and accommodate effect heterogeneity, and we recommend that their results be reported more routinely in meta-analyses.<sup>2</sup> However, these methods have their own limitations and assumptions. In particular, most selection models assume that the true effect sizes (before any publication bias occurs) are normally distributed and independent, and the models can perform poorly when these assumptions are violated.<sup>10</sup> Also, these methods can require a large number of studies to perform well—this is especially the case for selection models that accommodate more complex forms of publication bias.<sup>14,15</sup>

Other methods for publication bias are essentially hybrids of methods based on the funnel plot and selection models.<sup>16,17</sup> For example, robust bayesian model-averaging (RoBMA) involves specifying prior beliefs in the plausibility of various publication bias models (including selection models and the models underlying PET and PET-PEESE) and essentially

averaging over the results in a manner that accounts for model fit given the data.<sup>16</sup> These hybrid methods can provide some additional flexibility compared with funnel plot methods or selection models alone, but they also share many of their limitations. For example, RoBMA can perform worse than an uncorrected meta-analysis in the presence of P-hacking in addition to publication bias.<sup>18</sup> The relative performance of funnel plot methods, selection models, and hybrid methods is further discussed below.

All of these standard methods specifically assess publication bias that acts as a filter for which studies are published and ultimately included in the meta-analysis. However, selective reporting can also occur within studies. Researchers might P-hack by fitting different models to the same dataset or by analyzing several outcomes in an attempt to obtain an affirmative estimate.<sup>19–21</sup> In fact, such P-hacking is quite common, even according to researchers’ self-admissions.<sup>21</sup> When there is both traditional publication bias and P-hacking, standard methods for publication bias can be severely biased in either direction.<sup>18</sup>

This article describes a simple and complementary method known as a meta-analysis of non-affirmative studies (MAN). MAN is an additional method to assess how robust a meta-analysis might be to publication bias as well as many forms of P-hacking. This method is not intended to replace existing ones but rather to serve as a useful addition that can provide complementary insights while allowing for small meta-analyses, non-normal effects, and dependent effect sizes. This article describes why conducting a meta-analysis of non-affirmative results is a conservative sensitivity analysis; that is, it characterizes robustness to a hypothetical, worst case form of publication bias. The article then describes how to conduct and interpret MAN in practice, illustrated by reanalysis of three previously published meta-analyses.

### Meta-analysis of non-affirmative results

Conducting MAN involves simply meta-analyzing only the non-affirmative studies and excluding all affirmative studies.<sup>18,22</sup> No special statistical methods are required—a standard meta-analysis is simply conducted of only the non-affirmative studies. This method is a sensitivity analysis rather than a method for bias correction—that is, the method does not estimate the actual strength of publication bias, but instead assesses how results might change for a certain amount of publication bias. Specifically, MAN provides a meta-analytical estimate that corrects for a worst case form of publication bias in which affirmative studies (studies with significant P values and estimates in the desired direction) are more likely to be published than non-affirmative studies (studies with non-significant P values or estimates in the undesired direction). Instead of estimating how strong publication bias actually is in the meta-analysis, the meta-analytical estimate from MAN corrects for a hypothetical worst case situation in which publication bias favors affirmative studies almost infinitely more than non-affirmative studies.<sup>22</sup>

To understand how this works, assume that affirmative studies were known to be three times as likely to be published as non-affirmative studies. To counteract this threefold favoring of affirmative studies in the publication process, a bias-corrected meta-analysis would need to weight the published non-affirmative studies three times as much as the affirmative studies.<sup>22</sup> This weighting method is essentially the same as methods routinely used to correct survey samples for non-representative sampling. In practice, however, it is not known how much publication bias has occurred, so instead, the worst case publication bias can be conservatively considered. Following the same logic, if affirmative studies were infinitely more likely to be published than non-affirmative studies, a bias-corrected meta-analysis would need to weight the published non-affirmative studies infinitely more than the affirmative studies—which corresponds to analyzing only the non-affirmative studies.<sup>22</sup>

