Addressing Outcome Reporting Bias in Meta-analysis: A Selection Model Perspective

Alessandra Gaia Saracini and Leonhard Held

Abstract

Outcome Reporting Bias (ORB) poses significant threats to the validity of meta-analytic findings. It occurs when researchers selectively report outcomes based on the significance or direction of results, potentially leading to distorted treatment effect estimates. Despite its critical implications, ORB remains an under-recognized issue, with few comprehensive adjustment methods available. The goal of this research is investigate ORB-adjustment techniques through a selection model lenses, thereby extending some of the existing methodological approaches available in the literature. To gain a better insight into the effects of ORB in meta-analysis of clinical trials, specifically in the presence of heterogeneity, and to assess the effectiveness of ORB-adjustment techniques, we apply the methodology to real clinical data affected by ORB and conduct a simulation study focusing on treatment effect estimation with a secondary interest in heterogeneity quantification.

1 Introduction

Meta-analysis is a powerful statistical tool used to combine evidence from multiple studies investigating the same research question [11, 6]. It plays a crucial role in clinical research by providing a more comprehensive and robust analysis of treatment effects, especially when individual studies have limited statistical power. However, like any statistical method, meta-analysis is prone to biases that can affect its validity and reliability [6, 29, 12]. While publication bias (PB) is a well-known issue, with various statistical methods developed to address it, outcome reporting bias (ORB) is less explored but equally problematic [6, 29, 12, 21]. PB occurs when entire studies are not present in the literature due to the lack of significance or direction of results. On the other hand, ORB occurs when reporting decisions within published studies are influenced by results' significance or direction, leading to selective reporting of outcomes [19, 7, 9, 29, 20, 12, 32]. Therefore, unlike PB, studies affected by ORB may still be published, but certain outcomes, especially those unfavorable, may be omitted or reporting may be impartial, leading to inability to include the study outcome in a meta-analysis.

Studies have shown that ORB is prevalent in the meta-analysis literature, affecting reviews where both primary and secondary outcomes are often inadequately reported [4, 19, 27]. An investigation on a cohort of Cochrane systematic reviews by Kirkham *et al.* [19] found that more than half of the reviews did not include full data for the primary outcome of interest from eligible trials, and over a third contained at least one trial with high suspicion of ORB [19]. An investigation by Saini *et al.* [27], with a focus on meta-analyses where the primary outcome was a harmful one, found that 86% of Cochrane cohort reviews did not include full outcome data for the main adverse event of the trial, and ORB was suspected in nearly two thirds of the reviews [27]. A study by Chan *et al.* [4], inspecting 1402 outcomes from 48 trials with 68 publications, quantified the association between inadequate reporting of outcomes and statistical significance. They concluded that statistically significant beneficial outcomes have odds of being fully reported which are 2.7 times that of non-significant ones, with a 95% CI from 1.5 to 5.0 [4]. ORB poses a substantial threat to the integrity of meta-analyses, emphasizing the need for increased awareness and methods to mitigate its impact.

The current statistical methodologies to adjust for ORB, which differ in nature and underlying assumptions, can be summarized with the works of Kirkham *et al.* [20], Copas *et al.* [7, 9], Bay *et al.* [1], van Aert and Wicherts [32], including a bivariate meta-analysis adjustment of two correlated outcomes [20], a Bayesian extension of it [1], and a meta-regression approach [32]. The most established ORB-adjustment method, i.e., that of Copas *et al.* [9], relies on categorizing unreported outcomes into risk of bias categories - no risk (NR), low risk (LR), and high risk (HR) - based on the Outcome Reporting Bias in Trials (ORBIT) methodology. Given the classification, assumed to be correct, Copas *et al.* [9] developed a

likelihood-based ORB-adjustment method by adding a contribution from unreported study outcomes classified as HR of bias to the likelihood function, under the assumption that these were originally non-significant. In the Copas *et al.* [9] method, it is assumed that treatment effects, and possibly standard errors, are unreported, while sample sizes of the studies are known, and the adjustment is done separately for each outcome in the meta-analysis.

Our work can be seen as an extension of the Copas *et al.* [9] method by presenting ORB-adjustment through a selection model perspective, a framework typically used for PB adjustment [10, 14, 31]. The proposed approach for ORB adjustment offers a more flexible framework that does not require the ORBIT classification system, includes contribution from all unreported study outcomes, and allows for different assumptions on the missing data mechanism. We further consider the impact of heterogeneity between studies on ORB and ORB-adjustment, a novel aspect in the context of ORB, and conduct a simulation study investigating the impact of ORB and the effectiveness of ORB-adjustment, focusing on treatment effect estimation, with a secondary focus on heterogeneity, under different meta-analytic settings.

