Statistics and Its Interface Volume 13 (2020) 475–484

Small-study effects: current practice and challenges for future research

Arielle Marks-Anglin* and Yong Chen

Meta-analyses and systematic reviews are highly valued as evidence for clinical decision- and policy-making. However, inference in these settings may be invalid if the studies do not come from the same underlying distribution. Small study effects is one form of heterogeneity that can lead to biased estimates, particularly if it arises due to the selective publishing of studies, a phenomenon known as publication bias. In this paper we discuss landmark methods for diagnosing the presence of small-study effects and correcting for them, as well as the limitations of each method. We also identify ongoing challenges and key areas in need of methodological innovation.

Keywords and phrases: Small-study effects, Publication bias, Meta-analysis.

1. INTRODUCTION

Meta-analyses and systematic reviews synthesize results from individual trials and provide a strong source of evidence for treatment evaluation and clinical decision-making. A key assumption for valid inference in meta-analysis is that the component study estimates represent a random sample from a common, symmetric distribution centered on the true treatment effect. Random variability about this mean does not constitute a violation of this assumption, however heterogeneity that is tied to intrinsic study characteristics such as patient demographics, experimental design or results can lead to bias in the summary treatment effect estimate. The phenomenon whereby estimates from smaller studies included in a meta-analysis are on average different, and often further from the null, compared to those from larger studies is termed “small-study effects” (SSE) [69].

SSE may arise from a variety of sources. A particularly concerning and well-known cause of SSE is the selective publishing of studies based on the significance or favorability of results, an issue known as ‘publication bias’. This leads to an incomplete and biased evidence-base for inference and decision making. Because smaller studies require a larger treatment effect to achieve statistical significance, publication bias often manifests as SSE. However, SSE may also be due to differences in the design and conduct of smaller trials compared to larger studies. For example, smaller studies may enroll high-risk populations that are more responsive to treatment [31]. They may also implement the intervention more carefully than larger, multicenter trials [19], resulting in larger treatment effects. On the other hand, small trials are more likely to be of lower methodologic quality (eg. inadequate allocation concealment and blinding) and produce exaggerated effect estimates [58, 39]. Furthermore, the choice of effect-measure used may be inherently correlated with precision and therefore study size, inducing SSE [49]. Understanding the source of SSE in a meta-analysis is important for knowing how to address it. While some sources of SSE may be dealt with through transformation of the outcome variable or subgroup analysis to improve generalizability of results, publication bias is a missing-not-at-random problem and therefore requires assumptions regarding the missing studies in order to correct for it. Furthermore, distinguishing SSE from other sources of heterogeneity or chance variability of studies can be difficult.

In this paper we provide an overview of popular approaches to handling SSE. A number of methods have been developed for detecting and correcting for SSE that fall under two predominant categories: graph-based methods (including visual inspection, statistical testing and regression of plotted study estimates) and weighted distribution methods, also known as selection models. As we will discuss in this paper, many of the more commonly used graphical approaches, while having improved the quality and interpretation of meta-analyses in the last 30 years, can be limited in their ability to distinguish between sources of SSE, as well as disentangle SSE from random or explainable heterogeneity. Selection models overcome some of the limitations of graph-based approaches but may be more difficult to implement. An overview of the most well-known methods has previously been provided by Rücker et al. [53], followed by a chapter in the educational text *Meta-Analysis with R* [62]. In addition to discussing these landmark methods for SSE, we seek to provide an updated review including recent developments, particularly in the area of selection models, as well as directions for future work. In Section 1 we give an overview of select graph-based approaches to SSE, along with a discussion of their strengths and weaknesses. In Section 2 we offer a similar review of selection models for publication bias. Finally, in Section 3 we draw attention to more recent methods as well as new frontiers and challenges/opportunities for methods development in SSE.

[*] Corresponding author.

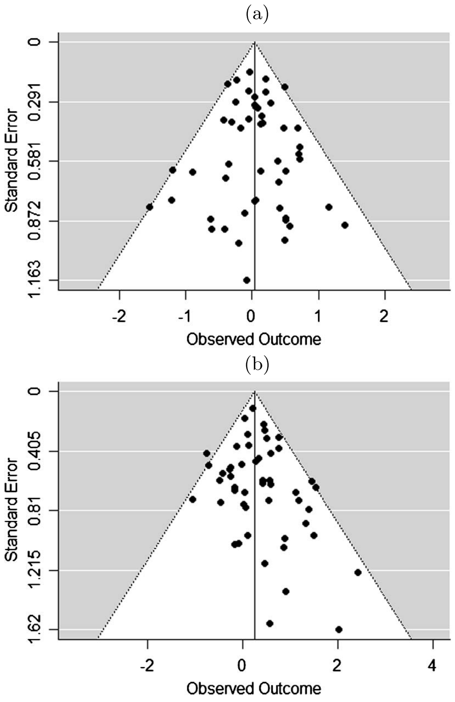

Figure 1. Simulated funnel plots with (a) no SSE present and (b) SSE present.

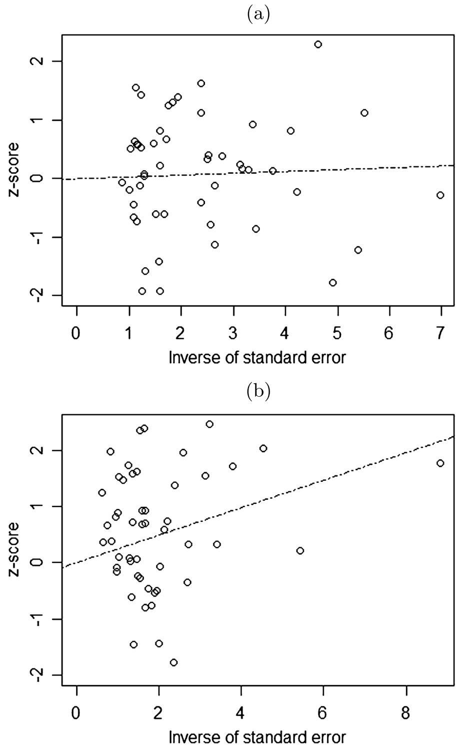

Figure 2. Simulated radial plots with (a) no SSE present and (b) SSE present.