#### Advantages of conducting a meta-analysis of non-affirmative studies

Reporting the worst case estimate from MAN, along with existing publication bias methods, has several advantages. First, the worst case estimate allows for a conservative consideration of how results might change in the presence of worst case publication bias. In practice, publication bias will rarely be this severe, but if the worst case estimate remains in the same direction as the uncorrected estimate, and remains of clinically meaningful size (ideally also with a reasonably precise confidence interval that excludes the null), then this provides strong evidence that the results would not be overturned by any amount of publication bias that might be present. Although the worst case estimate is highly conservative by design, it is often still informative in practice: for 66% of meta-analyses sampled across scientific disciplines, the worst case estimate agreed in direction with the uncorrected estimate, and for 25% of meta-analyses, its confidence interval also excluded the null.<sup>23</sup> If the worst case estimate is near the null or is in the opposite direction from the uncorrected estimate, however, this suggests that the meta-analysis might not be robust to worst case publication bias, and then robustness to less extreme publication bias could be assessed by conceptually related sensitivity analyses (section 2, supplement).<sup>22</sup>

As noted previously, MAN can be applied to meta-analyses with characteristics that can compromise the performance of standard methods for publication bias (eg, due to heterogeneity, non-normal effects, small number of studies, or dependent effects).<sup>1 4 8 24</sup> Additionally, applying MAN can help deal with not only publication bias but also other forms of selective reporting that occur within rather than across studies. In particular, if the studies are P-hacked in a manner that favors affirmative results, then the MAN estimate is still conservative (ie, attenuated towards the null). In contrast, standard methods for publication bias could be severely biased either towards or away from

the null.<sup>18</sup> Thus, MAN can provide a more holistic assessment of robustness to multiple forms of selective reporting than do standard methods for publication bias. For example, a large simulation study assessed the performance of MAN in over 80 scenarios exhibiting numerous forms of selective reporting (comprising both publication bias and P-hacking), including many that were more complicated than the simple model assumed by MAN.<sup>18</sup> These simulations found that the performance of MAN was robust to numerous plausible departures from its assumptions, whereas comparison methods (two-step selection models, PET-PEESE, and RoBMA) often had substantially compromised performance in scenarios with P-hacking.<sup>18</sup> Section 1 of the supplement describes details of these results.

#### How to conduct a meta-analysis of non-affirmative studies

Conducting MAN in practice is straightforward. Meta-analysts first need to decide on substantive grounds which direction of estimates—those greater than the null or those less than the null—are likely to be published preferentially. For example, if the exposure or intervention of interest is thought to be positively associated with the outcome, then publication bias might favor estimates that are greater than the null. In this case, affirmative studies would be defined as those with significant P values and estimates greater than the null, and non-affirmative studies would be those with non-significant P values or estimates less than the null. Alternatively, if the exposure is thought to be negatively associated with the outcome, then publication bias might favor estimates that are less than the null. In this case, affirmative studies would be those with significant P values and estimates less than the null. The next step would be to conduct a standard meta-analysis that includes only the non-affirmative studies. In practice, MAN can be conducted using any standard method for random effects or fixed effects meta-analysis.<sup>25-27</sup> We would generally suggest using robust methods for random effects meta-analysis that do not require population effect sizes to be normal, accommodate dependent effect sizes, and perform well for small meta-analyses.<sup>28-30</sup> These methods, known as robust variance estimation, are easy to implement in R using the packages `robumeta`<sup>31</sup> or `clubSandwich`.<sup>32</sup>

Along with the numerical results of the MAN analysis, we suggest that meta-analysts create a modified funnel plot, known as the significance funnel. This plot can be created using the R package `PublicationBias`<sup>22</sup> or an online web tool (<https://metabias.io/>). Like a standard funnel plot, the significance funnel plot shows studies' point estimates against their standard errors (eg, fig 2). But whereas a standard funnel plot focuses on detecting correlation between studies' estimates and their standard errors, the significance funnel plot focuses on portraying the extent to which the worst case MAN estimate differs from the uncorrected estimate. Thus, the significance funnel differentiates affirmative studies (yellow points) from non-affirmative studies (blue points). The significance funnel also