Throughout this work, we consider a random effects meta-analysis setting on a single beneficial outcome, i.e., an outcome for which a positive value indicates a beneficial direction of treatment. We assume normality and hence the following model:

$$y_i \sim \mathcal{N}(\theta_i, \sigma_i^2) \qquad \theta_i \sim \mathcal{N}(\mu, \tau^2), \tag{1}$$

where y_i and σ_i^2 are the observed treatment effect and standard error, respectively, for each study i, and the parameters of interest are μ, the treatment effect, and τ^2, the heterogeneity variance.

As a motivating example of ORB in meta-analysis, we consider the data used by Copas *et al.* [9], wherein a meta-analysis of 12 studies was conducted separately for 14 different outcomes, 2 considered beneficial and 12 harmful. The meta-analysis, originally by Bresnah *et al.* [3], includes studies investigating the effect of Topiramate, an antiepileptic drug first marketed in 1996, when used as an add-on treatment for drug-resistant focal epilepsy. Given that in this paper we focus on ORB and ORB correction for beneficial outcomes, we consider the 2 outcomes of the data assumed to have a positive effect, i.e., 50% seizure frequency reduction, and seizure freedom, illustrated in Table 1. We observe that all of the 12 studies in the meta-analysis report sizes; however, some studies do dot reporting the event frequencies, from which the log OR is computed and used as the normally distributed treatment effect, using a continuity correction in case of empty cell counts [9].

Table 1: Example meta-analysis data of beneficial outcomes affected by ORB [9, 3].

	Sample Size		50% Seizure Reduction		Seizure Freedom	
	T	C	T	C	T	C
Ben-Menachem 1996	28	28	12	0	Unrep	Unrep
Elterman 1999	41	45	16	9	4	2
Faught 1996	136	45	54	8	Unrep	Unrep
Guberman 2002	171	92	77	22	10	2
Korean 1999	91	86	45	11	7	1
Privitera 1996	143	47	58	4	Unrep	Unrep
Rosenfeld 1996	167	42	86	8	Unrep	Unrep
Sharief 1996	23	24	8	2	2	0
Tassinari 1996	30	30	14	3	0	0
Yen 2000	23	23	11	3	Unrep	Unrep
Zhang 2011	46	40	22	3	0	0
Coles 1999	52	51	Unrep	Unrep	Unrep	Unrep

This paper is organized as follows: Section 2 introduces the selection model framework typically used for PB and illustrates how this framework can be adapted to address ORB, considering various possible missing data mechanisms inspired by PB literature. Section 3 presents a simulation study investigating the impact of ORB and the effectiveness of the proposed ORB-adjustment method, with a focus on its application within a random effects meta-analysis model. Finally, Section 4 summarizes the proposed methodology and findings in a discussion, including limitations and conclusions.

2 Selection Models

Selection models have gained popularity in the PB adjustment literature [11, 14, 10, 31, 6], as they aim at correcting for the bias in treatment effect estimation by directly modelling the assumed missing data mechanism. Let y_i be the observed treatment effect estimate for study i in the meta-analysis, with distribution $f(y_i; \theta)$, assumed to be normal, where we denote θ as the unknown parameter of interest - in the context of the random effects meta-analysis of (1), θ is μ and τ^2.

The general form of a selection model in the PB literature involves the use of a weighted likelihood function which takes into account the observations y_i from published studies $i \in \{\text{Pub}\}$ by weighing them with a selection function $w(y_i)$ which describes the probability that study i is published/selected based on its significance [14, 10, 31]. By using the following relation:

$$f\left(y_i \mid i \in \{\text{Pub}\}\right) = \frac{f(y_i;\theta)\cdot w(y_i)}{\int_{-\infty}^{+\infty} f(y;\theta)\cdot w(y)dy,} \quad (2)$$

the PB-adjusted log-likelihood $\ell_{\text{Adj}}^{\text{PB}}(\theta)$ is derived [18, 15, 14, 10] as

$$\begin{aligned}\ell_{\text{Adj}}^{\text{PB}}(\theta) &= \sum_i \log f\left(y_i;\theta \mid i \in \{\text{Pub}\}\right) \\ &= \sum_{i\in\{\text{Pub}\}} \log f(y_i;\theta) \\ &\quad - \sum_{i\in\{\text{Pub}\}} \log\left[\int_{-\infty}^{\infty} f(y;\theta)\cdot w(y)dy\right].\end{aligned} \quad (3)$$