2. GRAPH-BASED EVALUATION OF SSE

2.1 Graphing of study results

Graph-based methods to detect the presence of SSE in meta-analyses look for an association between effect size and precision among the component studies. This is most clearly illustrated in the funnel plot, which is also one of the most commonly used visualization methods among investigators. First introduced by Light and Pillemer [41], the funnel plot is a scatter plot of individual study estimates against the inverse of their standard errors. Since larger studies have greater precision, the estimates at the top of the plot will scatter more closely around the mean than those at the bottom, leading to an inverted funnel shape. In the absence of any SSE, the estimates should be scattered symmetrically about the center of the plot, as in Figure 1(a). However, when SSE are present we expect to see the smaller studies scatter around a mean that is different from the larger studies (see Figure 1(b)), with empty space on one side of the funnel plot. This leads to an asymmetric pattern trending away from the mean as precision decreases.

Another graphical tool is Galbraith's radial plot [21], which plots standardized treatment effects (z scores) on the y-axis against the inverse of their standard errors on the x-axis. In the absence of SSE, the individual estimates should be scattered randomly above and below a regression line with zero-intercept and slope equal to the overall effect-size from a fixed-effect model (see Figure 2(a)). In the presence of SSE, the true intercept for the data will be away from the origin. As Figure 2(b) shows, in the presence of SSE, the points are no longer scattered symmetrically about this line, and the true intercept may be closer to $z = 1$. Furthermore, the estimated treatment effect represented by the slope is > 0, while the true effect is 0. A regression line with intercept closer to 1 will have an attenuated slope, reflecting a treatment effect closer to the truth. Note that the asymmetry about the zero-intercept line in Figure 2(b))is similar to the asymmetry about the vertical line in Figure 1(b).

Though certainly useful as visual tools for preliminary assessment of SSE, these plots are often misinterpreted under subjective evaluation, particularly when fewer than 10 studies are included in a meta-analysis [64]. One empirical evaluation study found that respondents correctly identified asymmetry in only 52.5% of simulated funnel plots on average, as funnel plots could appear asymmetric by chance even in the absence of publication bias, and the existence of publication bias may lead to only small levels of asymmetry [71]. Furthermore, the plots are limited in helping to distinguish between possible sources of asymmetry [70]. For example, treatment effects based on dichotomous outcomes, such as odds ratios, are inherently correlated with measures of precision, since the variance is a function of the mean. Sterne et al. [69] showed that in such cases, funnel plot asymmetry may be present even in the absence of publication bias. Furthermore, subgroup effects are often misinterpreted as asymmetry, when in fact the study estimates are forming symmetric funnel shapes about different means [71, 51]. Plotting results within subgroups defined by study-level covariates can therefore aid in interpretation. Overall the funnel plot remains a useful and common tool for preliminary evaluation.

2.2 Testing approaches for SSE

Hypothesis tests have been developed to provide greater objectivity for graphical evaluation, formally assessing the evidence for asymmetry beyond random chance. Consider a meta-analysis of K studies seeking to estimate the true treatment effect θ. Each of the K studies reports an estimate of θ, denoted as y_k, and standard error, s_k, $k = 1, \ldots, K$. If SSE are present, then we expect there to be an association between y_k and s_k. Begg's rank correlation statistic [2, 3] formally tests for this through a nonparametric measure of correlation (based on Kendall's tau) between the standardized deviates $\{(y_k - \bar{y})/s_k^*\}$ and $\{s_k\}$ (or equivalently, study size n_k), where $\bar{y}$ is the weighted average of the treatment effect estimates and $(s_k^*)^2 = s_k^2 - (\sum_{j=1}^{K} 1/s_j^2)^{-1}$. Another, parametric regression test was proposed by Egger [19]. Having similar intuition to the radial plot, it tests for association through the intercept term in the model

$$E[y_k/s_k] = \beta_0 + \beta_1 1/s_k$$

$H_0 : \beta_0 = 0$ corresponds to a regression line through the origin of the radial plot. The greater the association between observed treatment effects and their standard errors, the farther above or below the origin the line moves. Thus a rejection of H_0 would offer evidence in support of SSE being present. Egger's test is the second most widely used assessment tool for SSE, second only to the funnel plot [22, 48]. Both offer an intuitive, user-friendly means of investigating the quality of meta-analytic samples. However, as with the funnel plot, its use can be limited or misinterpreted in certain settings. Egger's test and Begg's test have been shown to have low power to detect publication bias when the number of studies is small [69, 19, 20]. Furthermore, since these are tests for the association between treatment effects and standard errors, they have inflated type 1 error when binary outcomes are used [69, 59, 50], particularly if the event is rare and/or the true treatment effect is large.

In light of these limitations, alternative tests have been proposed for binary outcome data. Macaskill et al. [43] and Peters et al. [50] presented modifications to Egger's test, in which the log odds ratios are either regressed against the study sample size (n_k) or the inverse of the sample size ($1/n_k$) respectively, with weights equal to the inverse of the pooled variance. In doing so they avoid the correlation induced by binary outcome measures, therefore having the appropriate false-positive rates and allowing for some distinction between SSE due to choice of outcome and other possible causes. However, Macaskill's test has been shown to have lower power compared to Egger's test [43]. Other modifications for binary outcome data include Harbord's score-based test, where the variance measure depends only on the marginal totals from a study's 2x2 table [24], Schwarzer's rank-correlation test [60], and Rücker's tests [54] based on the variance-stabilizing arcsine transformation. Among the aforementioned testing methods, Egger's and Peter's tests were found to have the greatest power, while Harbord's and Peter's tests best maintained type 1 error [54]. However, a more recent study involving over 5,000 meta-analyses from Cochrane reviews showed substantial loss of power for p-value based methods (including Egger's, Harbord's, Peter's and Begg's tests) when the median number of studies decreased from large to moderate [20].

