A Historical Review of Publication Bias

Arielle Marks-Anglin^{1*} and Yong Chen¹

August 23rd, 2020

Abstract

Publication bias is a well-known threat to the validity of meta-analyses and, more broadly, the reproducibility of scientific findings. When policies and recommendations are predicated on an incomplete evidence-base, it undermines the goals of evidence-based decision-making. Great strides have been made in the last fifty years to understand and address this problem, including calls for mandatory trial registration and the development of statistical methods to detect and correct for publication bias. We offer an historical account of seminal contributions by the evidence synthesis community, with an emphasis on the parallel development of graph-based and selection model approaches. We also draw attention to current innovations and opportunities for future methodological work.

Keywords: evidence-based medicine, meta-analysis, publication bias, selection bias, reproducibility

¹ Author affiliation: Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania

* Corresponding author: University of Pennsylvania School of Medicine
423 Guardian Drive
Philadelphia, PA 19104
anglinar@pennmedicine.upenn.edu

1 Introduction

Publication bias (PB), defined as the selective publishing of research based on the nature and direction of findings, occurs when studies with significant or favorable results are more likely to be published than those with non-significant or unfavorable findings. Not only does this phenomenon waste resources invested in scientific research (Freedman *et al.*, 2015), it also poses a major threat to the reproducibility of scientific findings from meta-analysis (Johnson *et al.*, 2017) and the practice of evidence-based medicine (Crawford *et al.*, 2010; Murad *et al.*, 2018). Policy- and decision-making that rely on recommendations from systematic reviews and meta-analyses may be particularly impacted by the selective dissemination of findings, as recommendations will be based on a biased sample of the studies conducted.

In this paper, we offer an historical overview of important developments in the study, prevention, assessment and correction of PB in biomedical research, and attempt to highlight the key challenges of addressing this systemic threat. The book, *Publication Bias in Meta-Analysis*, by Rothstein *et al.* (2005) provides a comprehensive coverage of research leading up to the early 2000s, followed by a review by Song *et al.* (2010). However, much progress has been made since then. Our hope is that this updated review can highlight the contributions made by many great researchers in the past decades and provide perspective and motivation for future methodological work in this area. It should be noted that given the large number of literatures on the topic of PB, our review is not exhaustive and may not include all existing related work, but rather seeks to critically appraise major advancements within their historic context. We also acknowledge that there are other forms of dissemination bias such as outcome reporting bias, citation bias and time lag bias that similarly skew the results of meta-analyses. These are only briefly highlighted in our work, with the focus being PB. A helpful discussion on reporting biases in general can be found in Song *et al.* (2010) and Dickersin and Chalmers (2011), as well as a comprehensive reference list of methods for dissemination bias by Mueller *et al.* (2016).

In Section 2, we provide a brief history on the evidence of PB and preventive measures, followed by early methodological work in Section 3. In Sections 4 and 5 we give an overview of ‘graph-based’ approaches and ‘selection models’ respectively, which account for the great majority of methods for handling PB. In Section 6 we present more recent developments and directions for future work, and finally in Section 7 we list software packages that have been released for implementing PB methods. We offer a timeline illustration in Figure 1 that highlights major milestones from 1979 to present, including the seminal papers on funnel plot, Egger's test, Trim-and-Fill and selection models, with their chronicle order.

2 A brief history on discourse, evidence and prevention of publication bias

Summary points

- Selective publishing practices have been influenced by negative perceptions regarding statistical non-significance
- Cohort studies of approved and grant-funded projects offer clear evidence of the presence and prevalence of PB
- Trial registries were created to encourage transparency and reproducibility of findings. They have also been used to study the impact of PB on meta-analytic results
- Inadequate registration of trials, along with the absence of a central database for accessing all studies, warrants the use of post-hoc methods for assessing and correcting for potential PB in meta-analyses

The term ‘publication bias’ started appearing in scientific literature around the 1980s (Smith, 1980; Begg, 1985). However, documented awareness of a preference for the dissemination of ‘successful’ research, or sharing of knowledge that will be received ‘favorably’, dates much earlier (Editorial, 1909; Dickersin, 2005). It can be argued that this phenomenon was not discovered in the late 20th century, but rather the accumulation of studies over time, coupled with the development of research synthesis methods, allowed for formal study of its existence and effects.

Empirical findings from the field of psychology provided the driving force for early research on PB, beginning with Sterling’s (1959) observations of a severe overrepresentation of published studies rejecting the null hypothesis and a corresponding lack of replicated studies (a phenomenon which persisted over 30 years later in his updated review (1995)). This was followed by additional evidence from Smart (1964) and Bozarth and Roberts (1972), among others. In 1975, Greenwald produced a condemning review of prevailing cultural attitudes on the null hypothesis among behavioral scientists, including the belief that only findings that rejected the null could advance science, while a null result only conferred uncertainty (Greenwald, 1975). Indeed, the popularization of p-values (Kennedy-Shaffer, 2019) in the early 20th century led to a misuse of and overdependence on tests for significance by the scientific community (Bakan, 1966, Ioannidis, 2005; Johnson *et al.*, 2017). Furthermore, the uncertainty associated with failure to reject the null, as it cannot be ‘proved or established’ (*Design of Experiments*, Fisher, 1949), contributed towards prejudice against null findings (Angell, 1989), leading to an abundance of false-positive, non-replicable results that persists in the 21st century (Szucs and Ioannidis, 2017).

Early awareness of selective publishing practices was later supported by cohort studies of biomedical research. These looked at the eventual publication status of projects that received ethics approval by review boards (Easterbrook *et al.*, 1991; Dickersin *et al.*, 1992;

Stern and Simes, 1997; Cooper *et al.*, 1997; Decullier 2005), externally-funded projects (Dickersin and Min, 1993; Ioannidis, 1998; Franco *et al.*, 2014; Driessen *et al.*, 2015), reports to licensing authorities (Bardy, 1998) and conference abstracts (Scherer *et al.*, 2018). The studies, together with meta-analytic reviews pooling their findings (Hopewell *et al.*, 2009; Song *et al.*, 2009; Dwan *et al.*, 2013; Schmucker *et al.*, 2014), showed that projects with ‘positive’ or significant results were more likely to be published than those with nonsignificant results. They also shed light on additional forms of bias in the publishing process, including time-lag bias, in which the time to publication is shorter for studies with favorable results (Sterne and Simes, 1997; Ioannidis, 1998; Hopewell *et al.*, 2005; Jones *et al.*, 2013; Kasenda *et al.*, 2014); citation bias, where trials showing favorable results are cited more frequently in papers than those showing unfavorable results (Nieminen *et al.*, 2007); and sponsorship (Easterbrook *et al.*, 1991; Dickersin *et al.*, 1992; Lexchin *et al.*, 2003; Jones *et al.*, 2013; Kasenda *et al.*, 2014), where industry-sponsored studies are less likely to be published than government-funded research. Together these findings offer a picture of the prevalence and sources of bias in the literature. Evidence of impact, on the other hand, has been produced with the aid of trial registries.

Trial registries were developed as a preventive measure against PB and, relatedly, outcome reporting bias, which is caused by the selective reporting of outcomes *within* studies. Prospective registration of trials before results became available would ensure that relevant information is publicly available regardless of eventual publication status. Over the last two decades, regulatory and funding agencies have made steps to mandate trial registration, including the International Committee of Medical Journal Editors (ICMJE) (De Angelis *et al.*, 2005), the U.S. Food and Drug Administration (FDA) (FDA, 2007; Zarin *et al.*, 2016), the National Institutes of Health (NIH) (Hudson *et al.*, 2016), among others. The availability of unpublished data allows for the most complete evidence base for conducting systematic reviews and meta-analyses (Simes, 1986). This has also enabled key studies revealing and quantifying the impact of PB on meta-analytic results (Simes, 1986; Turner *et al.*, 2008; Schmucker *et al.*, 2017), some of which found larger pooled treatment effects in the published data compared to the complete set of studies.

Registries alone are not sufficient for combating PB. While registry databases are plentiful now, grey literature was not readily available twenty years ago, and many meta-analyses include studies spanning decades of work. Even after mandates for prospective registration, a large proportion of published trials are unregistered, retrospectively registered, or inadequately report outcomes (Mathieu *et al.*, 2009; Scott *et al.*, 2015; Harriman and Patel, 2016; Farquhar *et al.*, 2017; Trinquart *et al.*, 2018), and trial registration often does not lead to publication (Jones *et al.*, 2013; Al-durra *et al.*, 2018). While this does not necessarily induce bias, as registration and publication can be unrelated to study results, the burden falls on the systematic reviewers and decision-makers to determine whether a body of published evidence is subject to bias.

Ideally, a thorough search for evidence would entail combing through a plethora of trial registries and other sources of grey literature to identify unpublished results and obtain the most complete evidence-base. The process of conducting a systematic review is already very research intensive and time-consuming, particularly as publications continue to accumulate, as evidenced by the recent need for ‘rapid’ reviews (Gannan *et al.*, 2010; Wang *et al.*, 2019), and some databases are difficult to access. Since there is no central repository for all studies conducted, inclusion of unpublished studies in practice may not identify a full, representative sample. Ferguson and Brannick (2012) found evidence of bias in meta-analyses that included unpublished results, due to selection bias in grey literature searches and overrepresentation of unpublished work produced by the very authors who are conducting the meta-analysis. Furthermore, the motivation for including grey literature hinges on the assumption that publication processes are indeed biased against non-significant results. If this isn’t the case, the peer review process is meant to guard against the publishing of studies with low methodological rigor, which could bias results if they were to be included in a meta-analysis (Cook *et al.*, 1993). Finally, as not all trials are registered, there will always be missing data that cannot be recovered.

Post-hoc methods for handling bias allow reviewers to assess the potential for bias in a time-sensitive manner and help with the interpretation of results. Researchers recognized early on the crucial role of meta-analysis in summarizing knowledge and influencing practice and treatment decision-making, and therefore emphasized the importance of methodological rigor for protecting the validity of conclusions (Cooper, 1982). This, combined with the ability of meta-analyses to investigate reproducibility among independent studies and show the underlying empirical effect beyond random variation of studies, motivated the development of a host of statistical methods for detecting and correcting for PB in meta-analysis, which we outline in Figure 1 and discuss in the remaining sections.