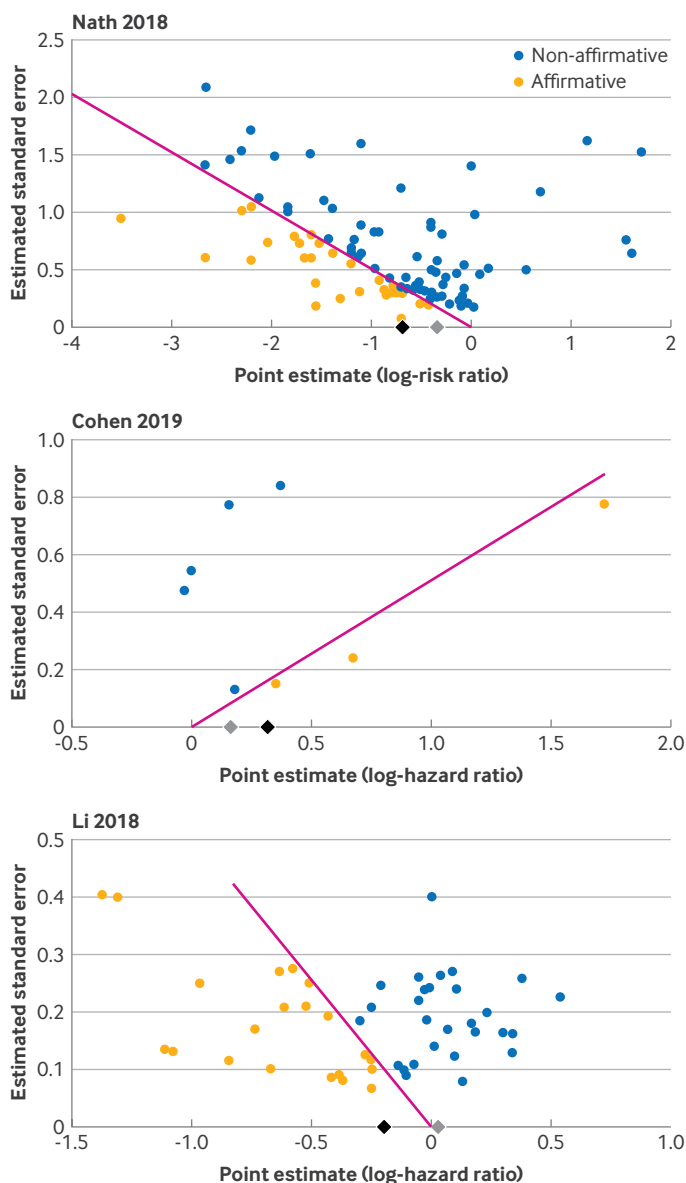


Fig 2 | Significance funnel plots for three example meta-analyses.<sup>33 34</sup> Studies on the pink diagonal lines have P values exactly at the threshold  $\alpha=0.05$ . Black diamond on x axis=uncorrected estimate in all studies; grey diamond on x axis=worst case estimate from meta-analysis of non-affirmative studies; yellow points=affirmative studies; blue points=non-affirmative studies

shows the uncorrected meta-analytical estimate (black diamond on x axis) and the MAN estimate (grey diamond on x axis). As a simple rule of thumb, if the diamonds are close to one another, this suggests that the meta-analysis is relatively robust to worst case publication bias. In contrast, if the diamonds are far apart or if the grey diamond represents an effect size that is too small to be clinically meaningful, the meta-analysis might be sensitive to worst case publication bias. However, the meta-analysis might still be robust to less severe publication bias, which researchers can assess using follow-up sensitivity analyses (section 2, supplement).<sup>22</sup>

Although our focus is not on the design of meta-analyses in general, a few points merit special attention

in the context of publication bias and MAN. First, as a general principle for meta-analysis, inclusion criteria for primary analyses should generally be designed to exclude studies of low methodological quality. This practice is not yet commonplace, especially in meta-analyses of non-randomized studies.<sup>35 36</sup> However, using methodological inclusion criteria can help reduce internal bias in the meta-analysis overall, which can occur independent of publication bias.<sup>36</sup> In the context of conducting MAN, this practice also helps alleviate the possibility that the non-affirmative studies will be of low quality. Second, to characterize differences between affirmative and non-affirmative studies, meta-analysts could report risk-of-bias ratings (eg, ROBINS-I<sup>37</sup>) and key methodological characteristics,<sup>36</sup> key study level moderators (eg, patient population), and typical sample sizes. These characteristics could be reported separately for all studies, for affirmative studies, and for non-affirmative studies, which would be a simple extension for existing best practices.<sup>38</sup> More technical points regarding heterogeneity are described in the supplement (section 3).

### Examples

We illustrate applying and interpreting MAN using three published meta-analyses (fig 2). First, a meta-analysis by Nath et al assessed the effect of using atraumatic needles versus conventional needles for lumbar puncture on multiple outcomes.<sup>33</sup> Our reanalysis of studies on one of the outcomes—the occurrence of any type of headache—included 99 randomized studies, of which 31 were affirmative (favoring atraumatic needles) and 68 were non-affirmative. The uncorrected estimate was risk ratio 0.50 (95% confidence interval 0.44 to 0.58;  $P<0.001$ ). The worst case MAN estimate was 0.71 (0.64 to 0.79;  $P<0.001$ ), which suggests that both the point estimate and confidence interval are quite robust to even worst case publication bias that favors protective effects of atraumatic needles.