Another consideration is that the tests have inflated type 1 error and decreased power when between-study heterogeneity (represented by τ^2 in random-effects models) increases. To help address this, Thompson and Sharp extended Egger's test [73] and Rücker's test (for binary outcomes) to allow for heterogeneity through a multiplicative overdispersion error term [73, 54, 70]. However, they may be slightly conservative and lack power to detect SSE if τ^2 is in fact small. Attempts have also been made to account for the explainable portion of heterogeneity and distinguish it from residual (or unexplained) heterogeneity by including observed study-level covariates in extended versions of the regression tests [51]. However, their performance was limited if both explainable and residual heterogeneity were present. An alternative approach is to perform tests for SSE within subgroups, particularly if differential publication bias is suspected, however this would significantly reduce the sample size used for each test [51].

2.3 Correcting for SSE

In addition to testing procedures, bias correction methods have been developed in this area. Natural extensions of the regression-based tests (including Egger's, Harbord's and Peter's tests) have been proposed to adjust for publication bias/SSE [46]. By extrapolating the regression lines to a study of infinite sample size (or zero standard error), according to the principle of funnel plot symmetry, this study will be centered on the true mean adjusting for SSE. The 'Trim-and-Fill' method is another approach based on funnel plot symmetry that was proposed by Duval and Tweedie [16, 17]. This iterative algorithm involves rank-based estimation of the number of missing studies according to asymmetry patterns about the mean effect, followed by a 'trimming' of the most extreme observed studies and replacement with imputed values to improve symmetry. Though simple and easy to implement, this nonparametric approach has been shown to be sensitive to outliers, leading to inflated standard errors and conservative inference [61].

Both trim-and-fill and Moreno's adjustment methods are subject to the same limitations of funnel plots, leading to spurious adjustment and poor coverage probabilities in the presence of between-study heterogeneity [72, 46]. As with testing, if subgroup effects are suspected that can be explained through study-level variables, a more appropriate method for bias correction would be subgroup analysis or, if the variables are continuous, meta-regression adjusting for study-level covariates, with studies weighted by the inverse of their variances. Meta-regression may also include random-effects to account for residual heterogeneity [73, 68]. However, this may not be feasible in small samples.

3. PUBLICATION BIAS AS A MISSING DATA PROBLEM AND SELECTION MODELS

3.1 p-value based selection models

As an alternative to the more general assessment of SSE through graph-based approaches, selection models focus on publication bias by explicitly characterizing the probability of study selection into meta-analyses as a function of the significance of study results [75, 9, 8]. Hence they are also known as 'weighted distribution' models. The formulation typically consists of an outcome model describing the distribution of effect estimates in all studies ever conducted (corresponding to a standard meta-analysis of the complete data), and a second latent model outlining the selection process.

Most formulations of selection models relate publication probability to the p-value of a study and critical value for significance, α, ranging from a fully deterministic framework whereby studies are published if and only if $p \leq \alpha$ [40, 25], to more gradual weight functions based on how far the p-value deviates above α [32], to step-weight functions defined by ranges of p-values [15, 26]. These intuitive approaches are particularly useful for modeling the 'thresholding' effect of publication decisions, whereby study results just below critical values are more likely to be published (eg. $p = 0.49$ when $\alpha = 0.05$) compared to those above, and these thresholds become points of discontinuity in step-weight functions. On the other hand, there can be an oversimplicity in associating the weight functions with the p-value, which is a combined measure of effect size and precision. For example, if publication only depends on study precision and not effect size, this does not constitute publication bias or SSE. Most p-value based models also do not account for the direction of the effect size in publication mechanisms.

More recent iterations of p-value based selection models include the p-curve [66, 67] and p-uniform methods [74] for testing and correction of publication bias. Like Hedges' model [25], these approaches only consider the distribution of significant, published p-values and assume that all studies with non-significant p-values are unpublished. p-curve offers a test for true treatment effect in the presence of publication bias, which is equivalent to testing for right skewness of the p-curve, or deviation from a uniform shape (which corresponds to a null effect). p-uniform instead tests for publication bias, assuming the distribution of p-values is uniform conditional on the true effect size. By focusing on the distribution of significant results as opposed to different weight functions, estimation is made easier with these methods. However, they rely on strong and possibly untenable assumptions regarding the unpublished studies, as non-significant results also get published.

3.2 Copas' selection model

In a series of papers [10, 9, 11, 12], Copas and colleagues proposed a more flexible framework, allowing for selection to depend on both the effect size and precision through separate parameters. This enables us to both account for directionality and model processes whereby larger studies are more likely to be published irrespective of the size estimated treatment effect. The formulation consists of a standard random effects model for all M $(M \geq K)$ studies conducted in a particular area, including those not observed in the meta-analysis,

$$\begin{gathered} y_m = \theta_m + \sigma_m \epsilon_m, \\ \epsilon_m \sim N(0,1), \\ \theta_m \sim N(\theta, \tau^2), \\ m = 1, \ldots, M \geq K \end{gathered}$$

In this formulation, each study's estimate y_m is centered on the study-specific mean θ_m, which itself varies randomly across studies with true mean equal to θ. The parameter τ^2 represents the degree of heterogeneity among the studies. The number of studies conducted will always be greater than or equal to the number observed in the meta-analysis.