Figure 1: Timeline of methodological development for publication bias

3 Fail-safe N and other ‘file drawer’ methods

The earliest attempts to quantify the impact of selective publishing assumed a very extreme form of bias (Lane and Dunlap, 1978; Rosenthal, 1979), in which only statistically significant results were published and all non-significant results were unpublished. The premise for the fail-safe N method (or ‘file-drawer’ method) developed by Rosenthal (1979) was the most extreme, assuming the true underlying effect was null and that journals were filled with the 5% of studies with Type-1 errors. It estimates the number of unpublished studies averaging a null result required to reverse the statistical significance of a summary result in a meta-analysis. The smaller the number of unpublished studies

required to reverse rejection of the null, the more doubt the reviewer would cast on results of the meta-analysis. Lane and Dunlap (1978) similarly assumed that all nonsignificant studies were unpublished but allowed the true effect to be different from the null.

As the first iteration of methods for handling PB, these approaches are limiting in their assumptions regarding the publication process, as non-significant results may still be published. They also lack flexibility by relating publication only to p-values, effectively combining the impact of effect size and precision while potentially ignoring the directionality of results. Orwin (1983) proposed an alternate fail-safe N based on effect sizes which could account for true effects different from the null, but it does not account for uncertainty in the estimates and similarly assumes a mean zero effect of the unpublished results. Furthermore, neither fail-safe N method accounts for weighting of the observed or unpublished studies by study-size, a modification that was later made by Rosenberg (2005).

Fail-safe N approaches generally offer a measure of confidence in meta-analytic conclusions, but in the absence of a statistical model or distributional assumptions for the unpublished studies, little more can be said regarding the presence or size of bias. From here, methods development expanded to incorporate assumptions regarding unpublished data and to investigate increasingly complex publication processes.

4 Graph-based approaches to publication bias

Summary points

- Graph-based tools for PB are both intuitive and easy to implement, making them popular among systematic reviewers
- Interpretations of funnel plots can be misleading, as PB is only one source of funnel plot asymmetry, along with heterogeneity and other ‘small-study effects’
- Regression tests and their modifications are useful for initial assessment of potential PB
- Meta-regression could be used to adjust for PB in the presence of explainable heterogeneity, but may not account for other sources of asymmetry

Graph-based approaches to PB evaluate whether observed study outcomes in a meta-analysis are systemically associated with their precisions. Since smaller studies require larger treatment effects to achieve statistical significance, selective publishing of significant results can induce a correlation between observed effect sizes and standard errors, a phenomenon known as “small-study effects” (Sterne et al., 2000, Marks-Anglin and Chen, 2020).

4.1 Visualizing and testing for asymmetry

Light and Pillemer's funnel plot (1984) was a key development that motivated a slew of testing and correction methods for PB, one of which (Egger *et al.*, 1997) catapulted it to become the most widely used visual tool for assessment of PB among medical researchers (Gerber *et al.*, 2007; Papageorgiou *et al.*, 2014). Based on the intuition that all component studies in a meta-analysis measure the same underlying effect with greater variability about the true value for smaller studies, a funnel plot graphs measures of precision (eg. inverse of standard error, or trial size) for individual studies against their corresponding effect sizes, leading to a symmetric inverted funnel shape. Selective publication of 'positive' or significant results can cause deviation from symmetry, since smaller studies requiring larger effect sizes to reach significance will be scattered further from the null than studies closer to the top. It is an intuitive approach, but subjective assessment often leads to errors in interpretation (Terrin *et al.*, 2005), particularly with a small number of studies (Lau *et al.*, 2006), since plots can appear asymmetric by chance even in the absence of publication bias.

The need for more objective means of detecting asymmetry beyond random chance led researchers to develop statistical tests for correlation between observed effect sizes and precisions. Begg and Mazumdar (1994) first proposed an adjusted rank correlation test, followed by Egger *et al.*'s (1997) regression test based on Galbraith's radial plot for heterogeneity (Galbraith, 1988), equivalent to a regression of treatment effects on standard errors weighted by the inverse of the standard errors. Later, Thompson and Sharp (1999) introduced a meta-regression testing approach accounting for between-study heterogeneity. A nonparametric, rank-based correction method, called 'trim-and-fill', was proposed by Duval and Tweedie (2000a; 2000b) to obtain adjusted summary estimates, by 'trimming' observed studies and subsequently imputing missing studies in order to recover symmetry.

Egger's seminal work on a simple, easily implemented regression test for asymmetry gained widespread popularity, becoming the second most cited article on systematic reviews and meta-analysis published between 1977 and 2008 (Lau *et al.*, 2006; Uthman *et al.*, 2013). Egger's paper also cautioned against always equating asymmetry (and equivalently, small-study effects) with PB. Egger and Sterne asserted that PB was merely one possible source of small-study effects, along with lower methodological quality of smaller trials, enrollment of more high-risk populations, clinical heterogeneity among subgroups, confounding, the effect measure used or even chance. Therefore, a limitation of the funnel graph and related methods is their inability to distinguish between PB and other sources of asymmetry and random heterogeneity, as evidenced by the lack of power of graph-based tests to detect bias in heterogeneous meta-analyses (Peters *et al.*, 2010), or 'trim-and-fill' leading to spurious adjustment in the presence of between-study heterogeneity when there is no PB (Terrin *et al.*, 2003; Peters *et al.*, 2007).

Evaluating the source of asymmetry is important for determining the most appropriate means of addressing it (Rücker *et al.*, 2011, Marks-Anglin and Chen, 2020). For example, in cases where clinical heterogeneity is the cause, a subgroup analysis may be more appropriate than correction, as a combined estimate is less clinically useful. Funnel plots and their associated methods have thus been promoted as tools for visualizing meta-analytic data and detecting initial evidence of small-study effects or PB. However, investigations based on funnel plots alone can be misleading and inadequate, especially in the presence of large between-study heterogeneity, and should be used for exploratory data analysis rather than primary data analysis (Tang and Liu, 2000; Terrin *et al.*, 2005). A survey study by Ioannidis and Trikalinos (2007a) found that reviewers often apply these methods inappropriately. We refer the reader to Sterne *et al.* (2011) for helpful guidelines on the interpretation of funnel plots and their associated tests, as well as Ioannidis (2006) for greater discussion on other possible sources of bias/asymmetry.

Modifications to the aforementioned tests have been proposed to rule out certain sources of small-study effects and explainable heterogeneity when visualizing and testing for PB. For example, effect measures for binary data (eg. log-odds ratios or diagnostic test accuracy results) are statistically correlated with their corresponding variance estimators, leading to funnel-plot asymmetry even in the absence of other sources of small-study effects. This was evidenced by inflated type I errors when binary outcomes were used, particularly for rare events and large treatment effects, and diagnostic test studies under heterogeneity (Sterne *et al.*, 2000; Schwarzer *et al.*, 2002; Peters *et al.*, 2006; Deeks *et al.*, 2005). Proposed solutions to avoid such induced correlation included modifications to Egger's test by Macaskill *et al.* (2001) and Peters *et al.* (2006), Harbord's score-based test (Harbord *et al.*, 2006), Schwarzer's rank-correlation test (Schwarzer *et al.*, 2007), and Rücker *et al.*'s (2008) variance-stabilizing arcsine transformations of effect measures. Other modifications tried to distinguish between explainable heterogeneity and unknown heterogeneity, by incorporating study-level covariates in regression tests or performing tests within subgroups defined by those covariates (Peters *et al.*, 2010). Modifications to funnel plots have also aided in interpretation, including the addition of contour lines representing significance regions, such that PB can be better identified if regions with observed studies correspond to areas of high statistical significance (Peters *et al.*, 2008; Lin, 2019).

While these modifications have served to improve type I error rates, other issues persist. Firstly, power continues to be an issue for the p-value based tests, particularly when the sample size is small (Sterne *et al.*, 2000; Furuya-Kanamori *et al.*, 2020) as is common in meta-analyses (Davey *et al.*, 2011). They also often lead to inconsistent conclusions, as no one test is optimal in all meta-analytic settings (Lin *et al.*, 2018), although Lin (2020) very recently proposed a hybrid test that incorporates the strengths of all of these tests to maximize power across different settings. Furthermore, some began to question the utility of testing for PB, as this conferred limited additional information beyond the possible

presence of selective publication (Sutton *et al.*, 2000). As such, emphasis began to shift towards sensitivity analysis and adjustment to better understand the extent to which PB impacts upon the results of a meta-analysis.

4.2 Graph-based correction for PB and meta-regression

The nonparametric trim-and-fill approach (Terrin *et al.*, 2003; Peters *et al.*, 2007) is the most widely used correction method for PB, and is feasible even in small datasets. However, as previously discussed it can lead to inappropriate adjustment in the presence of heterogeneity. It may also result in inflated standard errors since it imputes studies with the most extreme values (Schwarzer *et al.*, 2010). Zhu *et al.* (2018) also proposed a parametric approach for frequentist estimation of the adjusted treatment effect, formulating study omission as truncation of a normal distribution. They assume that studies with effect sizes below a certain threshold (or p-values above a threshold) are truncated, where thresholds are estimated using the most extreme observed study in the sample. Therefore, as with trim-and-fill, this approach may be susceptible to outliers. However, unlike trim-and-fill, it can provide some distinction between heterogeneity and PB when formulated as a random effects model.

Meta-regression has also been proposed as a means of obtaining adjusted summary effects while accounting for true effect heterogeneity through the inclusion of study-level covariates. A generalization of trim-and-fill for application to meta-regression was developed by Weinhandl and Duval (2012), allowing for clinical heterogeneity of the treatment effect as a function of study-level variables. Moreno *et al.* (2009) studied the performance of meta-regression-based adjustment of PB. Under the assumption that smaller studies are more biased, the authors extrapolate the regression lines to a study of infinite sample size representing an unbiased, adjusted effect estimate. Their methods displayed promising results compared to trim-and-fill, although performance similarly suffered under high levels of unexplained heterogeneity. Rücker *et al.* (2011) provided a similar adjustment approach (termed ‘limit meta-analysis’) that better distinguishes between small-study effects and heterogeneity, by diminishing within-study variability but retaining between-study variability in their extrapolations.