Second, Cohen et al's meta-analysis assessed the association of white coat hypertension (ie, high blood pressure occurring only when the patient is at the doctor's office) with cardiovascular events.<sup>34</sup> Their analysis included eight non-randomized studies, of which three were affirmative (suggesting a detrimental association) and five were non-affirmative. On reanalysis, the uncorrected estimate was hazard ratio 1.37 (95% confidence interval 0.95 to 1.98;  $P=0.07$ ); the worst case estimate was 1.18 (0.84 to 1.66;  $P=0.12$ ). The authors correctly reported that funnel plot methods can perform poorly for this relatively small number of studies, so they did not assess publication bias. Nevertheless, MAN can be applied and suggests that even under worst case publication bias, the pooled estimate would remain in the detrimental direction (hazard ratio 1.18), although the wide confidence interval includes the null due to the small number of non-affirmative studies.

Third, Li et al's meta-analysis<sup>3</sup> assessed the effect of PI3K/AKT/mTOR pathway inhibitors compared with various control treatments on progression-free



survival in patients with advanced solid tumors. Their analysis included 50 estimates that were clustered within 39 randomized studies. Of the estimates, 22 were affirmative (suggesting protective effects) and 28 were non-affirmative. On reanalysis, the uncorrected estimate was hazard ratio 0.82 (95% confidence interval 0.74 to 0.91;  $P < 0.001$ ), but the worst case estimate was close to the null (1.03; 0.94 to 1.12;  $P = 0.49$ ). This suggests that the results might not be robust to worst case publication bias that favors protective effects; follow-up sensitivity analyses could then be conducted (section 2, supplement).<sup>22</sup>

## Discussion

Key advantages of MAN are its applicability to meta-analyses that have heterogeneous and even non-normal effects, dependent effects, a small number of studies, and P-hacking in addition to publication bias. Funnel plot methods and other standard methods for publication bias can perform poorly in these situations.<sup>1 4 8 18 24</sup> MAN also has certain conceptual strengths. MAN assumes that publication bias favors significant studies with estimates in the desired direction and that all studies, regardless of size, can be affected by publication bias. In contrast, funnel plot methods effectively assume that publication bias favors large point estimates, rather than significant P values, and that the largest studies are not affected by publication bias at all.<sup>4</sup> Because MAN considers a different form of publication bias—one that is supported by empirical findings on how investigators interpret and report P values<sup>39-41</sup>—it can provide different insights than do funnel plots. Selection models can also consider this form of publication bias,<sup>11</sup> and as described in the introduction, we recommend routinely reporting results of such models as well.<sup>2</sup>

MAN also has limitations. The method assumes that publication bias (and P-hacking, if present) favor affirmative results but that, among non-affirmative results, it does not favor larger point estimates.<sup>18 22</sup> In other words, non-affirmative results with larger point estimates are no more likely to be published than other non-affirmative results with smaller point estimates. The simulation results mentioned above do suggest that the method is quite robust to numerous departures from these assumptions, but simulations cannot be exhaustive. In practice, we suggest examining diagnostic plots, such as the density of studies' z scores, to help assess whether this assumption is plausible.<sup>18 23</sup> Additionally, MAN cannot be applied to meta-analyses that contain only affirmative studies, and if there are very few non-affirmative studies, then its confidence interval could be wide. Last, because MAN considers only worst case publication bias, if the worst case estimate is near the null, the meta-analysis might nevertheless be robust to less severe publication bias. Such an analysis is inconclusive, and in this case, we would suggest conducting the sensitivity analyses mentioned above to consider less extreme publication bias and describe the amount of publication bias that would be required to explain away the results (section

2, supplement).<sup>22</sup> However, again, empirical evidence suggests that for many meta-analyses—although certainly not all—MAN might in fact suggest robustness to even worst case publication bias.<sup>23</sup>

In summary, MAN could be routinely reported in meta-analyses to help assess robustness to worst case publication bias or P-hacking that favors affirmative results, ideally along with a significance funnel plot. MAN complements an uncorrected meta-analysis and standard publication bias analyses by accommodating effects that differ across studies, small meta-analyses, non-normal effects, and additional forms of selective reporting.