A separate model is given for the latent variable z_m, where study m is published if and only if $z_m > 0$.

$$\begin{gathered} z_m = a + b/s_m + \delta_m, \\ \delta_m \sim N(0,1), \\ \mathrm{corr}(\delta_m, \epsilon_m) = \rho \end{gathered}$$

The two models are related through a single correlation parameter, ρ, between their error terms ϵ_m and δ_m. According to this model, $\rho > 0$ suggests a greater probability of selection for more positive treatment effects among studies with similar precision, while $\rho < 0$ indicates selection probability increases with more negative treatment effects. Thus the K observed treatment effects follow a conditional distribution, $f(y_k|z_k > 0)$, where $f(y_k|z_k > 0) = f(y_k)$ when $\rho = 0$. The selection model parameters a and b control the marginal rate of publication. a represents the mean value of z_m for a study m with infinite variance, while b relates publication probability to a study's precision.

A key property is the distinction between the true within-study variance, $\sigma_m^2 = var(y_m|\theta_m)$ and the conditional observed within-study variance, s_m^2, and the inclusion of the latter in the selection model but not the mean model. This avoids any spurious correlation between y_m and s_m^2 due to choice of outcome or clinical heterogeneity, as here they are only related through the publication bias mechanism (i.e. when $\rho \neq 0$), unlike the regression models described in Section 2.2.

A limitation of most selection models is the lack of data from unobserved studies to inform estimation of the weights or selection model parameters, and often a sensitivity analysis approach is recommended. For example, Copas and Li [10] proposed sensitivity analysis to assess the robustness of meta-analysis conclusions to different values of a and b, reflecting various rates of non-publication. Given fixed values of a and b, the remaining parameters (θ, τ, ρ) can be estimated through maximizing the profile likelihood,

$$L_{a,b}(\theta) = \max_{\theta,\rho,\tau|a,b} L(\theta, \rho, \tau, a, b).$$

One can then assess the sensitivity of θ to varying combinations of a and b. In an effort to guide the choice of the selection model parameters, Copas and Shi [11] propose choosing a and b to reflect a range of values for the marginal probability of selection for a study with standard error s, where

$$P(z > 0|s, a, b) = \Phi(a + b/s)$$

Though intuitive to some degree, this indirect relation between probability of publication and the model parameters can be seen as unnecessarily complex and discourages the model's use in practice.

Alternative solutions to sensitivity analysis have included trading model accuracy to improve identifiability. For instance, instead of stepped-wedge selection models that require estimation of weights for several ranges of the p-value, a beta-approximation was proposed that requires estimation of only two parameters [27, 8]. Bayesian approaches have also been developed to aid estimation of the model parameters [23, 65]. Mavridis et al. [45] proposed a Bayesian framework for an extended Copas model applied to the network meta-analysis setting, which involves multiple treatment comparisons. In similar fashion to Copas and Shi's approach to sensitivity analysis, they place priors on $P(z > 0|s_{large}, a, b)$ and $P(z > 0|s_{small}, a, b)$, rather than on a and b directly. These quantities correspond to the probability of publication for studies with the largest and smallest standard errors respectively in a meta-analysis. With priors placed on the remaining model parameters, inference on the true treatment effect can then be made using Markov Chain Monte-Carlo (MCMC) draws from the posterior distribution, as opposed to maximizing the complex likelihood. The use of prior knowledge in obtaining an adjusted treatment effect makes the Bayesian approach an attractive alternative to sensitivity analysis.

In the absence of prior information, a frequentist approach to data augmentation was proposed by Ning et al. [47]. By incorporating the symmetry assumption of funnel plots, they impute the missing counterparts to the observed studies through an Expectation Maximization (EM) algorithm, which converges to a final adjusted estimate of the treatment effect under Copas' model. This has shown promising results for moderately sized meta-analyses. Its dependence on the Copas model additionally makes it more robust to outliers than trim-and-fill [61]. However, the incorporation of the symmetry assumption may be controversial if publication bias does not necessarily induce asymmetry among the observed studies. For example, if $b = 0$ in Copas' selection model, then the publication process will be unrelated to study precision, but may still depend on the treatment effect. Since Copas' model also assumes marginal independence of y_m and s_m^2, this will induce a funnel plot that is symmetric about a biased estimate treatment effect, and the EM-algorithm will not lead to adjustment.

Overall, selection models may be preferred over graph-based approaches for handling publication bias specifically. However, as they are more sophisticated than graph-based methods, they are less accessible and therefore not as popular. Recent developments of statistical packages such as `metasens` [63] and `selecMeta` [55] may serve to increase their use in practice.

4. FUTURE WORK AND DIRECTION

4.1 Outcome reporting bias (ORB)

Relative to publication bias, less attention has been given to the selective reporting of outcomes within trials, which can also give rise to SSE and is no less problematic [38, 29]. Evidence has shown that significant outcomes are more likely to be fully reported compared to nonsignificant outcomes [18], while secondary endpoints tend to be underreported compared to primary endpoints [44].

An Outcome Reporting Bias in Trials (ORBIT) study aimed to estimate the prevalence and impact of ORB on meta-analyses for benefit outcomes [38] followed by a later study for harm outcomes [56]. To achieve this the authors developed a classification tool to assess the risk of ORB when a trial is excluded from a meta-analysis. Their method uses the information contained within the published trial report, along with expert opinion, to determine whether the absence or underreporting of the outcome is at high, low, or no risk of being due to reporting bias on the part of the investigators. Such a risk assessment considers the likelihood that the unreported outcome was measured and/or analyzed, in which case the failure to report results may be due to bias against the analyzed outcome. Separate criteria are developed for benefit and harm outcomes, since trialists are more likely to suppress a significant harm profile, in contrast to the preference for significant results for efficacy outcomes. Through contacting trialists directly, they showed that the tool had high sensitivity and specificity for detecting bias. However, their validation assumes no response bias from the trialists, which would likely lead to underestimation of sensitivity and overestimation of specificity [38].