From the field of economics, Stanley and Doucouliagos wrote a series of working papers that were eventually published (Stanley, 2005; 2008; Stanley and Doucouliagos, 2014) on the FAT-PET-PEESE approach, a family of meta-regression approaches that test for funnel plot asymmetry (FAT; equivalent to Egger’s test), test for empirical effects adjusted for PB (PET), and a new model based on an approximation of Heckman’s regression for producing adjusted effect estimates (PEESE), as well as hybrids between these approaches that improve bias reduction compared to Eggers’ regression. We refer the reader to their papers and book, *Meta-Regression in Economics and Business* (Stanley and Doucouliagos, 2012), for greater detail on these methods.

5 Selection models for publication bias

Summary points

- Selection models seek to directly characterize the selective publication process to correct for PB
- Most models consider the probability of publication as a function of p-values
- Copas' model is more flexible, allowing publication to depend separately on outcomes and precisions and thus modeling the impact of 'directionality' of study results on publication
- Selection models suffer from weak identifiability due to the small number of studies and limited information on selection parameters, and have instead been used in sensitivity analysis
- Recent work improving identifiability of Copas' model may increase its use in practice

Another line of research involving selection models grew parallel to the development of graph-based methods. These models aim to provide corrected estimates in meta-analyses through simultaneous estimation of an outcome model and corresponding selection model or weight functions that characterize the probability of observing the included studies. By explicitly modelling the selection process, in addition to a random-effects formulation of the outcome model, they avoid the ambiguities associated with interpreting graph-based methods, particularly for heterogeneous data (Terrin *et al.*, 2003). However, along with issues regarding identifiability due to the complexity of some models and small sample sizes, their performance is dependent on arguably idealistic assumptions, and they are often recommended for sensitivity analysis (Vevea and Woods, 2005; Copas and Shi, 2001; McShane *et al.*, 2016) by fixing the values of the selection model parameters. As we will discuss, both prior and ongoing work aim to help address these issues and increase their use in practice.

5.1 p-value based models

Motivated by the simulations performed by Lane and Dunlap (1978) under an extreme form of selective publication, Hedges (1984) proposed the first model for study selection. Their model assumes that non-significant studies are published with zero probability, and significant results published with a probability of 1. They derived the conditional distribution of homogeneous effect sizes under this restriction in order to estimate an adjusted summary effect. Like the earliest fail-safe N methods, Hedge's model assumes an extreme, or 'worst-case' form of bias. This was followed by a less simplistic model by

Iyengar and Greenhouse (1988) which allows for the selection of non-significant studies through selection weights that monotonically decreased with increasing p-values, but similarly assumed homogeneous effect sizes.

The next stage of development involved random effects formulations of outcome models and left-continuous step functions for selection weights defined by ranges of p-values (Dear and Begg, 1992; Hedges, 1992; Vevea and Hedges, 1995), which are incorporated via inverse-probability weighting. Step functions have the advantage of incorporating ‘thresholding’ or ‘cliff’ effects in the publication process, where p-values below certain thresholds are perceived as more conclusive than those above the thresholds (Hedges, 1992), leading to clustering of studies around these points of discontinuity. There is evidence of this phenomenon in the psychology literature, where p-values cluster just below the .05 threshold (Masicampo et al., 2012). On the other hand, frequentist estimation of multiple weights is difficult in practice and only feasible with a large number of studies. A non-parametric class of weight functions was developed by Silliman (1997b) for sensitivity analysis, as well as Bayesian hierarchical modelling approaches for step weight functions by Silliman (1997a) and Givens *et al.* (1997). For frequentist alternatives with smaller datasets, Rufibach (2011) developed a regularized alternative to the purely nonparametric method by Dear and Begg (1992) by assuming monotonicity of the weight function. Vevea and Woods (2005) proposed a sensitivity analysis approach to Vevea and Hedges’ model (1995) using a priori fixed weights, and Hedges and Vevea (1996) and Citkowicz and Vevea (2017) replaced the step function mechanism with a more parsimonious, yet still flexible, beta distribution, governed by only two parameters to improve tractability in small samples. Gleser and Olkin (1996) also proposed a p-value based selection modelling approach to estimate the number of unpublished studies, with similar motivation to Rosenthal’s (1979) fail-safe N method and also assuming the null hypothesis is true.

By modelling the selection process as a function of p-values, the aforementioned weights depend jointly on effect size and standard errors. However, if selection depends only on the sample size of studies, then there is no bias introduced by this selection process if the average treatment effect is similar between large and small studies. Furthermore, models that employ a two-tailed selection mechanism (i.e. significant studies have a greater likelihood of publication regardless of the direction of results) (Dear and Begg, 1992; Hedges, 1992) do not account for the directionality of the estimated effects which may be important in practice, for example when a benefit or harm outcome is of primary interest.

5.2 Copas’ selection model

A more flexible model for characterizing the publication process was developed by John Copas and colleagues in a series of papers at the turn of the century (Copas and Li, 1997; Copas, 1999; Copas and Shi, 2000, 2001). Their model allows study selection to depend separately on effect sizes and standard errors, enabling researchers to characterize a wider

variety of selection mechanisms and model the impact of directionality of the study outcomes on publication probability. It also distinguishes between true within-study variance in the outcome model and the conditional observed variance in the latent-variable model for publication. In doing so they exclude spurious correlation between the observed effect sizes and standard errors (i.e. sources of asymmetry not due to PB).

While a promising tool to correct for PB, the small number of studies included in a meta-analysis lead to weak identifiability for the selection model parameters, and it was promoted as a sensitivity analysis approach by the authors (Copas and Li, 1997), in which estimation of adjusted effects proceeds using the profile likelihood of pooled effect size under assumed values for the selection model parameters. Carpenter *et al.* (2009) developed an R package (which has since been removed and the methods added to the `metasens` package (Schwarzer *et al.*, 2016)) to implement this sensitivity analysis and found it provided useful summaries in 80% of meta-analyses with binary outcomes. However, like trim-and-fill it underestimated the true standard error of the summary effect, since the authors did not account for the extra variability associated with the sensitivity parameters. Schwarzer *et al.* (2010) found that Copas' model was less sensitive to outliers compared to trim-and-fill, and less likely to overcorrect.

As a more sophisticated approach compared to graph-based methods, Copas' model has gained less traction in practice. Efforts have been made to address some of the challenges associated with its use, in particular the weak identifiability of the model and the need to choose selection model parameters in sensitivity analysis. Copas and colleagues proposed a simple, easily implemented formulation for a maximum bias bound over all possible selection functions as an alternative to sensitivity analysis, only requiring specification of the overall probability of publication and assuming that the probability of study selection is a nondecreasing function of study precision (Copas and Jackson, 2004). Henmi *et al.* (2007) extended this approach to deriving expanded confidence intervals around the treatment effect estimate that account for possible selection bias, as well as an upper bound on the p-value. While useful for assessing the robustness of results to potential 'worse-case' scenarios, these approaches provide a limited amount of information and are not ideal if something is in fact known about the publication process.

Within the last ten years, new data-augmentation approaches have been developed that overcome the weak identifiability of Copas' model to estimate adjusted treatment effects. A Bayesian framework was proposed by Mavridis *et al.* (2013; 2014) for the expanded network meta-analysis (NMA) setting, in which prior knowledge on the missingness process can be used. For a frequentist alternative, Ning *et al.* (2018) and Piao *et al.* (2019) offered expectation-maximization (EM) algorithm approaches for maximum likelihood estimation, using the symmetry property of funnel plots to impute missing studies, but basing the probability of missingness on Copas' model.

6 Recent development and future needs

Summary points

- Recent interpretable measures for quantification of PB are useful alternatives to testing.
- Outcome reporting bias (ORB) is equally concerning and harder to detect than PB. It may also impact the availability of published results, and therefore should be considered jointly with PB in missing data models
- More complex meta-analytic frameworks including network meta-analysis (NMA) and multivariate meta-analysis (MMA) require innovative methods to overcome the unique challenges of addressing PB in these settings
- A new class of ‘goodness of fit’ methods using selection models offer an attractive alternative to existing methods, but more work is needed to study their validity in real world settings
- Newly proposed utilities of selection models hold promise for more rigorous sensitivity analysis and testing of PB

In recent years, new directions have been pursued in relation to PB, and some important methodological gaps remain to be filled by researchers.

6.1 Quantification of PB

Firstly, in addition to sensitivity analysis and correction methods, quantification has been proposed as an alternative to testing, in the same way point estimates and confidence intervals convey magnitude as opposed to simply statistical significance. A simple candidate measure is the intercept term in Egger’s regression test (Egger et al., 1997), as an intercept of 0 corresponds to perfect funnel plot symmetry. Unfortunately, this is not a readily interpretable quantity and not comparable across meta-analyses. It is also dependent on the scale of the outcome used in a meta-analysis. Lin and Chu (2018) developed a more intuitive skewness measure for quantifying asymmetry in funnel plots using the study-specific standardized deviates from the mean, where a skewness of zero indicates symmetry. This measure is independent of both the scale of the outcome measure and the number of studies included in the meta-analysis. However, as with the graph-based tests, both measures rely on the symmetry assumption for PB (Schmid, 2017), whereas PB is only one source of asymmetry among other causes such as small-study effects and between-study heterogeneity.

Recently, Bai et al. (2020) developed a Bayesian Copas Selection model, where they introduced a new measure based on the Hellinger distance between the posterior distributions of the effect size before and after bias correction. Such a measure directly

quantifies the impact of publication bias on the inference on effect size, and takes the entire posterior distribution (rather than a point estimate) into consideration. Furthermore, this new measure is defined based on distributions, hence it is scale invariant, and naturally takes values between 0 and 1, with 0 being no difference in posterior densities (i.e. no impact due to PB), and 1 being vastly different (i.e. substantial impact due to PB). With this measure, they studied the prevalence of PB in 1500 randomly selected meta-analysis datasets from the Cochrane Database of Systematic Reviews, and found that 65.6% of them had small amounts of PB, 27.4% had moderate PB, 6.1% had high PB, and only 0.9% had substantial PB.