There are numerous forms which the selection function can take in the context of PB, with the general intuition being that in a meta-analysis of a beneficial outcome, for larger p-values, the probability of publication/selection decreases. Of note, in the case of a meta-analysis of a harmful outcome, we would expect the inverse, i.e., small, significant $p-$values to be likely not reported, as they would indicate harm [27, 9]. In the following sections we defined the selection functions assuming a beneficial outcomes and thus positive direction of treatment. The selection function $w(y_i)$ is thus often defined as a function of the p-value p_i, which constitutes an intuitive way of understanding the relationship between significance and probability of selection, [14, 10, 31]. Given that the p-value is simply a transformation of the observed treatment effect y_i and standard error σ_i, we use $w(y_i)$ for ease of notation.

2.1 Selection Models for ORB

In the PB selection model setting one takes into account only the non-missing studies $i \in \{\text{Pub}\}$ by defining the conditional log-likelihood, i.e., conditional on the studies being published. In the context of ORB adjustment methods, according to the framework developed by Copas *et al.* [9], the likelihood function takes into account studies for which we have both non-missing and missing outcome information. The studies have different log-likelihood contributions, depending on whether a study i reports the outcome, i.e., $i \in \{\text{Rep}\}$, or the study i does not report the outcome, i.e., $i \in \{\text{Unrep}\}$. The full ORB-adjusted log-likelihood, where $K = K_{\text{Rep}} + K_{\text{Unrep}}$ is the total number of studies, can be seen as

$$\begin{aligned}\ell_{\text{Adj}}^{\text{ORB}} &= \sum_{i=1}^{K} \ell(\theta) \\ &= \sum_{i\in\{\text{Rep}\}} \ell(\theta) + \sum_{i\in\{\text{Unrep}\}} \ell(\theta) \\ &= \sum_{i\in\{\text{Rep}\}} \log f(y_i;\theta) + \sum_{i\in\{\text{Unrep}\}} \log f(y_i;\theta).\end{aligned} \quad (4)$$

We can then adapt the formulation of equation (2) for ORB, by considering, for reported studies $\{i \in \text{Rep}\}$, the probability $w(y_i)$ of a study reporting an outcome, instead of the probability of a study being published. The following thus holds:

$$f\left(y_i \mid i \in \{\text{Rep}\}\right) = \frac{f(y_i;\theta)\cdot w(y_i)}{\int_{-\infty}^{\infty} f(y;\theta)\cdot w(y)dy}. \quad (5)$$

Similarly, for the unreported studies $i \in \{\text{Unrep}\}$, we can use the formulation (2) and consider the probability $1-w(y_i)$ of a study not reporting an outcome. We hence obtain

$$f\left(y_i \mid i \in \{\text{Unrep}\}\right) = \frac{f(y_i;\theta)\cdot(1-w(y_i))}{\int_{-\infty}^{\infty} f(y;\theta)\cdot(1-w(y))\,dy}. \quad (6)$$

Using (5) and (6), and solving for $f(y_i;\theta)$, we can re-write the ORB-adjusted log-likelihood (4) as

$$\begin{aligned}\ell_{\text{Adj}}^{\text{ORB}}(\theta) &= \sum_{i\in\{\text{Rep}\}} \log f(y_i;\theta) \\ &\quad - \sum_{i\in\{\text{Rep}\}} \log\left[\int_{-\infty}^{\infty} f(y;\theta)\cdot w(y)dy\right] \\ &\quad + \sum_{i\in\{\text{Unrep}\}} \log\left[\int_{-\infty}^{\infty} f(y;\theta)\cdot(1-w(y))\,dy\right].\end{aligned} \quad (7)$$

The likelihood (7) is the generic setting using a weight function for the probability of reporting, i.e., for $i \in \{\text{Rep}\}$, and a weight function for the probability of not reporting, i.e, for $i \in \{\text{Unrep}\}$. In the Copas *et al.* [9] model formulation, specific assumptions were made regarding the missing data mechanism, which result in a simplification of (7). For the reported outcomes, Copas *et al.* [9] implicitly do not assume any

selection process, i.e., $w(y_i) = 1$ when $i \in \{\text{Rep}\}$. This means that no weight function representing the reporting probability based on the p-value is associated with the reported observations. In light of this assumption, (7) can be further simplified to

$$\ell_{\text{Adj}}^{\text{ORB}}(\theta) = \sum_{i\in\{\text{Rep}\}} \log f(y_i;\theta) + \sum_{i\in\{\text{Unrep}\}} \log\left[\int_{-\infty}^{\infty} f(y;\theta)\cdot(1-w(y))\,dy\right]. \tag{8}$$

The log-likelihood (8) is thus the generic form for ORB adjustment, which has different shapes depending on the selection function $w(y_i)$ used, representative of the missing data mechanism assumed for unreported study outcomes. Given the alignment of our ORB adjustment with the PB framework of selection models, one can use similar selection functions which are typically found in the PB literature.