Copas et al. [13] proposed a likelihood-based sensitivity analysis using the ORBIT classification system. It incorporates two sensitivity parameters defining the quality of risk assessment. They include ρ_1, the probability that a measured, non-significant and unreported outcome is classified correctly as high risk, and ρ_2, the probability that an unmeasured and unreported outcome is correctly identified as low risk. Under perfect risk assessment, $\rho_1 = \rho_2 = 1$. Inference for θ proceeds by maximizing the profile log-likelihood given (ρ_1, ρ_2). More recently they offered a simplified approach under the assumption of perfect risk assessment, using a modified likelihood that includes studies with missing/underreported outcomes that were identified as high risk by ORBIT assessment [14]. For efficacy endpoints, the underreported studies are included in the likelihood function through the probability that their values are non-significant. The formulation is easily extended harm endpoints, where the underreported studies contribute to the likelihood through the probability that $y_m > 0$ (i.e. the treatment increases the risk of the adverse event).

A limitation is that the likelihood functions include σ_m, the true within-study variance for study m, and this is generally not observed for unreported outcomes. The authors propose imputing σ_m using the total sample size within a study and a proportionality constant estimated using the observed studies. This of course assumes that, on average, the studies with reported and unreported outcomes have similar designs.

More work is needed in this area, particularly ones that can jointly model publication and selective reporting processes, as these mechanisms each contribute to missingness of results in meta-analyses.

4.2 SSE in multivariate and network meta-analysis

Multivariate and network meta-analysis (MMA and NMA) are more recent extensions of the univariate framework [33]. MMA simultaneously models multiple outcomes, accounting within-study correlation, and NMA models a network of treatment comparisons, combining direct and indirect effects against common comparators [42, 1, 57]. With the added complexity involved with modeling MMAs and NMAs, methods for SSE and publication bias remain understudied in these settings.

For MMAs, detection of SSE is complicated by the fact that not only might studies be selectively published, but outcomes within a study may be differentially impacted. For example, safety endpoints may be underreported and the relative risk biased closer to the null, while efficacy estimates tend to be biased away from the null. One should also account for the correlation between multiple outcomes in a study in order to increase power in any testing procedure, as opposed to separately testing for SSE for each outcome. However, correlation is often unreported in multi-outcome studies [52, 6]. Simulation studies have shown that when the outcomes are moderately or strongly correlated, performing MMA (using a modified model that doesn't require specification of within-study correlation) can reduce the impact of ORB within trials [37, 30]. However, they assume that at least one outcome is always fully reported, and found that MMA may slightly increase the bias of the estimated treatment effect for the fully reported outcome. Fitting a standard MMA model also does not account for selective publishing of entire studies.

Hong et al. [28] recently developed a score test to detect SSE in MMAs. Their approach accounts for the multivariate nature of MMAs, thereby improving power over univariate methods. To overcome the underreporting of within-study correlation, they base inference on the pseudolikelihood, as proposed by Chen et al. [5]. Since it is a natural extension of Egger's test (they are equivalent when the number of outcomes equals 1), it has similar strengths and limitations. In particular, type 1 error may be inflated when the outcome is binary. To reduce the correlation between the treatment effect and precision for binary data, the authors proposed a smoothed version of the test, where s_k^2 is replaced by a smoothed variance (first proposed by Jin et al. [34] for testing for publication bias in meta-analyses of observational studies). Since the existing methods for MMA are limited to graph-based approaches, there is room for future work on selection modeling in the MMA setting.

NMAs are arguably more complex than MMAs, due to the inclusion of multiple study designs and treatments. Unlike univariate and multivariate MAs which restrict analysis to a single treatment comparison, it is difficult to graphically display a network of direct and indirect contrasts, let alone provide a reference line from which to assess asymmetry and thus visually detect SSE. Chaimani et al. [4] proposed a 'comparison-adjusted' funnel plot, where effect estimates are centered on their comparison-specific means and plotted against the inverse of their standard errors. However, it requires investigators to first develop a meaningful ordering of the treatments and make assumptions regarding the direction of SSE. It also doesn't visualize indirect effects, which can form the entire base of evidence for novel treatment comparisons. Rather, this strategy offers an overall aggregated evidence of SSE, equivalent to evaluating separate funnel plots for each univariate contrast. Further work is needed to detect and visualize evidence of SSE in the NMA setting.

The difficulty in developing graph-based approaches for NMAs has led to a focus on selection models for detecting publication bias. Copas' model in particular has been extended to the NMA framework [7], with each study design having its own selection model and design-specific parameters. An obvious challenge to fitting this model is that model complexity increases with the size of the network, leading to identifiability issues even with a large number of studies. Mavridis et al. [45] proposed a Bayesian approach to improve identifiability in this setting, though this relies on prior knowledge. There remains a need for frequentist alternatives that do not require prior information, but future developments should also consider more parsimonious models in the NMA setting. A simulation study can shed light on whether separate selection models are needed for each study design, to reflect differential reporting bias, or if they can be collapsed to reduce the number of model parameters.

4.3 Alternatives to Copas' selection model

Copas' model provides greater flexibility in modeling the selective publishing process compared to earlier formulations. However, this flexibility comes with the added cost of model complexity. Intractability of the likelihood and convergence issues are encountered even in the univariate setting, making extensions to MMA and NMA settings far less feasible. It is also difficult to build upon this model to include additional sources of publishing bias. For example, industry sponsored trials are less likely to be published than investigator-sponsored trials [35, 36]. In NMA's, multi-arm trials may have a greater probability of publication compared to two-arm studies. Mavridis et al. [45] proposed including such variables in the model for z_m, however this would add yet more non-identifiable parameters to the framework. Bayesian approaches may be flexible enough to allow for such additions and still enable identifiability, but more parsimonious models should be considered to increase ease-of-use among clinicians.

Furthermore, Copas' model may not be intuitive enough for investigators. The parameter a in the latent variable model is interpreted as the marginal publication rate for a study with infinite variance, which is not a meaningful quantity in practice. The indirect means of selecting a and b for sensitivity analysis (through the inverse relation with $P(z > 0|s, a, b)$) can also be a barrier to use. A more direct formulation with meaningful parameters can increase utility. Another drawback is it postulates a continuous distribution for the latent variable model, which does not reflect any sense of thresholding at a critical value, a property seen in earlier selection models.