6.2 Outcome reporting bias

Selective outcome reporting is another form of reporting bias in meta-analyses, in which only certain evaluated outcomes in a study are reported, while other outcomes are either unreported or less fully reported (eg. secondary outcomes or subgroup analyses) due to the significance and direction of results (Hutton and Williamson, 2000). Additionally, the primary outcome in a study may change or be relegated to a secondary outcome (Vedula *et al.*, 2009), and clinical trials may include multiple measures and analyses for the outcome of interest which are then selectively reported (Mayo Wilson *et al.*, 2017). There is strong evidence for the existence of selective outcome reporting in biomedical literature across a wide spectrum of areas (Hahn *et al.*, 2002, Chan *et al.*, 2004, Chan and Altman, 2005; Dwan *et al.*, 2008; 2011; 2013; Smyth *et al.*, 2011; Franco *et al.*, 2016), as well as the impact of outcome reporting bias (ORB) (Kirkham *et al.*, 2010; Frosi *et al.*, 2015) on conclusions in meta-analyses.

A limited number of methods exist for handling ORB, as the vast majority of testing and correction methods have focused on PB as the reason for an incomplete data frame. Hutton and Williamson (2000) and Hahn *et al.* (2000) proposed sensitivity analysis approaches to correcting for ORB based on ‘worst-case’ scenarios, where only significant outcomes or subgroup effects are reported in a study. These were followed by a selection model approach by Jackson *et al.* (2005), an imputation method for binary outcomes by Williamson and Gamble (2005) and later a maximum bias bound (2007) in similar fashion to Copas and Jackson’s (2004) bound for PB. Following the development of a classification system for missing or incomplete outcome reporting in randomized trials (ORBIT; Kirkham *et al.*, 2010; Kirkham *et al.*, 2018), Copas and colleagues developed a selection model based on the ORBIT system for performing sensitivity analysis for benefit (2014) and harm (2018) outcomes.

6.3 Multivariate meta-analysis

Relatedly, selective reporting of outcomes implies that multiple outcomes are of interest in a study. Prior approaches have tested for small-study effects and reporting bias for each outcome separately (Chan *et al.*, 2004; Jun *et al.*, 2010; Kavalieratos *et al.*, 2016),

but this causes multiple testing issues. Furthermore, Jackson *et al.* (2011) and (Kirkham *et al.*, 2012) found that by borrowing strength across correlated outcomes, as is done in multivariate meta-analysis (MMA), investigators can learn more about the unreported outcomes and reduce the impact of ORB. MMA simultaneously models multiple outcomes (eg. safety and efficacy) in order to provide a full picture on decision-making to balance between benefits and risks. A key challenge for modelling MMA's is that within-study correlations between outcomes are generally unreported in practice. Recently, Hong *et al.* (2020) proposed a multivariate Egger's test for small-study effects using a pseudolikelihood approach to handle within-study correlations (Chen *et al.*, 2014), as well as a visualization tool called the 'galaxy plot' for bivariate meta-analyses, akin to the univariate funnel plot (Hong *et al.*, 2020). Continuing this line of work, a recent nonparametric, bivariate trim-and-fill correction method based on the galaxy plot has been developed by Luo *et al.* (2020).

Measures that are not directly affected by sample size and can quantify the overall evidence of reporting bias, as well as those that specifically evaluate PB as opposed to small-study effects, still need to be developed. Furthermore, because selective publishing and selective reporting can both lead to missingness of results and biased conclusions in a meta-analysis, more methodological work is needed to jointly consider these processes in the missing data mechanism.

6.4 Network meta-analysis

The increasing popularity of more complex meta-analytic frameworks for informing treatment decisions, such as network meta-analysis (NMA) (Lumley *et al.* 2002; Lu and Ades, 2004), have necessitated the development of novel methods for handling PB in these settings. NMAs enable simultaneous comparisons of multiple interventions by pooling evidence from direct comparisons in head-to-head trials with evidence from indirect comparisons under the assumption of evidence consistency.

Methods for evaluating PB in this setting have typically taken a contrast-based approach, while an arm-based approach has been increasingly studied as an alternative (Zhang *et al.*, 2015; Hong *et al.*, 2016; Lin *et al.*, 2016; Wang *et al.*, 2020). As such, additional considerations need to be accounted for including correlations between treatments in multi-arm studies, different selection mechanisms for different study designs or treatment contrasts, and the inclusion of indirect evidence. While efforts have been made to visualize NMAs using funnel graphs (Chaimani *et al.*, 2013), displaying indirect comparisons for the purpose of assessing symmetry is difficult, and extending graph-based methods to an NMA framework remains challenging. Trinquart *et al.* (2012) proposed an extended meta-regression approach for NMAs along the lines of Moreno *et al.* (2009) using only direct, pairwise treatment comparisons, and assuming the selection mechanism is shared across the network. However, it doesn't take advantage of information from indirect effects or account for within-study correlation in multi-arm studies.

Selection models can more easily incorporate indirect comparisons and within-study correlations through multivariable modelling, making extensions to the multiple treatment setting more feasible. Existing formulations (Chootrakool *et al.*, 2011) assume separate selection mechanisms for each study design in an NMA, but this poses the problem of increasing model complexity as the number of treatments and study designs increases, leading to weak identifiability and non-convergence with frequentist estimation. Mavridis *et al.*'s (2014) Bayesian framework is the only solution to date for estimating selection models in the NMA setting. However, the specification of prior distributions for multiple selection probabilities can be cumbersome, and the investigation of sensitivity to various types of priors can be complicated. Furthermore, the data applications in Chootrakool's and Mavridis' works have only illustrated the feasibility of their method in small networks of three treatments. Larger networks are common, particularly in the field of psychiatry (Cipriani *et al.*, 2009; 2011; 2016; 2018; Huhn *et al.*, 2019), which often include more than 6 treatments under comparison. Furthermore, some study designs in the networks only contain a single data point (i.e., only 1 head-to-head comparison).

More work is needed to successfully apply Copas' model to a larger network of treatments, perhaps using a reduced model that assumes a shared selection mechanism within a disease area for better model parsimony.

6.5 Advances in selection modeling

Finally, recent advances have promoted new utilities of selection models, including testing and a new approach to sensitivity analysis.

6.5.1 'Goodness of fit' methods

Firstly, methods have been proposed that focus on the distribution of the observed values for testing and correction, simplifying the assumptions needed. Ioannidis and Trikalinos (2007b) developed a test for excess number of significant studies, by comparing the expected number of significant studies (calculated under the assumption that the estimated summary effect is the true effect) to the observed number of significant studies in the meta-analysis. This generally only works for homogeneous effects, and the authors propose performing the tests within subgroups defined by study-level covariates if heterogeneity of the treatment effect is suspected.

The 'p-curve' (Simonsohn *et al.*, 2014a) and 'p-uniform' (van Assen *et al.*, 2015) approaches were later developed as alternative estimation strategies to Hedges' (1984) model, and involve testing whether the distribution of published p-values significantly deviate from a uniform distribution. For the p-curve method, this is a test for right skewness as evidence of a true treatment effect, under the assumption that the shape of the p-curve will be uniform if the true treatment effect is null, or left-skewed if there are large amounts of selective publishing or p-hacking (performing multiple different analyses of data collected in a study until a significant p-value is obtained, with selective publishing of the

one significant analysis). For p-uniform, the authors assume that the conditional distribution of p-values on the true effect size is uniform, and so a test for deviation from the uniform distribution is a test for PB, rather than a test for a true effect. Simonsohn *et al.* (2014b) and van Assen *et al.* (2015) also proposed effect size estimation corrected for PB using p-curve and p-uniform respectively.

These methods hold promise for increased utility over traditional selection models that are more difficult to implement, as they do not require estimation of selection model parameters which can be difficult in small samples. However, there is disagreement over the validity of their use in real world settings. Critiques of p-curve and p-uniform largely center on their inability to generalize to real-world settings with between-study heterogeneity, publication of non-significant results, and multiple-testing (Ulrich and Miller, 2014; McShane *et al.*, 2016; Carter *et al.*, 2019). Some of these concerns have been addressed by the authors (Simonsohn *et al.*, 2015; van Aert *et al.*, 2016; Simmons *et al.*, 2018), who offer solutions to improve robustness in real world settings and argue for the unbiasedness of p-curve methods for estimating the average treatment effect even in the presence of heterogeneity. A recent paper by Rice *et al.* (2018) describing the relevance and utility of fixed-effect models for heterogeneous settings may offer support for this argument. van Aert and van Assen (2018) offered a revised method, p-uniform*, that includes non-significant studies, though they assume that these studies share the same probability of publication; Simmons *et al.* (2017), on the other hand, defended the exclusion of nonsignificant results.

6.5.2 Testing with Copas' model

Copas' model has generally not been used for testing, as under the null hypothesis some of the selection model parameters are nonidentifiable and the Fisher information matrix is singular, making traditional testing (eg. Wald, Score test) infeasible. Recently, Duan *et al.* (2020) constructed a score test to overcome these non-regularity conditions, fixing the nonidentifiable parameters at specific values such that the submatrix of the Fisher information is non-singular. By calculating a test statistic for each pair of parameter values over a prespecified grid, one can then select the largest statistic for inference. They showed in simulation that this approach had controlled type I error, and although power was dependent on the number of grid points chosen, even with a small number of grid points it had competitive power compared to Egger's test in large samples. As with most selection-based methods, and particularly Copas' model, its use in practice may be limited by the need for investigators to specify a range of parameter values, necessitating a certain level of knowledge regarding selection models for missing data.