2.1.1 Selection Functions

We present a series of selection functions, defined as functions of the one-sided p-value, $p_i = \Phi(-y_i/\sigma_i)$, where α is the threshold for significance, e.g., $\alpha = 0.05$. We use a one-sided p-value to model the probability of selection, in alignment with selection models of beneficial outcomes in PB [15, 31, 26]. One of the simplest selection functions used for PB is:

$$w_A(y_i) = \begin{cases} 1 & \text{if } p_i \le \alpha \\ 0 & \text{if } p_i > \alpha, \end{cases} \tag{9}$$

While this selection function can be found in the PB literature [18, 14, 31], we note that it is also the one implicitly used in the Copas *et al.* [9] adjustment, although the authors do not explicitly frame the ORB adjustment via a selection model framework. Of note, in Copas *et al.* [9], ORB-adjustment is applied by including only the unreported study outcomes classified at HR of bias by the ORBIT classification system. They thus omit the unreported study outcomes classified e.g., at LR of bias, and regard them as missing at random. Furthermore, the authors use the two-sided $p-$value $p_i = 2\cdot\Phi(y_i/\sigma_i)$ instead of the one-sided one proposed in this work. We deem a one-sided p-value to be more appropriate to model the underlying missing data mechanism for a beneficial effect of treatment, as it would be unlikely for significant outcomes, but in the wrong direction, to be reported [15, 31, 26].

Using the log-likelihood (8) and the selection function (9) for a subset of the unreported studies, i.e., those classified as HR of bias, along with a two-sided significance test, we can easily see how we obtain the simplified ORB-adjusted log-likelihood presented for the random effects model in Copas *et al.* [9], namely:

$$\begin{aligned}\ell_{\text{Adj}}^{\text{ORB}}(\theta) &= \sum_{i\in\{\text{Rep}\}} \log f(y_i;\theta) + \sum_{i\in\{\text{HR}\}} \log\left[\int_{-\infty}^{+\infty} f(y;\theta)(1-w(y))dy\right] \\ &= -\frac{1}{2}\sum_{i\in\{\text{Rep}\}}\left[\log(\sigma_i^2+\tau^2) + \frac{(y_i-\mu)^2}{\sigma_i^2+\tau^2}\right] \\ &\quad + \sum_{i\in\{\text{HR}\}} \log\left[\Phi\left(\frac{z_\alpha\sigma_i-\mu}{\sqrt{\sigma_i^2+\tau^2}}\right) - \Phi\left(\frac{-z_\alpha\sigma_i-\mu}{\sqrt{\sigma_i^2+\tau^2}}\right)\right].\end{aligned} \tag{10}$$

The selection function (9) results in a simple shape of the ORB-adjusted log-likelihood; however, the underlying assumption regarding the missing data mechanism is somewhat strict, and extensions which relax its assumption are commonly found in the PB literature [14, 31]. One example is the function $w_B(y;\beta)$ with tuning parameter $\beta > 0$:

$$w_B(y;\beta) = \begin{cases} 1 & \text{if } p \le \alpha \\ \frac{p^{-\beta}}{\alpha^{-\beta}} & \text{if } p > \alpha \end{cases} \tag{11}$$

The idea of this selection function in the context of PB is that the associated probability of publishing which weights observations is greater than 0 for non-significant outcomes. Here, for the non-significant study outcomes, the associated probability of reporting is a decreasing function of the p-value, while significant study outcomes have an associated probability of reporting of 1.

In the context of ORB we further propose a different selection function, $w_C(y;\gamma)$ with tuning parameter $\gamma > 0$ in (12), for which the rationale is inverted compared to $w_B(y;\beta)$ in (11), as we assume that non-significant study outcomes have an associated probability of reporting 0, while significant study outcomes have an associated probability of reporting which is a decreasing function of the p-value. This can be motivated by scenarios where ORB results from prioritizing more impactful or clinically relevant findings in a published study [5, 30], leading to only highly significant outcomes being reported. This could be interpreted as a lower threshold for not reporting compared to PB, and thus a higher level of bias. At the same time, given that the selection function allows for significant unreported outcomes, it can also account

for settings in which outcomes are missing because they were deemed less relevant, resulting in a more random pattern of missing data and less bias [28]. Understanding the exact cause of unreporting can be challenging, and information on the strength of evidence for other outcomes in the meta-analysis could help clarify the likely cause of unreporting.