Finally, the distinction between σ_m^2, the true within-study variance, and s_m^2, the observed variance of y_m conditional on the study being published, makes model-based data generation difficult, as investigators do not condition estimates on study publication. In fact, existing estimation approaches require a modification to the outcome model, where σ_m^2 is replaced by s_m^2 [47, 45], under the assumption that they are approximately equal if all the component studies are large. This also reduces the number of unknown nuisance parameters in the models, since σ_m^2 is study specific. However, not only might this approximation be invalid if some studies are small, but the modification negates a key property of Copas' model that distinguishes it from graph-based approaches, which is the independence of y_m and s_m^2 apart from publication bias. Such a modification no longer avoids other sources of correlation. Therefore, while Copas is presented as a working model, alternative formulations can improve both interpretation and tractability for more widespread use in meta-analyses.

4.4 Distinction between sources of SSE

Much of the existing work has focused on detecting the presence of SSE and publication bias, with less discussion surrounding how to proceed with meta-analyses if they exist. We believe such a discussion should give weight to understanding the sources of SSE and the relationships between them. For example, publication bias does not always give rise to SSE, since studies with similar effect estimates could have the same probability of censoring regardless of precision. In such cases symmetry-based approaches for correcting for bias like trim-and-fill and the EM-algorithm for Copas' model will not lead to adjusted estimates. Perhaps sensitivity analysis using Copas' model could be guided by visual assessment of funnel plots. If publication bias is suspected but asymmetry is not present, parameter b could be fixed at 0 in the profile likelihood for sensitivity analysis, or priors placed on b that are heavily weighted at 0 in a Bayesian analysis.

Additionally, not all sources of SSE require correction. If SSE are due to clinical heterogeneity, such as more severe patients being recruited in smaller studies and exhibiting stronger responses to treatment, then a subgroup analysis by clinical severity is more appropriate than correction. This also improves the generalizability of results. Similarly, when SSE are induced by the choice of outcome alone, or by chance, then the estimate is unbiased and does not need correction.

A more complex, but common, phenomenon occurs when publication bias and other sources of SSE simultaneously impact the results of a meta-analysis. Copas' model can aid in correcting for bias due to selective publication alone, but asymmetry may still remain due to other sources of correlation. Recommendations should be made on how to proceed in this setting.

5. CONCLUSION

With increasing awareness of the presence and impact of SSE, significant advances have been made in detecting and correcting for bias in meta-analyses. Here we have provided a critical review of the more commonly used methods and recent developments. Even with decades of methodological work, this remains an active research area in statistics with substantial room for improvement and innovation. There is particularly a need for more intuitive, flexible selection models that can be readily implemented and extended to more complex meta-analytic frameworks, as well as more robust testing procedures.

Alongside efforts to mitigate the effects of SSE (and especially publication bias) at the analysis stage, emphasis should be placed on prevention of selective reporting through proper registration and reporting of outcomes at the trial stage. Since any approach to missing data relies on untestable assumptions regarding the missing studies, we gain far more by eliminating the missing-not-at-random process altogether than trying to model it. A concerted effort at both the study-level and meta-analytic level can lead to valid inference and unbiased recommendations for treatment decision-making and best practices in the medical community.

ACKNOWLEDGEMENTS

We thank Dr. Rui Duan (Harvard University) for her contribution to the discussion of this paper. A. Marks-Anglin's effort was supported in part by NIH grant 1R01LM012607. Y. Chen's effort was supported in part by NIH grants 1R01LM012607 and 1R01AI130460.

Received 12 December 2019

REFERENCES

[1] Ades, A. (2003). A chain of evidence with mixed comparisons: models for multi-parameter synthesis and consistency of evidence. *Statistics in Medicine* **22** 2995–3016.

[2] Begg, C. B. (1994). Publication bias. *The Handbook of Research Synthesis* **25** 299–409.

[3] Begg, C. B. and Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics* 1088–1101.

[4] Chaimani, A., Higgins, J. P., Mavridis, D., Spyridonos, P. and Salanti, G. (2013). Graphical tools for network meta-analysis in STATA. *PloS One* **8** e76654.

[5] Chen, Y., Hong, C. and Riley, R. D. (2015). An alternative pseudolikelihood method for multivariate random-effects meta-analysis. *Statistics in Medicine* **34** 361–380. MR3301582

[6] Chen, Y., Cai, Y., Hong, C. and Jackson, D. (2016). Inference for correlated effect sizes using multiple univariate meta-analyses. *Statistics in Medicine* **35** 1405–1422. MR3513459

[7] Chootrakool, H., Shi, J. Q. and Yue, R. (2011). Meta-analysis and sensitivity analysis for multi-arm trials with selection bias. *Statistics in Medicine* **30** 1183–1198. MR2828931

[8] Citkowicz, M. and Vevea, J. L. (2017). A parsimonious weight function for modeling publication bias. *Psychological Methods* **22** 28.

[9] Copas, J. (1999). What works?: Selectivity models and meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **162** 95–109.

[10] Copas, J. B. and Li, H. (1997). Inference for non-random samples. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **59** 55–95. MR1436555

[11] Copas, J. and Shi, J. Q. (2000). Meta-analysis, funnel plots and sensitivity analysis. *Biostatistics* **1** 247–262.

[12] Copas, J. and Shi, J. Q. (2001). A sensitivity analysis for publication bias in systematic reviews. *Statistical Methods in Medical Research* **10** 251–265.

[13] Copas, J., Dwan, K., Kirkham, J. and Williamson, P. (2014). A model-based correction for outcome reporting bias in meta-analysis. *Biostatistics* **15** 370–383.