6.5.3 'Fail-safe' sensitivity analysis

Lastly, a sensitivity analysis approach using one-tailed p-value based selection has been developed by Mathur and Vanderweele (2020) that offers an intuitive interpretation in

similar vein to the fail-safe N approaches. It allows investigators to determine the number of nonaffirmative (non-positive or non-significant) studies required to ‘nullify’ an observed summary effect, effectively showing the robustness of results to possible selection bias. By incorporating a one-tailed selection mechanism, it accounts for the directionality of results while also relaxing the strong parametric distributional assumptions of earlier p-value-based models through use of a proportionality constant for the probability of publishing affirmative results (positive and affirmative) relative to nonaffirmative results.

This sensitivity analysis is complemented with a ‘significance funnel plot’, which adds a lens of interpretation to the original funnel plot by distinguishing between affirmative and non-affirmative studies and plotting their respective fixed-effects estimates. Under the most extreme form of the file-drawer problem, where published studies represent the 5% of studies that are significant by chance, the non-affirmative fixed-effect estimate would correspond to the adjusted treatment effect, and thus large discrepancies between this and the affirmative fixed-effects estimate may suggest susceptibility to PB. The use of this ‘worst-case’ assumption warrants careful interpretation of the plot as a sensitivity analysis, as opposed to evidence in support of or against the existence of PB. Furthermore, these tools are most useful when nonaffirmative studies are observed, which may not be the case if the true treatment effect is large.

Overall, we believe these are promising directions for rigorous assessment of PB, with additional consideration needed for applications to heterogeneous settings. More innovative modeling strategies that are intuitive and flexible in capturing differential impacts of study-size and direction of results on the publication process, while still being parsimonious and identifiable, are also useful. Finally, extensions of these methods to more complicated settings such as MMA, NMA and meta-analysis of diagnostic tests are needed.

7 Software

In the age of computing, the utility of the methods reviewed in this paper are largely limited by the availability of software to implement them, and many developments in this area have occurred within the last ten years. As such, we wish to briefly highlight the advances in statistical computing that have made these methods more accessible to investigators. Many cutting-edge methods developed by statisticians are released in the open source statistical software R, while STATA is popular among clinicians and epidemiologists. In Table 1 we offer a listing of packages available in these programs to implement several of the methods discussed. In addition to these platforms, p-curve has a dedicated R-based application which can be found at <http://www.p-curve.com/app4/>. Furthermore, fail-safe N and popular graph-based approaches can also be implemented using the user-friendly commercial software *Comprehensive Meta-Analysis*.

Notably in Table 1, the most widely used graph-based methods such as the funnel plot, Egger’s regression and trim-and-fill are available in both STATA and R, while

packages for selection models and newer methods are found only in R. This gap between the limited functionality in STATA/SAS (which are used more frequently in applied research (Dembe et al., 2011)) and the greater variety of functions in R (which requires more programming skills) may partly explain the discrepancy in use between graph-based and selection models. More robust and user-friendly platforms with wide offerings of functions are needed to support methods dissemination and rigorous meta-analysis implementation in clinical practice.

Package	Method
R	
meta	Funnel plot, Asymmetry tests (Begg, Egger, Macaskill, Harbord, Peter), Trim-and-Fill
metafor	Funnel plot, asymmetry tests (Begg, Egger), Trim-and-Fill, Fail-safe N (Rosenthal, Orwin, Rosenberg)
metamisc	Asymmetry tests (Egger, Macaskill, Peters, Debray (for survival data))
metasens	Copas sensitivity analysis, Copas and Jackson upper bound, Rucker's limit meta-analysis
PubBias	Ioannidis and Trikalinos excess significance test
puniform	p-uniform
PublicationBias	Mathur and Vanderweele sensitivity analysis for PB
publipha	Hedges (1992)
selectMeta	Selection models (Iyengar and Greenhouse (1988), Dear and Begg (1992), Rufibach (2011))
weightr	Vevea and Hedges (1995) selection model
xmeta	Test of PB in Copas model (Duan et al., 2020), Multivariate Egger's test (Hong et al., 2019), Galaxy plot (Hong et al., 2020), bivariate trim-and-fill method (Luo et al., 2020)
STATA	
meta bias	Asymmetry tests (Egger, Harbord, Peters, Begg)
meta funnel plot	Funnel plot
meta trim fill	Trim-and-Fill

8 Conclusion

As we have shown in this historical review, modern methods for evidence synthesis facilitated the formal study of PB and its impact, and approaches to handling PB have greatly advanced since the earliest ‘file-drawer’ attempts. However, no one method is optimal across all settings (Lin *et al.*, 2018; Carter *et al.*, 2019), and all require assumptions regarding the nature of the published and/or unpublished studies. It is important to therefore understand their appropriate uses in practice, within the context of sample size, outcome measure used, the presence of random and systemic heterogeneity, direction of potential bias and, when relevant, the assumptions we are willing to make for identifiability of the true effect.

With increasingly complex meta-analytic frameworks to better inform treatment decisions and policy, innovative solutions are needed for these settings that are robust, intuitive and easily implemented by reviewers. Furthermore, future methodological work should ideally target PB specifically, as opposed to the more encompassing phenomenon of small-study effects, or at least encourage reviewers to investigate the source of small-study effects before proceeding with bias reduction methods. Researchers should also be mindful of whether the meta-analysis models (i.e., fixed-effect vs random-effects) are appropriate in their specific settings. In particular, the re-evaluation of fixed-effects meta-analysis (Rice *et al.*, 2018) introduces new promise to detect, quantify and correct for PB in this new framework.

Acknowledgements

We would like to thank Drs. Chris Schmid, Kay Dickersin and Lifeng Lin for their assistance in identifying important contributions for inclusion in this review.

Funding

A.M. and Y.C.'s research was supported in part by NIH funds 1R01LM012607, 1R01AI130460 and 1R01HD099348.

Conflicts of Interest:

The authors declare no conflict of interest.

Highlights

- Selective publishing practices that are dependent on study results undermine evidence-based practice and impacts the reproducibility of scientific findings
- Cohort studies of approved and grant-funded projects offer clear evidence of the presence and prevalence of publication bias (PB)

- Trial registries were created to encourage transparency and reproducibility of findings. They have also been used to study the impact of PB on meta-analytic results
- Graph-based methods are useful for detecting initial evidence of publication bias or small-study effects, but can be misinterpreted or used inappropriately, particularly in heterogeneous settings
- Selection models seek to directly characterize the selective publication process to correct for PB, but may suffer from weak-identifiability and are often used for sensitivity analysis. Recent work has sought to address this challenge
- Innovations are needed for more complex settings such as network meta-analysis and multivariate meta-analysis, including interpretable quantifications of PB

References

- [1] Al-Durra M, Nolan RP, Seto E, Cafazzo JA, Eysenbach G. Nonpublication rates and characteristics of registered randomized clinical trials in digital health: Cross-sectional analysis. *Journal of medical Internet research*. 2018;20(12):e11924.
- [2] Angell M. Negative Studies. *New England Journal of Medicine*. 1989;321:464-466.
- [3] Bai R, Lin L, Boland MR, Chen Y. A Robust Bayesian Copas Selection Model for Quantifying and Correcting Publication Bias. *arXiv preprint arXiv:2005.02930*. 2020 May 6.
- [4] Bakan D. The test of significance in psychological research. *Psychological bulletin*. 1966;66(6):423.
- [5] Bardy AH. Bias in reporting clinical trials. *British journal of clinical pharmacology*. 1998;46(2):147-50.
- [6] Begg CB. A measure to aid in the interpretation of published clinical trials. *Statistics in medicine*. 1985;4(1):1-9.
- [7] Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. *Biometrics*. 1994:1088-101.
- [8] Bozarth JD, Roberts RR. Signifying significant significance. *American Psychologist*. 1972;27(8):774.
- [9] Carpenter JR, Rücker G, Schwarzer G. copas: An R package for fitting the Copas selection model. *The R Journal*. 2009;2:31-6.
- [10] Carter EC, Schönbrodt FD, Gervais WM, Hilgard J. Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*. 2019;2(2):115-44.
- [11] Chaimani A, Higgins JP, Mavridis D, Spyridonos P, Salanti G. Graphical tools for network meta-analysis in STATA. *PloS one*. 2013;8(10).
- [12] Chan AW, Altman DG. Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors. *BMJ*. 2005;330(7494):753.
- [13] Chan AW, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *Jama*. 2004;291(20):2457-65.
- [14] Chan AW, Krleža-Jerić K, Schmid I, Altman DG. Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. *Cmaj*. 2004;171(7):735-40.
- [15] Chen Y, Hong C, Riley RD. An alternative pseudolikelihood method for multivariate random-effects meta-analysis. *Statistics in medicine*. 2015;34(3):361-80.
- [16] Chootrakool H, Shi JQ, Yue R. Meta-analysis and sensitivity analysis for multi-arm trials with selection bias. *Statistics in medicine*. 2011;30(11):1183-98.