$$w_C(y;\gamma) = \begin{cases} 1 - \frac{p^\gamma}{\alpha^\gamma} & \text{if } p \leq \alpha \\ 0 & \text{if } p > \alpha \end{cases} \tag{12}$$

Based on the selection functions $w_B(y;\beta)$ in (11) and $w_C(y;\gamma)$ in (12) one could envisage a combination of these by using e.g., selection function $w_D(y;\beta,\gamma)$ in (13). In this case, one can flexibly specify both γ and β parameters, as well as the probability of reporting assumed for a study outcome at the significance threshold α, which we note ω_α. In the case of (11), ω_α was implicitly 1 and in case of (12) this was set to 0. Here, we set $\omega_\alpha = 0.5$, as a middle value between (11) and (12). The selection function $w_D(y;\beta,\gamma)$ has the potential of being used to conduct extensive sensitivity analyses when adjusting for ORB.

$$w_D(y;\beta,\gamma) = \begin{cases} 1 - (1 - w_\alpha)\left(\frac{p^\gamma}{\alpha^\gamma}\right) & \text{if } p \leq \alpha \\ w_\alpha \left(\frac{p^{-\beta}}{\alpha^{-\beta}}\right) & \text{if } p > \alpha \end{cases} \tag{13}$$

The selection functions proposed above, namely $w_A(y)$ in (9), $w_B(y;\beta,\gamma)$ in (11), $w_C(y;\gamma)$ in (12) and $w_D(y;\beta,\gamma)$ in (13) are plotted in Figure 1 for some example values of the $\gamma > 0$ and $\beta > 0$ parameters. Further rationale for the parameter choices are discussed in the simulation study protocol, available in the OSF project repository.

2.1.2 Imputation of Missing Variances

When utilizing any of the selection functions presented in the ORB-adjusted log-likelihood (8), we require knowledge of the standard error of the unreported study outcome, which is generally missing. This value hence needs to be imputed; we do so following the methodology of Copas *et al.* [9], used also in the ORB-adjustment approach of Bay *et al.* [1]

$$\sigma_i^2 \approx \frac{1}{\hat{k} n_i}, \tag{14}$$

where n_i is the sample size of each study i and $\hat{k}$ is:

$$\hat{k} = \frac{\sum_{i \in \{\text{Rep}\}} \sigma_i^{-2}}{\sum_{i \in \{\text{Rep}\}} n_i}. \tag{15}$$

With the selection model framework for ORB adjustment presented in this work, one is thus able to include a likelihood contribution from unreported study outcomes, by specifying the desired missing data assumption via a selection function, representative of the assumed probability of reporting. This framework enables the joint estimation, via maximum likelihood (ML), of the ORB-adjusted parameters of interest in the model, in our case treatment effect, as well the heterogeneity variance.