[14] Copas, J., Marson, A., Williamson, P. and Kirkham, J. (2019). Model-based sensitivity analysis for outcome reporting bias in the meta analysis of benefit and harm outcomes. *Statistical Methods in Medical Research* **28** 889–903. MR3922897

[15] Dear, K. B. and Begg, C. B. (1992). An approach for assessing publication bias prior to performing a meta-analysis. *Statistical Science* 237–245.

[16] Duval, S. and Tweedie, R. (2000a). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association* **95** 89–98. MR1803144

[17] Duval, S. and Tweedie, R. (2000b). Trim and fill: a simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics* **56** 455–463.

[18] Dwan, K., Gamble, C., Williamson, P. R. and Kirkham, J. J. (2013). Systematic review of the empirical evidence of study publication bias and outcome reporting bias—an updated review. *PloS One* **8** e66844.

[19] Egger, M., Smith, G. D., Schneider, M. and Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ* **315** 629–634.

[20] Furuya-Kanamori, L., Xu, C., Lin, L., Doan, T., Chu, H., Thalib, L. and Doi, S. A. (2020). P value–driven methods were underpowered to detect publication bias: analysis of Cochrane review meta-analyses. *Journal of Clinical Epidemiology* **118** 86–92.

[21] Galbraith, R. (1988). Graphical display of estimates having differing standard errors. *Technometrics* **30** 271–281.

[22] Gerber, S., Tallon, D., Trelle, S., Schneider, M., Jüni, P. and Egger, M. (2007). Bibliographic study showed improving methodology of meta-analyses published in leading journals 1993–2002. *Journal of Clinical Epidemiology* **60** 773–780.

[23] Givens, G. H., Smith, D. and Tweedie, R. (1997). Publication bias in meta-analysis: a Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. *Statistical Science* 221–240.

[24] Harbord, R. M., Egger, M. and Sterne, J. A. (2006). A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Statistics in Medicine* **25** 3443–3457. MR2252403

[25] Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics* **9** 61–85.

[26] Hedges, L. V. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science* 246–255.

[27] Hedges, L. V. and Vevea, J. L. (1996). Estimating effect size under publication bias: Small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics* **21** 299–332.

[28] Hong, C., Salanti, G., Morton, S., Riley, R., Chu, H., Kimmel, S. E. and Chen, Y. (2018). Testing small study effects in multivariate meta-analysis. *arXiv preprint arXiv:1805.09876*.

[29] Howard, B., Scott, J. T., Blubaugh, M., Roepke, B., Scheckel, C. and Vassar, M. (2017). Systematic review: Outcome reporting bias is a problem in high impact factor neurology journals. *PLoS One* **12** e0180986.

[30] Hwang, H. and DeSantis, S. M. (2018). Multivariate network meta-analysis to mitigate the effects of outcome reporting bias. *Statistics in Medicine* **37** 3254–3266. MR3853281

[31] Ioannidis, J. P. and Lau, J. (1997). The impact of high-risk patients on the results of clinical trials. *Journal of Clinical Epidemiology* **50** 1089–1098.

[32] Iyengar, S. and Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science* 109–117.

[33] Jackson, D., Riley, R. and White, I. R. (2011). Multivariate meta-analysis: potential and promise. *Statistics in Medicine* **30** 2481–2498. MR2843472

[34] Jin, Z.-C., Wu, C., Zhou, X.-H. and He, J. (2014). A modified regression method to test publication bias in meta-analyses with binary outcomes. *BMC Medical Research Methodology* **14** 132.

[35] Jones, C. W., Handler, L., Crowell, K. E., Keil, L. G., Weaver, M. A. and Platts-Mills, T. F. (2013). Non-publication of large randomized clinical trials: cross sectional analysis. *BMJ (Clinical research ed.)* **347** f6104.

[36] Kasenda, B., von Elm, E., You, J., Blümle, A., Tomonaga, Y., Saccilotto, R., Amstutz, A., Bengough, T., Meerpohl, J. J.,

Stegert, M., Tikkinen, K. A. O., Neumann, I., Carrasco-Labra, A., Faulhaber, M., Mulla, S. M., Mertz, D., Akl, E. A., Bassler, D., Busse, J. W., Ferreira-González, I., Lamontagne, F., Nordmann, A., Gloy, V., Raatz, H., Moja, L., Rosenthal, R., Ebrahim, S., Schandelmaier, S., Xin, S., Vandvik, P. O., Johnston, B. C., Walter, M. A., Burnand, B., Schwenkglenks, M., Hemkens, L. G., Bucher, H. C., Guyatt, G. H. and Briel, M. (2014). Prevalence, characteristics, and publication of discontinued randomized trials. *JAMA* **311** 1045–1051.

[37] Kirkham, J. J., Riley, R. D. and Williamson, P. R. (2012). A multivariate meta-analysis approach for reducing the impact of outcome reporting bias in systematic reviews. *Statistics in Medicine* **31** 2179–2195. MR2967921

[38] Kirkham, J. J., Dwan, K. M., Altman, D. G., Gamble, C., Dodd, S., Smyth, R. and Williamson, P. R. (2010). The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. *BMJ* **340**.

[39] Kjaergard, L. L., Villumsen, J. and Gluud, C. (2001). Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. *Annals of Internal Medicine* **135** 982–989.

[40] Lane, D. M. and Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology* **31** 107–112.

[41] Light, R. J. and Pillemer, D. B. (1984). Summing up; the science of reviewing research.

[42] Lumley, T. (2002). Network meta-analysis for indirect treatment comparisons. *Statistics in Medicine* **21** 2313–2324.

[43] Macaskill, P., Walter, S. D. and Irwig, L. (2001). A comparison of methods to detect publication bias in meta-analysis. *Statistics in Medicine* **20** 641–654.

[44] Matthews, J. H., Bhanderi, S., Chapman, S. J., Nepogodiev, D., Pinkney, T. and Bhangu, A. (2016). Underreporting of secondary endpoints in randomized trials. *Annals of Surgery* **264** 982–986.