I thank Vandana Mathur for helpful comments. All code and data required to reproduce the applied examples are publicly available and documented (<https://osf.io/8nqwy/>).

**Contributors:** MBM conducted this research and is the guarantor. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

**Funding:** This research was supported by US National Institutes of Health grants R01 LM013866, UL1TR003142, P30CA124435, and P30DK116074. The funders had no role in the design, conduct, or reporting of this research.

**Competing interests:** All authors have completed the ICMJE uniform disclosure form at [www.icmje.org/disclosure-of-interest/](http://www.icmje.org/disclosure-of-interest/) and declare: financial support from the National Institutes of Health for the submitted work; no financial relationships with any organizations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

**Provenance and peer review:** Not commissioned; externally peer reviewed.

This is an Open Access article distributed in accordance with the terms of the Creative Commons Attribution (CC BY 4.0) license, which permits others to distribute, remix, adapt and build upon this work, for commercial use, provided the original work is properly cited. See: <http://creativecommons.org/licenses/by/4.0/>.

- Jin Z-C, Zhou X-H, He J. Statistical methods for dealing with publication bias in meta-analysis. *Stat Med* 2015;34:343-60. doi:10.1002/sim.6342
- Maier M, VanderWeele TJ, Mathur MB. Using selection models to assess sensitivity to publication bias: A tutorial and call for more routine use. *Campbell Syst Rev* 2022;18:e1256. doi:10.1002/cl2.1256
- Li X, Dai D, Chen B, Tang H, Xie X, Wei W. Efficacy of PI3K/AKT/mTOR pathway inhibitors for the treatment of advanced solid cancers: A literature-based meta-analysis of 46 randomised control trials. *PLoS One* 2018;13:e0192464. doi:10.1371/journal.pone.0192464
- Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;315:629-34. doi:10.1136/bmj.315.7109.629
- Duval S, Tweedie R. Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics* 2000;56:455-63. doi:10.1111/j.0006-341X.2000.00455.x
- Stanley TD, Doucouliagos H. Meta-regression approximations to reduce publication selection bias. *Res Synth Methods* 2014;5:60-78. doi:10.1002/jrsm.1095
- Sterne JA, Sutton AJ, Ioannidis JP, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ* 2011;343:d4002. doi:10.1136/bmj.d4002
- Lau J, Ioannidis JP, Terrin N, Schmid CH, Olkin I. The case of the misleading funnel plot. *BMJ* 2006;333:597-600. doi:10.1136/bmj.333.7568.597
- Ropovik I, Adamkovic M, Greger D. Neglect of publication bias compromises meta-analyses of educational research. *PLoS One* 2021;16:e0252415. doi:10.1371/journal.pone.0252415
- Rodgers MA, Pustejovsky JE. Evaluating meta-analytic methods to detect selective reporting in the presence of dependent effect sizes. *Psychol Methods* 2020;26:141. doi:10.1037/met0000300
- Vevea JL, Hedges LV. A general linear model for estimating effect size in the presence of publication bias. *Psychometrika* 1995;60:419-35. doi:10.1007/BF02294384.