2.1.3 Application to Motivating Example

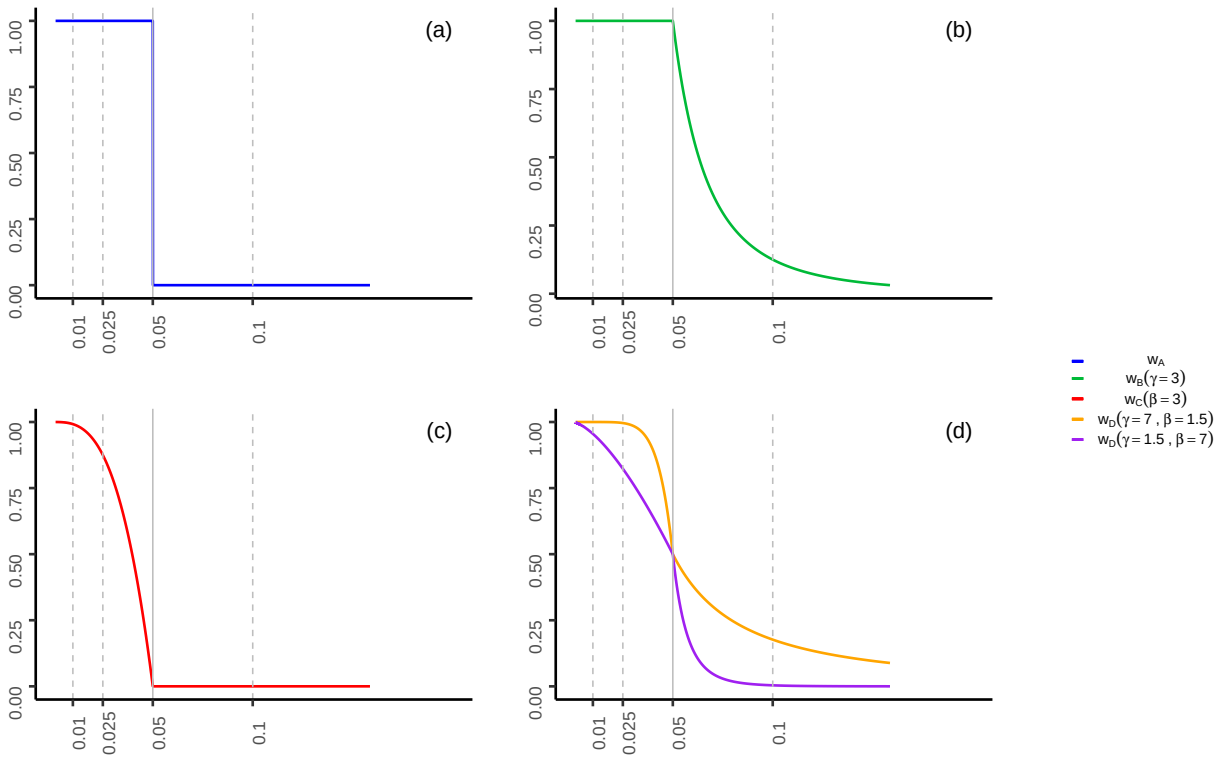
We applied the ORB-adjustment framework to the Epilepsy data from Bresnah *et al.* [3], Copas *et al.* [9], using the selection functions proposed in this study. These selection functions utilize a one-sided p-value for significance with a threshold of $\alpha = 0.05$, in contrast to Copas *et al.* [9], which used a two-sided p-value. While a one-sided threshold defines the underlying missing data mechanism, two-sided significance is used to construct profile likelihood (PL) confidence intervals (CIs) for the treatment effect estimate.

Figure 2 presents the point estimates and 95% CI for the log RR of the treatment effect for two beneficial outcomes in the meta-analysis: a 50% reduction in seizure frequency and seizure freedom. For both outcomes, the naive log RR estimate shows a significant positive treatment effect compared to the control.

For the 50% seizure frequency reduction, the ORB-adjusted estimates are slightly shifted towards the null value and are consistent across different selection functions. This minor shift is expected since only one study had unreported outcomes. However, for the seizure freedom outcome, with several unreported study outcomes, the ORB-adjusted estimates show a substantial shift towards the null, even negating the significance of the results. The differences between the ORB-adjusted estimates using various selection functions are more pronounced here.

The strictness of different ORB-adjustments, derived from the selection functions, is intuitive and stems from their underlying assumptions about unreported study outcomes. The estimate obtained using $w_B(\gamma = 3)$ is more conservative than that obtained with w_A. The selection function w_A assumes a probability of unreporting of 1 for non-significant studies, regardless of the p-value magnitude, while $w_B(\gamma = 3)$ assumes a higher probability of unreporting for larger

Figure 1: Possible Selection Functions for ORB-adjustment. Function $w_A(y)$ from equation (9) in (a), function $w_B(y;\beta = 3)$ from equation (11) in (b), function $w_C(y;\gamma = 3)$ from equation (12) in (c), and functions $w_D(y;\beta = 1.5,\gamma = 7)$ and $w_D(y;\beta = 7,\gamma = 1.5)$ from equation(13) shown in (d).