- [17] Cipriani A, Barbui C, Salanti G, Rendell J, Brown R, Stockton S, Purgato M, Spineli LM, Goodwin GM, Geddes JR. Comparative efficacy and acceptability of antimanic drugs in acute mania: a multiple-treatments meta-analysis. *The Lancet*. 2011;378(9799):1306-15.
- [18] Cipriani A, Furukawa TA, Salanti G, Chaimani A, Atkinson LZ, Ogawa Y, Leucht S, Ruhe HG, Turner EH, Higgins JP, Egger M. Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. *Focus*. 2018;16(4):420-9.
- [19] Cipriani A, Furukawa TA, Salanti G, Geddes JR, Higgins JP, Churchill R, Watanabe N, Nakagawa A, Omori IM, McGuire H, Tansella M. Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. *The Lancet*. 2009;373(9665):746-58.
- [20] Cipriani A, Zhou X, Del Giovane C, Hetrick SE, Qin B, Whittington C, Coghill D, Zhang Y, Hazell P, Leucht S, Cuijpers P. Comparative efficacy and tolerability of antidepressants for major depressive disorder in children and adolescents: a network meta-analysis. *The Lancet*. 2016;388(10047):881-90.
- [21] Citkowitz M, Vevea JL. A parsimonious weight function for modeling publication bias. *Psychological Methods*. 2017;22(1):28.
- [22] Cook DJ, Guyatt GH, Ryan G, Clifton J, Buckingham L, Willan A, McIlroy W, Oxman AD. Should unpublished data be included in meta-analyses?: Current convictions and controversies. *JAMA*. 1993;269(21):2749-53.
- [23] Cooper H, DeNeve K, Charlton K. Finding the missing science: The fate of studies submitted for review by a human subjects committee. *Psychological Methods*. 1997;2(4):447.
- [24] Cooper HM. Scientific guidelines for conducting integrative research reviews. *Review of educational research*. 1982;52(2):291-302.
- [25] Copas J. What works?: Selectivity models and meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 1999;162(1):95-109.
- [26] Copas J, Dwan K, Kirkham J, Williamson P. A model-based correction for outcome reporting bias in meta-analysis. *Biostatistics*. 2014;15(2):370-83.
- [27] Copas J, Jackson D. A bound for publication bias based on the fraction of unpublished studies. *Biometrics*. 2004 Mar;60(1):146-53.
- [28] Copas J, Li HG. Inference for non-random samples. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 1997;59(1):55-95.
- [29] Copas J, Marson A, Williamson P, Kirkham J. Model-based sensitivity analysis for outcome reporting bias in the meta analysis of benefit and harm outcomes. *Statistical methods in medical research*. 2019;28(3):889-903.

- [30] Copas J, Shi JQ. Meta-analysis, funnel plots and sensitivity analysis. *Biostatistics*. 2000;1(3):247-62.
- [31] Copas J, Shi JQ. A sensitivity analysis for publication bias in systematic reviews. *Statistical methods in medical research*. 2001;10(4):251-65.
- [32] Crawford JM, Briggs CL, Engeland CG. Publication bias and its implications for evidence-based clinical decision making. *Journal of dental education*. 2010;74(6):593-600.
- [33] Davey J, Turner RM, Clarke MJ, Higgins JP. Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: a cross-sectional, descriptive analysis. *BMC medical research methodology*. 2011;11(1):160.
- [34] DeAngelis CD, Drazen JM, Frizelle FA, Haug C, Hoey J, Horton R, Kotzin S, Laine C, Marusic A, Overbeke AJ, Schroeder TV. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. *Archives of dermatology*. 2005;141(1):76-7.
- [35] Dear KB, Begg CB. An approach for assessing publication bias prior to performing a meta-analysis. *Statistical Science*. 1992;7:237-45.
- [36] Debray TP, Moons KG, Riley RD. Detecting small-study effects and funnel plot asymmetry in meta-analysis of survival data: a comparison of new and existing tests. *Research synthesis methods*. 2018;9(1):41-50.
- [37] Decullier E, Lhéritier V, Chapuis F. Fate of biomedical research protocols and publication bias in France: retrospective cohort study. *BMJ*. 2005;331(7507):19.
- [38] Dickersin K, Chalmers I. Recognizing, investigating and dealing with incomplete and biased reporting of clinical research: from Francis Bacon to the WHO. *Journal of the Royal Society of Medicine*. 2011;104(12):532-8.
- [39] Dickersin K, Min YI, Meinert CL. Factors influencing publication of research results: follow-up of applications submitted to two institutional review boards. *JAMA*. 1992;267(3):374-8.
- [40] Dickersin K, Min YI. NIH clinical trials and publication bias. *The Online journal of current clinical trials*. 1993;Doc No 50.
- [41] Dickersin K. Publication bias: Recognizing the problem, understanding its origins and scope, and preventing harm. Publication bias in meta-analysis: Prevention, assessment and adjustments. 2005:11-33.
- [42] Driessen E, Hollon SD, Bockting CL, Cuijpers P, Turner EH. Does publication bias inflate the apparent efficacy of psychological treatment for major depressive disorder? A systematic review and meta-analysis of US National Institutes of Health-funded trials. *PloS one*. 2015;10(9):e0137864.

- [43] Duan R, Piao J, Marks-Anglin A, Tong J, Lin L, Chu H, Ning J, Chen Y. Testing for publication bias in meta-analysis under Copas selection model. *arXiv preprint arXiv:2007.00836*. 2020.
- [44] Duval S, Tweedie R. A nonparametric “trim and fill” method of accounting for publication bias in meta-analysis. *Journal of the american statistical association*. 2000a;95(449):89-98.
- [45] Duval S, Tweedie R. Trim and fill: a simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*. 2000b;56(2):455-63.
- [46] Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan AW, Cronin E, Decullier E, Easterbrook PJ, Von Elm E, Gamble C, Gherzi D. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PloS one*. 2008;3(8).
- [47] Dwan K, Altman DG, Cresswell L, Blundell M, Gamble CL, Williamson PR. Comparison of protocols and registry entries to published reports for randomised controlled trials. *Cochrane Database of Systematic Reviews*. 2011(1).
- [48] Dwan K, Gamble C, Williamson PR, Kirkham JJ. Systematic review of the empirical evidence of study publication bias and outcome reporting bias—an updated review. *PloS one*. 2013;8(7).
- [49] Easterbrook PJ, Gopalan R, Berlin JA, Matthews DR. Publication bias in clinical research. *The Lancet*. 1991;337(8746):867-72.
- [50] Editors. The reporting of unsuccessful cases (editorial). *The Boston Medical and Surgical Journal*. 1909;161:263–264.
- [51] Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ*. 1997;315(7109):629-34.
- [52] Farquhar CM, Showell MG, Showell EA, Beetham P, Baak N, Mourad S, Jordan VM. Clinical trial registration was not an indicator for low risk of bias. *Journal of clinical epidemiology*. 2017;84:47-53.
- [53] Ferguson CJ, Brannick MT. Publication bias in psychological science: prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological methods*. 2012;17(1):120.
- [54] Fisher RA. The design of experiments. 1949.
- [55] Franco A, Malhotra N, Simonovits G. Publication bias in the social sciences: Unlocking the file drawer. *Science*. 2014;345(6203):1502-5.
- [56] Franco A, Malhotra N, Simonovits G. Underreporting in psychology experiments: Evidence from a study registry. *Social Psychological and Personality Science*. 2016;7(1):8-12.
- [57] Freedman LP, Cockburn IM, Simcoe TS. The economics of reproducibility in preclinical research. *PLoS Biol*. 2015;13(6):e1002165.

- [58] Frosi G, Riley RD, Williamson PR, Kirkham JJ. Multivariate meta-analysis helps examine the impact of outcome reporting bias in Cochrane rheumatoid arthritis reviews. *Journal of clinical epidemiology*. 2015;68(5):542-50.
- [59] Furuya-Kanamori L, Xu C, Lin L, Doan T, Chu H, Thalib L, Doi SA. P value-driven methods were underpowered to detect publication bias: analysis of Cochrane review meta-analyses. *Journal of clinical epidemiology*. 2020;118:86-92.
- [60] Galbraith RF. Graphical display of estimates having differing standard errors. *Technometrics*. 1988;30(3):271-81.
- [61] Ganann R, Ciliska D, Thomas H. Expediting systematic reviews: methods and implications of rapid reviews. *Implementation Science*. 2010;5(1):56.
- [62] Gerber S, Tallon D, Trelle S, Schneider M, Jüni P, Egger M. Bibliographic study showed improving methodology of meta-analyses published in leading journals 1993–2002. *Journal of clinical epidemiology*. 2007;60(8):773-80.
- [63] Givens GH, Smith DD, Tweedie RL. Publication bias in meta-analysis: a Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. *Statistical Science*. 1997:221-40.
- [64] Gleser LJ, Olkin I. Models for estimating the number of unpublished studies. *Statistics in medicine*. 1996;15(23):2493-507.
- [65] Greenwald AG. Consequences of prejudice against the null hypothesis. *Psychological bulletin*. 1975;82(1):1
- [66] Hahn S, Williamson PR, Hutton JL, Garner P, Flynn EV. Assessing the potential for bias in meta-analysis due to selective reporting of subgroup analyses within studies. *Statistics in medicine*. 2000;19(24):3325-36.
- [67] Hahn S, Williamson PR, Hutton JL. Investigation of within-study selective reporting in clinical research: follow-up of applications submitted to a local research ethics committee. *Journal of evaluation in clinical practice*. 2002;8(3):353-9.
- [68] Harbord RM, Egger M, Sterne JA. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Statistics in medicine*. 2006;25(20):3443-57.
- [69] Harriman SL, Patel J. When are clinical trials registered? An analysis of prospective versus retrospective registration. *Trials*. 2016;17(1):187.
- [70] Hedges LV. Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics*. 1984;9(1):61-85.
- [71] Hedges LV. Modeling publication selection effects in meta-analysis. *Statistical Science*. 1992;7:246-55.

- [72] Hedges LV, Vevea JL. Estimating effect size under publication bias: Small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics*. 1996;21(4):299-332.
- [73] Henmi M, Copas JB, Eguchi S. Confidence intervals and P-values for meta-analysis with publication bias. *Biometrics*. 2007;63(2):475-82.
- [74] Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Statistics in medicine*. 2002;21(11):1539-58.
- [75] Hong C, Duan R, Zeng L, Hubbard RA, Lumley T, Riley R, Chu H, Kimmel SE, Chen Y. Galaxy Plot: A New Visualization Tool of Bivariate Meta-Analysis Studies. *American Journal of Epidemiology*. 2020.
- [76] Hong C, Zhang J, Li Y, Elia E, Riley R, Chen Y. A regression-based method for detecting publication bias in multivariate meta-analysis. *arXiv preprint arXiv:2002.04775*. 2020.
- [77] Hong H, Chu H, Zhang J, Carlin BP. A Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons. *Research synthesis methods*. 2016;7(1):6-22.
- [78] Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. *Cochrane Database of Systematic Reviews*. 2009(1).
- [79] Hudson KL, Lauer MS, Collins FS. Toward a new era of trust and transparency in clinical trials. *JAMA*. 2016;316(13):1353-4.
- [80] Huhn M, Nikolakopoulou A, Schneider-Thoma J, Krause M, Samara M, Peter N, Arndt T, Bäckers L, Rothe P, Cipriani A, Davis J. Comparative efficacy and tolerability of 32 oral antipsychotics for the acute treatment of adults with multi-episode schizophrenia: a systematic review and network meta-analysis. *The Lancet*. 2019;394(10202):939-51.
- [81] Hutton JL, Williamson PR. Bias in meta-analysis due to outcome variable selection within studies. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. 2000;49(3):359-70.
- [82] Ioannidis JP. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. *JAMA*. 1998;279(4):281-6.
- [83] Ioannidis JP. Why most published research findings are false. *PLoS med*. 2005;2(8):e124.
- [84] Ioannidis JP. Differentiating biases from genuine heterogeneity: Distinguishing artifactual from substantive effects. *Publication bias in meta-analysis: Prevention, assessment and adjustments*. 2005:287-302.
- [85] Ioannidis JP, Trikalinos TA. The appropriateness of asymmetry tests for publication bias in meta-analyses: a large survey. *Cmaj*. 2007a;176(8):1091-6.