[45] Mavridis, D., Sutton, A., Cipriani, A. and Salanti, G. (2013). A fully Bayesian application of the Copas selection model for publication bias extended to network meta-analysis. *Statistics in Medicine* **32** 51–66. MR3017883

[46] Moreno, S. G., Sutton, A. J., Ades, A., Stanley, T. D., Abrams, K. R., Peters, J. L. and Cooper, N. J. (2009). Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Medical Research Methodology* **9** 2.

[47] Ning, J., Chen, Y. and Piao, J. (2017). Maximum likelihood estimation and EM algorithm of Copas-like selection model for publication bias correction. *Biostatistics* **18** 495–504. MR3799591

[48] Papageorgiou, S. N., Papadopoulos, M. A. and Athanasiou, A. E. (2014). Reporting characteristics of meta-analyses in orthodontics: methodological assessment and statistical recommendations. *European Journal of Orthodontics* **36** 74–85.

[49] Papageorgiou, S. N., Tsiranidou, E., Antonoglou, G. N., Deschner, J. and Jäger, A. (2015). Choice of effect measure for meta-analyses of dichotomous outcomes influenced the identified heterogeneity and direction of small-study effects. *Journal of Clinical Epidemiology* **68** 534–541.

[50] Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R. and Rushton, L. (2006). Comparison of two methods to detect publication bias in meta-analysis. *JAMA* **295** 676–680.

[51] Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., Rushton, L. and Moreno, S. G. (2010). Assessing publication bias in meta-analyses in the presence of between-study heterogeneity. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **173** 575–591. MR2758732

[52] Riley, R. D., Thompson, J. R. and Abrams, K. R. (2008). An alternative model for bivariate random-effects meta-analysis when the within-study correlations are unknown. *Biostatistics* **9** 172–186.

[53] Rücker, G., Carpenter, J. R. and Schwarzer, G. (2011). Detecting and adjusting for small-study effects in meta-analysis. *Biometrical Journal* **53** 351–368. MR2897405

[54] Rücker, G., Schwarzer, G. and Carpenter, J. (2008). Arcsine test for publication bias in meta-analyses with binary outcomes. *Statistics in Medicine* **27** 746–763. MR2418511

[55] Rufibach, K. (2011). Selection models with monotone weight functions in meta analysis. *Biometrical Journal* **53** 689–704.

[56] Saini, P., Loke, Y. K., Gamble, C., Altman, D. G., Williamson, P. R. and Kirkham, J. J. (2014). Selective reporting bias of harm outcomes within studies: findings from a cohort of systematic reviews. *BMJ* **349** g6501.

[57] Salanti, G., Ades, A. and Ioannidis, J. P. (2011). Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. *Journal of Clinical Epidemiology* **64** 163–171.

[58] Schulz, K. F., Chalmers, I., Hayes, R. J. and Altman, D. G. (1995). Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* **273** 408–412.

[59] Schwarzer, G., Antes, G. and Schumacher, M. (2002). Inflation of type I error rate in two statistical tests for the detection of publication bias in meta-analyses with binary outcomes. *Statistics in Medicine* **21** 2465–2477. MR2339170

[60] Schwarzer, G., Antes, G. and Schumacher, M. (2007). A test for publication bias in meta-analysis with sparse binary data. *Statistics in Medicine* **26** 721–733. MR2339170

[61] Schwarzer, G., Carpenter, J. and Rücker, G. (2010). Empirical evaluation suggests Copas selection model preferable to trim-and-fill method for selection bias in meta-analysis. *Journal of Clinical Epidemiology* **63** 282–288.

[62] Schwarzer, G., Carpenter, J. R. and Rücker, G. (2015). Small-study effects in meta-analysis. In *Meta-analysis with R* 107–141. Springer.

[63] Schwarzer, G., Carpenter, J. and Rüker, G. (2019). Metasens: advanced statistical methods to model and adjust for bias in meta-analysis. *R package version 0.4-0*.

[64] Sedgwick, P. and Marston, L. (2015). How to read a funnel plot in a meta-analysis. *BMJ* **351** h4718.

[65] Silliman, N. P. (1997). Hierarchical selection models with applications in meta-analysis. *Journal of the American Statistical Association* **92** 926–936. MR1482123

[66] Simonsohn, U., Nelson, L. D. and Simmons, J. P. (2014a). P-curve: a key to the file-drawer. *Journal of Experimental Psychology: General* **143** 534.

[67] Simonsohn, U., Nelson, L. D. and Simmons, J. P. (2014b). p-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science* **9** 666–681.

[68] Stanley, T. D. and Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods* **5** 60–78.

[69] Sterne, J. A., Gavaghan, D. and Egger, M. (2000). Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology* **53** 1119–1129.

[70] Sterne, J. A., Sutton, A. J., Ioannidis, J. P., Terrin, N., Jones, D. R., Lau, J., Carpenter, J., Rücker, G., Harbord, R. M., Schmid, C. H. et al. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ* **343**.

[71] Terrin, N., Schmid, C. H. and Lau, J. (2005). In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. *Journal of Clinical Epidemiology* **58** 894–901.

[72] Terrin, N., Schmid, C. H., Lau, J. and Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine* **22** 2113–2126.

[73] Thompson, S. G. and Sharp, S. J. (1999). Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in Medicine* **18** 2693–2708.

[74] Van Assen, M. A., van Aert, R. and Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods* **20** 293.

[75] Vevea, J. L. and Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika* **60** 419–435.

Arielle Marks-Anglin
Department of Biostatistics, Epidemiology and Informatics
University of Pennsylvania
Philadelphia, PA
USA
E-mail address: anglinar@pennmedicine.upenn.edu

Yong Chen
Department of Biostatistics, Epidemiology and Informatics
University of Pennsylvania
Philadelphia, PA
USA
E-mail address: ychen123@pennmedicine.upenn.edu