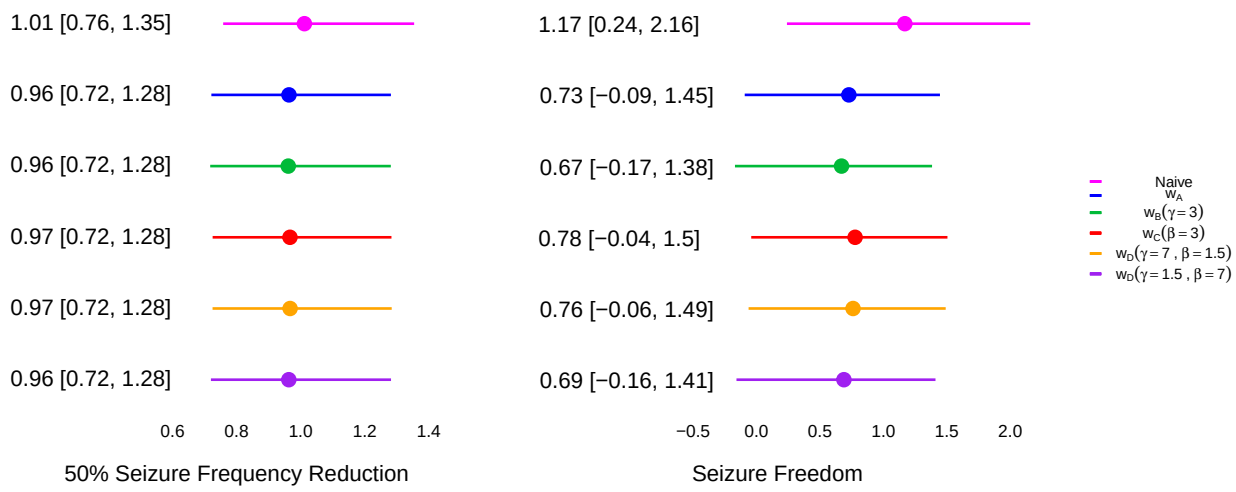
p-values, implying greater bias and thus stricter correction. Conversely, w_C is less conservative than w_A, as it assumes that some unreported outcomes may still be significant, indicating less bias and thus a less strict adjustment. Functions $w_D(\gamma = 1.5, \beta = 7)$ and $w_D(\gamma = 7, \beta = 1.5)$ align with $w_B(\gamma = 3)$ and $w_C(\beta = 3)$, respectively, suggesting that the larger parameter between γ and β drives the estimate.

Since the true underlying data mechanism is unknown in this example, we further investigate the ORB-adjustment effect in a simulation study. This allows us to evaluate the ORB adjustment implementation using both correctly specified and misspecified models of the missing data mechanism.

3 Simulation Study

It is of interest to assess the extent to which ORB negatively impacts meta-analytic findings and the extent to which the ORB-adjustment methodology presented in the previous section is effective in reducing this bias. Our primary interest lies in the bias detection and mitigation for treatment effect estimation under different meta-analysis settings, e.g., varying levels of heterogeneity and meta-analysis study sizes. A secondary interest of the investigation is the possible impact of ORB on heterogeneity variance estimation. To achieve this, we conduct a simulation study wherein we first simulate a random effects meta-analysis of a single beneficial outcome and subsequently mimic selective reporting by removing the observed treatment effect and standard error from the meta-analysis dataset based on the strength and/or direction of the results, favoring the reporting of studies with small p-values. We then utilize different estimation methods for the parameters of interest and assess the performance of the methods using performance measures on a large number of simulations. The details of the simulation study can be found in the simulation study protocol (already available in the OSF project repository) and are summarized in the following setting description subsection.

Figure 2: Application of ORB-adjustment to example data [9, 3], using the different selection functions showcased in Figure 1. In addition, the naive estimate, without ORB-adjustment, is shown for comparison.



3.1 Setting

The first step of the simulation process consists in simulating random-effects meta-analysis datasets in the presence of ORB. We first simulate a random effects meta-analysis study comprising K studies, each with treatment and control arms of equal sizes $n_i = n = 50$, and reported treatment effects y_i with standard errors σ_i. We first obtain the study-specific true treatment effects θ from

$$\theta_i \sim \mathcal{N}(\mu, \tau^2), \tag{16}$$

where μ is the overall treatment effect and τ^2 is the between-study heterogeneity variance. The observed treatment effects y_i are given by

$$y_i \sim \mathcal{N}(\theta_i, \sigma^2), \tag{17}$$

where $\sigma^2 = 2/n$, while the standard errors are generated from a scaled ^{2} distribution

$$\sigma_i^2 \sim \frac{\chi^2_{2n_i-2}}{(n_i-1)n_i}. \tag{18}$$

These values are generated independently for each study, assuming no correlation between studies. We then simulate ORB by selectively excluding certain studies from the meta-analysis based on the direction and significance of treatment effects. The ORB simulation process involves removing study outcomes with a probability of reporting determined by a decreasing function of the one-sided p-value, i.e., $p_i = \Phi(-y_i/\sigma_i)$. The function (19) used to simulate ORB is taken from simulation studies on PB, for consistency with our selection model approach, typical of PB settings. We simulate under two ORB settings, i.e., $\gamma = 1.5$ [10, 2, 22, 26] and $\gamma = 0.5$.

$$P(i \in \{\text{Rep}\}) = e^{-4 \cdot p_i^{\gamma}}. \tag{19}$$

Each meta-analysis dataset hence results in K or fewer of the original study outcomes. If for some meta-analysis datasets, less than two study outcomes are reported, the simulation is repeated until at least two reported study outcomes are obtained [10, 2, 22, 13]. The ORB-affected meta-analysis datasets are generated under different settings - we vary the number of studies in the meta-analysis, $K \in \{5, 15, 30\}$, the amount of between-study heterogeneity $I^2 \in \{0\%, 25\%, 50\%, 75\%, 90\%\}$ and the true underlying treatment effect $\mu \in \{0, ..., 0.8\}$ with increments of 0.1 [23, 17, 13].