- [86] Ioannidis JP, Trikalinos TA. An exploratory test for an excess of significant findings. *Clinical trials*. 2007b;4(3):245-53.
- [87] Iyengar S, Greenhouse JB. Selection models and the file drawer problem. *Statistical Science*. 1988;3:109-17.
- [88] Jackson D, Copas J, Sutton AJ. Modelling reporting bias: the operative mortality rate for ruptured abdominal aortic aneurysm repair. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 2005;168(4):737-52.
- [89] Jackson D, Riley R, White IR. Multivariate meta-analysis: potential and promise. *Statistics in medicine*. 2011;30(20):2481-98.
- [90] Johnson VE, Payne RD, Wang T, Asher A, Mandal S. On the reproducibility of psychological science. *Journal of the American Statistical Association*. 2017;112(517):1-0.
- [91] Jones CW, Handler L, Crowell KE, Keil LG, Weaver MA, Platts-Mills TF. Non-publication of large randomized clinical trials: cross sectional analysis. *BMJ*. 2013;347:f6104.
- [92] Jun M, Foote C, Lv J, Neal B, Patel A, Nicholls SJ, Grobbee DE, Cass A, Chalmers J, Perkovic V. Effects of fibrates on cardiovascular outcomes: a systematic review and meta-analysis. *The Lancet*. 2010 May 29;375(9729):1875-84.
- [93] Kasenda B, Von Elm E, You J, Blümle A, Tomonaga Y, Saccilotto R, Amstutz A, Bengough T, Meerpohl JJ, Stegert M, Tikkinen KA. Prevalence, characteristics, and publication of discontinued randomized trials. *JAMA*. 2014;311(10):1045-52.
- [94] Kavalieratos D, Corbelli J, Zhang D, Dionne-Odom JN, Ernecoff NC, Hanmer J, Hoydich ZP, Ikejiani DZ, Klein-Fedyshin M, Zimmermann C, Morton SC. Association between palliative care and patient and caregiver outcomes: a systematic review and meta-analysis. *Jama*. 2016;316(20):2104-14.
- [95] Kennedy-Shaffer L. Before $p < 0.05$ to beyond $p < 0.05$: Using history to contextualize p-values and significance testing. *The American Statistician*. 2019;73(sup1):82-90.
- [96] Kirkham JJ, Altman DG, Chan AW, Gamble C, Dwan KM, Williamson PR. Outcome reporting bias in trials: a methodological approach for assessment and adjustment in systematic reviews. *BMJ*. 2018;362:k3802.
- [97] Kirkham JJ, Dwan KM, Altman DG, Gamble C, Dodd S, Smyth R, Williamson PR. The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. *BMJ*. 2010 Feb 15;340:c365.
- [98] Kirkham JJ, Riley RD, Williamson PR. A multivariate meta-analysis approach for reducing the impact of outcome reporting bias in systematic reviews. *Statistics in medicine*. 2012;31(20):2179-95.
- [99] Lane DM, Dunlap WP. Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology*. 1978;31(2):107-12.

- [100] Lau J, Ioannidis JP, Terrin N, Schmid CH, Olkin I. The case of the misleading funnel plot. *BMJ*. 2006;333(7568):597-600.
- [101] Lexchin J, Bero LA, Djulbegovic B, Clark O. Pharmaceutical industry sponsorship and research outcome and quality: systematic review. *BMJ*. 2003;326(7400):1167-70.
- [102] Light RJ, Pillemer DB. Summing up; The Science of Reviewing Research. Harvard University Press; 1984.
- [103] Lin L. Graphical augmentations to sample-size-based funnel plot in meta-analysis. *Research synthesis methods*. 2019;10(3):376-88.
- [104] Lin L. Hybrid test for publication bias in meta-analysis. *Statistical Methods in Medical Research*. 2020; <https://doi.org/10.1177/0962280220910172>
- [105] Lin L, Chu H, Hodges JS. Sensitivity to excluding treatments in network meta-analysis. *Epidemiology (Cambridge, Mass.)*. 2016;27(4):562.
- [106] Lin L, Chu H. Quantifying publication bias in meta-analysis. *Biometrics*. 2018;74(3):785-94.
- [107] Lin L, Chu H, Murad MH, Hong C, Qu Z, Cole SR, Chen Y. Empirical comparison of publication bias tests in meta-analysis. *Journal of general internal medicine*. 2018;33(8):1260-7.
- [108] Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. *Statistics in medicine*. 2004;23(20):3105-24.
- [109] Lumley T. Network meta-analysis for indirect treatment comparisons. *Statistics in medicine*. 2002;21(16):2313-24.
- [110] Luo C, Marks-Anglin A, Duan R, Lin L, Hong C, Chu H, Chen Y. Accounting for small-study effects using a bivariate trim and fill meta-analysis procedure. medRxiv 2020.07.27.20161562. 2020.
- [111] Macaskill P, Walter SD, Irwig L. A comparison of methods to detect publication bias in meta-analysis. *Statistics in medicine*. 2001;20(4):641-54.
- [112] Marks-Anglin A, Chen Y. Small-study effects: current practice and challenges for future research. *Statistics and its Interface*. 2020.
- [113] Mathieu S, Boutron I, Moher D, Altman DG, Ravaud P. Comparison of registered and published primary outcomes in randomized controlled trials. *JAMA*. 2009;302(9):977-84.
- [114] Mathur M, VanderWeele T. Sensitivity analysis for publication bias in meta-analyses. OSF Preprints. 2020. doi:10.31219/osf.io/s9dp6.
- [115] Masicampo EJ, Lalande DR. A peculiar prevalence of p values just below. 05. *Quarterly journal of experimental psychology*. 2012;65(11):2271-9.

- [116] Mavridis D, Welton NJ, Sutton A, Salanti G. A selection model for accounting for publication bias in a full network meta-analysis. *Statistics in medicine*. 2014;33(30):5399-412
- [117] Mavridis D, Sutton A, Cipriani A, Salanti G. A fully Bayesian application of the Copas selection model for publication bias extended to network meta-analysis. *Statistics in medicine*. 2013;32(1):51-66.
- [118] Mayo-Wilson E, Fusco N, Li T, Hong H, Canner JK, Dickersin K. Multiple outcomes and analyses in clinical trials create challenges for interpretation and research synthesis. *Journal of clinical epidemiology*. 2017;86:39-50.
- [119] McShane BB, Böckenholt U, Hansen KT. Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*. 2016;11(5):730-49.
- [120] Moreno SG, Sutton AJ, Ades AE, Stanley TD, Abrams KR, Peters JL, Cooper NJ. Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC medical research methodology*. 2009;9(1):2.
- [121] Mueller KF, Meerpohl JJ, Briel M, Antes G, von Elm E, Lang B, Motschall E, Schwarzer G, Bassler D. Methods for detecting, quantifying, and adjusting for dissemination bias in meta-analysis are described. *Journal of clinical epidemiology*. 2016;80:25-33.
- [122] Murad MH, Chu H, Lin L, Wang Z. The effect of publication bias magnitude and direction on the certainty in evidence. *BMJ evidence-based medicine*. 2018;23(3):84-6.
- [123] Nieminen P, Rucker G, Miettunen J, Carpenter J, Schumacher M. Statistically significant papers in psychiatry were cited more often than others. *Journal of clinical epidemiology*. 2007;60(9):939-46.
- [124] Ning J, Chen Y, Piao J. Maximum likelihood estimation and EM algorithm of Copas-like selection model for publication bias correction. *Biostatistics*. 2017;18(3):495-504.
- [125] Orwin RG. A fail-safe N for effect size in meta-analysis. *Journal of educational statistics*. 1983;8(2):157-9.
- [126] Papageorgiou SN, Papadopoulos MA, Athanasiou AE. Reporting characteristics of meta-analyses in orthodontics: methodological assessment and statistical recommendations. *European Journal of Orthodontics*. 2014;36(1):74-85.
- [127] Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Comparison of two methods to detect publication bias in meta-analysis. *JAMA*. 2006;295(6):676-80.
- [128] Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. *Statistics in medicine*. 2007;26(25):4544-62.