- 12 Copas JB, Shi JQ. A sensitivity analysis for publication bias in systematic reviews. *Stat Methods Med Res* 2001;10:251-65. doi:10.1177/096228020101000402
- 13 Citkovic M, Vevea JL. A parsimonious weight function for modeling publication bias. *Psychol Methods* 2017;22:28-41. doi:10.1037/met0000119
- 14 McShane BB, Böckenholt U, Hansen KT. Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspect Psychol Sci* 2016;11:730-49. doi:10.1177/1745691616662243
- 15 Carter EC, Schönbrodt FD, Gervais WM, Hilgard J. Correcting for bias in psychology: A comparison of meta-analytic methods. *Adv Methods Pract Psychol Sci* 2019;2:115-44. doi:10.1177/2515245919847196
- 16 Bartoš F, Maier M, Wagenmakers EJ, Doucouliagos H, Stanley TD. Robust Bayesian meta-analysis: Model-averaging across complementary publication bias adjustment methods. *Res Synth Methods* 2023;14:99-116. doi:10.1002/jrsm.1594
- 17 Bom PRD, Rachinger H. A kinked meta-regression model for publication bias correction. *Res Synth Methods* 2019;10:497-514. doi:10.1002/jrsm.1352
- 18 Mathur MB. P-hacking in meta-analyses: A formalization and new meta-analytic methods. *Res Synth Methods* 2024. doi:10.1002/jrsm.1701
- 19 Brodeur A, Lé M, Sangnier M, Zylberberg Y. Star Wars: The empirics strike back. *Am Econ J Appl Econ* 2016;8:1-32. doi:10.1257/app.20150044
- 20 Jager LR, Leek JT. An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics* 2014;15:1-12. doi:10.1093/biostatistics/kxt007
- 21 John LK, Loewenstein G, Prelec D. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol Sci* 2012;23:524-32. doi:10.1177/0956797611430953
- 22 Mathur MB, VanderWeele TJ. Sensitivity analysis for publication bias in meta-analyses. *J R Stat Soc Ser C Appl Stat* 2020;69:1091-119. doi:10.1111/rssc.12440
- 23 Mathur MB, VanderWeele TJ. Estimating publication bias in meta-analyses of peer-reviewed studies: A meta-meta-analysis across disciplines and journal tiers. *Res Synth Methods* 2021;12:176-91. doi:10.1002/jrsm.1464
- 24 Pustejovsky JE, Rodgers MA. Testing for funnel plot asymmetry of standardized mean differences. *Res Synth Methods* 2019;10:57-71. doi:10.1002/jrsm.1332
- 25 Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. *Methods for meta-analysis in medical research*. John Wiley & Sons, 2000.
- 26 Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. *Stat Med* 2001;20:825-40. doi:10.1002/sim.650
- 27 DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;7:177-88. doi:10.1016/0197-2456(86)90046-2
- 28 Hedges LV, Tipton E, Johnson MC. Robust variance estimation in meta-regression with dependent effect size estimates. *Res Synth Methods* 2010;1:39-65. doi:10.1002/jrsm.5
- 29 Pustejovsky JE, Tipton E. Meta-analysis with robust variance estimation: Expanding the range of working models. *Prev Sci* 2022;23:425-38. doi:10.1007/s11121-021-01246-3
- 30 Tipton E. Small sample adjustments for robust variance estimation with meta-regression. *Psychol Methods* 2015;20:375-93. doi:10.1037/met0000011
- 31 Fisher Z, Tipton E. Robumeta: An R-package for robust variance estimation in meta-analysis. *arXiv* 2015;1503.02220.
- 32 Pustejovsky J. *clubSandwich: Cluster-Robust (Sandwich) Variance Estimators with Small-Sample Corrections*. R package version 0.5.10. 2023. <https://cran.r-project.org/web/packages/clubSandwich/index.html>.
- 33 Nath S, Koziarz A, Badhiwala JH, et al. Atraumatic versus conventional lumbar puncture needles: a systematic review and meta-analysis. *Lancet* 2018;391:1197-204. doi:10.1016/S0140-6736(17)32451-0
- 34 Cohen JB, Lotito MJ, Trivedi UK, Denker MG, Cohen DL, Townsend RR. Cardiovascular events and mortality in white coat hypertension: a systematic review and meta-analysis. *Ann Intern Med* 2019;170:853-62. doi:10.7326/M19-0223
- 35 Kraemer HC, Gardner C, Brooks J, Yesavage JA. Advantages of excluding underpowered studies in meta-analysis: Inclusionist versus exclusionist viewpoints. *Psychol Methods* 1998;3:23. doi:10.1037/1082-989X.3.1.23
- 36 Mathur MB, VanderWeele TJ. Methods to address confounding and other biases in meta-analyses: review and recommendations. *Annu Rev Public Health* 2022;43:19-35. doi:10.1146/annurev-publhealth-051920-114020
- 37 Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* 2016;355:i4919. doi:10.1136/bmj.i4919
- 38 Page MJ, Higgins JP, Sterne JAC. Chapter 13. *Cochrane handbook for systematic reviews of interventions*. 2023. <https://training.cochrane.org/handbook>.
- 39 Laber EB, Shedden K. Statistical significance and the dichotomization of evidence. *J Am Stat Assoc* 2017;112:902-4. doi:10.1080/01621459.2017.1311265
- 40 Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD. The extent and consequences of p-hacking in science. *PLoS Biol* 2015;13:e1002106. doi:10.1371/journal.pbio.1002106
- 41 Masicampo EJ, Lalande DR. A peculiar prevalence of p values just below .05. *Q J Exp Psychol (Hove)* 2012;65:2271-9. doi:10.1080/17470218.2012.711335

### Web appendix: Supplement