- [129] Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *Journal of clinical epidemiology*. 2008;61(10):991-6.
- [130] Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L, Moreno SG. Assessing publication bias in meta-analyses in the presence of between-study heterogeneity. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 2010;173(3):575-91.
- [131] Piao J, Liu Y, Chen Y, Ning J. Copas-like selection model to correct publication bias in systematic review of diagnostic test studies. *Statistical methods in medical research*. 2019;28(10-11):2912-23.
- [132] Rice K, Higgins JP, Lumley T. A re-evaluation of fixed effect (s) meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 2018;181(1):205-27.
- [133] Rosenberg MS. The file-drawer problem revisited: a general weighted method for calculating fail-safe numbers in meta-analysis. *Evolution*. 2005;59(2):464-8.
- [134] Rosenthal R. The file drawer problem and tolerance for null results. *Psychological bulletin*. 1979;86(3):638.
- [135] Rothstein HR, Sutton AJ, Borenstein M. Publication bias in meta-analysis: Prevention, assessment and adjustments. 2005:1-7.
- [136] Rücker G, Carpenter JR, Schwarzer G. Detecting and adjusting for small-study effects in meta-analysis. *Biometrical Journal*. 2011;53(2):351-68.
- [137] Rücker G, Schwarzer G, Carpenter JR, Binder H, Schumacher M. Treatment-effect estimates adjusted for small-study effects via a limit meta-analysis. *Biostatistics*;12(1):122-42.
- [138] Rufibach K. Selection models with monotone weight functions in meta analysis. *Biometrical Journal*. 2011;53(4):689-704.
- [139] Scherer RW, Meerpohl JJ, Pfeifer N, Schmucker C, Schwarzer G, von Elm E. Full publication of results initially presented in abstracts. *Cochrane Database of Systematic Reviews*. 2018(11).
- [140] Schmid CH. Discussion of “quantifying publication bias in meta-analysis” by Lin et al. *Biometrics*. 2018;74(3):797-9.
- [141] Schmucker C, Schell LK, Portalupi S, Oeller P, Cabrera L, Bassler D, Schwarzer G, Scherer RW, Antes G, Von Elm E, Meerpohl JJ. Extent of non-publication in cohorts of studies approved by research ethics committees or included in trial registries. *PLoS one*. 2014;9(12).
- [142] Schmucker CM, Blümle A, Schell LK, Schwarzer G, Oeller P, Cabrera L, von Elm E, Briel M, Meerpohl JJ, OPEN consortium. Systematic review finds that study data not published in full text articles have unclear impact on meta-analyses results in medical research. *PLoS one*. 2017;12(4):e0176210.

- [143] Schwarzer G, Antes G, Schumacher M. 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*. 2002;21(17):2465-77.
- [144] Schwarzer G, Antes G, Schumacher M. A test for publication bias in meta-analysis with sparse binary data. *Statistics in medicine*. 2007;26(4):721-33.
- [145] Schwarzer G, Carpenter J, Rücker G. Empirical evaluation suggests Copas selection model preferable to trim-and-fill method for selection bias in meta-analysis. *Journal of clinical epidemiology*. 2010;63(3):282-8.
- [146] Schwarzer G, Carpenter JR, Rücker G. Metasens: Advanced Statistical Methods to Model and Adjust for Bias in Meta-Analysis. R package version 0.4-0. 2016.
- [147] Scott A, Rucklidge JJ, Mulder RT. Is mandatory prospective trial registration working to prevent publication of unregistered trials and selective outcome reporting? An observational study of five psychiatry journals that mandate prospective clinical trial registration. *PloS one*. 2015;10(8).
- [148] Silliman NP. Hierarchical selection models with applications in meta-analysis. *Journal of the American Statistical Association*. 1997a;92(439):926-36.
- [149] Silliman NP. Nonparametric classes of weight functions to model publication bias. *Biometrika*. 1997b;84(4):909-18.
- [150] Simes RJ. Publication bias: the case for an international registry of clinical trials. *Journal of clinical oncology*. 1986;4(10):1529-41.
- [151] Simmons JP, Nelson LD, Simonsohn U. P-curve handles heterogeneity just fine. Data Colada blog (2018).
- [152] Simonsohn U, Nelson LD, Simmons JP. P-curve: a key to the file-drawer. *Journal of experimental psychology: General*. 2014a;143(2):534.
- [153] Simonsohn U, Nelson LD, Simmons JP. P-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*. 2014b;9(6):666-81.
- [154] Simonsohn U, Simmons JP, Nelson LD. Better P-curves: Making P-curve analysis more robust to errors, fraud, and ambitious P-hacking, a Reply to Ulrich and Miller. *Journal of experimental psychology: General*. 2015;144(6):1146.
- [155] Simonsohn U, Simmons JP, Nelson LD. Why p-curve excludes $p > .05$. Data Colada blog (2017)
- [156] Smart RG. The importance of negative results in psychological research. *Canadian Psychologist/Psychologie Canadienne*. 1964;5(4):225.
- [157] Smith ML. Publication bias and meta-analysis. *Evaluation in education*. 1980;4:22-4.

- [158] Smyth RM, Kirkham JJ, Jacoby A, Altman DG, Gamble C, Williamson PR. Frequency and reasons for outcome reporting bias in clinical trials: interviews with trialists. *BMJ*. 2011;342:c7153.
- [159] Song F, Parekh-Bhurke S, Hooper L, Loke YK, Ryder JJ, Sutton AJ, Hing CB, Harvey I. Extent of publication bias in different categories of research cohorts: a meta-analysis of empirical studies. *BMC medical research methodology*. 2009;9(1):79.
- [160] Song F, Parekh S, Hooper L, Loke YK, Ryder J, Sutton AJ, Hing C, Kwok CS, Pang C, Harvey I. Dissemination and publication of research findings: an updated review of related biases. *Health Technol Assess*. 2010;14(8):1-93.
- [161] Stanley TD. Beyond publication bias. *Journal of economic surveys*. 2005;19(3):309-45.
- [162] Stanley TD. Meta-regression methods for detecting and estimating empirical effects in the presence of publication selection. *Oxford Bulletin of Economics and statistics*. 2008;70(1):103-27.
- [163] Stanley TD, Doucouliagos H. Meta-regression analysis in economics and business. Routledge; 2012.
- [164] Stanley TD, Doucouliagos H. Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*. 2014;5(1):60-78.
- [165] Sterling TD. Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American statistical association*. 1959;54(285):30-4.
- [166] Sterling TD, Rosenbaum WL, Weinkam JJ. Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician*. 1995;49(1):108-12.
- [167] Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *BMJ*. 1997;315(7109):640-5.
- [168] Sterne JA, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *Journal of clinical epidemiology*. 2000;53(11):1119-29.
- [169] Sutton AJ, Song F, Gilbody SM, Abrams KR. Modelling publication bias in meta-analysis: a review. *Statistical methods in medical research*. 2000;9(5):421-45.
- [170] Szucs D, Ioannidis J. When null hypothesis significance testing is unsuitable for research: a reassessment. *Frontiers in human neuroscience*. 2017;11:390.
- [171] Tang JL, Liu JL. Misleading funnel plot for detection of bias in meta-analysis. *Journal of clinical epidemiology*. 2000;53(5):477-84.
- [172] Terrin N, Schmid CH, Lau J, Olkin I. Adjusting for publication bias in the presence of heterogeneity. *Statistics in medicine*. 2003 Jul 15;22(13):2113-26.

- [173] Terrin N, Schmid CH, Lau J. In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. *Journal of clinical epidemiology*. 2005;58(9):894-901.
- [174] Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in medicine*. 1999;18(20):2693-708.
- [175] Trinquart L, Abbe A, Ravaud P. Impact of reporting bias in network meta-analysis of antidepressant placebo-controlled trials. *PLoS One*. 2012;7(4).
- [176] Trinquart L, Dunn AG, Bourgeois FT. Registration of published randomized trials: a systematic review and meta-analysis. *BMC medicine*. 2018;16(1):173.
- [177] Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R. Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine*. 2008;358(3):252-60.
- [178] Ulrich R, Miller J. p-hacking by post hoc selection with multiple opportunities: Detectability by skewness test?: Comment on Simonsohn, Nelson, and Simmons. *Journal of experimental psychology*. 2015;144(6):1137-45
- [179] US Food and Drug Administration. Food and Drug Administration Amendments Act (FDAAA) of 2007. US Public Law 110-85 section 801.
- [180] Uthman OA, Okwundu CI, Wiysonge CS, Young T, Clarke A. Citation classics in systematic reviews and meta-analyses: who wrote the top 100 most cited articles?. *PLoS one*. 2013;8(10).
- [181] van Aert RC, Wicherts JM, van Assen MALM. Conducting meta-analyses based on p values: Reservations and recommendations for applying p-uniform and p-curve. *Perspectives on Psychological Science*. 2016;11(5):713-29.
- [182] van Aert RC, van Assen MALM. P-uniform*. MetaArXiv. 2018. doi:10.31222/osf.io/zqjr9.
- [183] van Assen MALM, van Aert R, Wicherts JM. Meta-analysis using effect size distributions of only statistically significant studies. *Psychological methods*. 2015;20(3):293.
- [184] Vedula SS, Bero L, Scherer RW, Dickersin K. Outcome reporting in industry-sponsored trials of gabapentin for off-label use. *New England Journal of Medicine*. 2009;361(20):1963-71.
- [185] Vevea JL, Hedges LV. A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*. 1995;60(3):419-35.
- [186] Vevea JL, Woods CM. Publication bias in research synthesis: sensitivity analysis using a priori weight functions. *Psychological methods*. 2005;10(4):428.
- [187] Wang L, Rouse B, Marks-Anglin A, Duan R, Shi Q, Quach K, Chen Y, Cameron C, Schmid CH, Li T. Rapid network meta-analysis using data from Food and Drug

- Administration approval packages is feasible but with limitations. *Journal of clinical epidemiology*. 2019;114:84-94.
- [188] Wang Z, Lin L, Hodges J, Chu H. The impact of covariance priors on arm-based Bayesian network meta-analyses with binary outcomes. *Statistics in Medicine*. 2020.
- [189] Weinhandl ED, Duval S. Generalization of trim and fill for application in meta-regression. *Research synthesis methods*. 2012;3(1):51-67.
- [190] Williamson PR, Gamble C. Identification and impact of outcome selection bias in meta-analysis. *Statistics in medicine*. 2005;24(10):1547-61.
- [191] Zarin DA, Tse T, Williams RJ, Carr S. Trial reporting in ClinicalTrials.gov—the final rule. *New England Journal of Medicine*. 2016;375(20):1998-2004.
- [192] Zhang J, Chu H, Hong H, Virnig BA, Carlin BP. Bayesian hierarchical models for network meta-analysis incorporating nonignorable missingness. *Statistical methods in medical research*. 2017;26(5):2227-43.
- [193] Zhu Q, Carriere KC. Detecting and correcting for publication bias in meta-analysis—A truncated normal distribution approach. *Statistical methods in medical research*. 2018;27(9):2722-41.