After having simulated ORB, hence resulting in some treatment effects and standard errors unreported, we use maximum likelihood (ML) estimation to obtain point estimates of the treatment effect μ and the heterogeneity variance τ^2, along with profile likelihood (PL) confidence intervals (CI) [33, 9, 16]. The ML estimate and PL CI for μ and τ^2 are obtained using different log-likelihoods, depending on the information and/or missing data mechanism assumed, leading to i) naive, ii) complete data, and iii) ORB-adjusted estimation methods. We further differentiate various ORB-adjusted estimates based on the selection function assumed for the probability of

reporting.

The naive log-likelihood (i) includes the contribution only from reported study outcomes and disregards the unreported ones. The naive estimate serves as a baseline for comparison of the ORB-adjustment methodologies and quantifies the negative impact of ORB when the latter is not accounted for [31, 9]. The complete data log-likelihood (ii) uses all studies in the meta-analysis before ORB is simulated, and is a proxy for the true treatment effect if there were no ORB. The various ORB-adjusted estimates (iii) are obtained by maximizing the ORB-adjusted log-likelihood (8) using the different selection functions: $w_A(y)$ from (9), $w_B(y;\beta = 3)$ from (11), $w_C(y;\gamma = 3)$ from (12), and $w_D(y;\beta = 1.5,\gamma = 7)$, $w_D(y;\beta = 7,\gamma = 1.5)$ from (13), as well as the selection function (19) used to simulate ORB, so as to include the correct model specification in the adjustment. Since the latter can be indeed seen a selection function, we note as $w_{DGM}(y)$. The parameters of the selection functions, i.e., β or γ used in the adjustment correspond to those illustrated in Figure 1.

For each parameter setting, the simulation process is repeated $\mathrm{N_{sim}} = 3200$ times; the simulation size $\mathrm{N_{sim}}$ is calculated based on the expected variance of the unknown parameter estimate [17, 24] and a desired Monte Carlo Standard Error (MCSE) of 0.005 from IntHout *et al.* [17], Morris *et al.* [24]. The performance measures recorded for the unknown parameter are bias, empirical standard error (SE), mean squared error (MSE), coverage and power, along with the MCSEs of each, as per IntHout *et al.* [17], Morris *et al.* [24].

3.2 Results

Firstly, the results indicate a substantial bias in the estimation of the treatment effect when using naive estimation methods that do not account for ORB. This aligns with existing literature [9, 1, 32] and prior exploratory analysis [28]. As the true treatment effect size increases, the bias diminishes, reflecting the reduced likelihood of unreported studies, given the higher change of statistically significant results for large treatment effect sizes. Study size variations ($K = 5, 15, 30$) do not significantly affect the bias, though confidence intervals are larger for smaller study sizes. High heterogeneity settings, particularly with $I^2 = 90$, exhibit larger biases, reinforcing findings from previous exploratory work [28]. Incorporating heterogeneity effects into the ORB framework offers novel insights compared to past investigations [9].

The observed patterns of naive estimation are consistent across both ORB simulation processes, regardless of the DGM parameter γ value, whether $\gamma = 1.5$ (mostly non-significant unreported studies) or $\gamma = 0.5$ (some significant unreported studies, with larger p-values).

When applying the ORB-adjustment framework using selection functions, the bias is eliminated when the selection function matches the ORB DGM, confirming the model's effectiveness when correctly specified. Different selection functions (w_A, w_B, w_C, w_D) show varying degrees of bias reduction. For the DGM with $\gamma = 1.5$, the ORB-adjusted estimates shift the bias towards the null but do not fully eliminate it unless the exact DGM function is used. w_B performs slightly better than w_A, and w_C performs the least well, with minimal differences noted among the functions, particularly in low heterogeneity settings.

In the $\gamma = 0.5$ setting, similar patterns are observed, with w_B being the least strict and w_C the most strict. The ORB-adjustment here tends to reduce the treatment effect size excessively, indicating potential overcorrection due to the steep p-value dependence.

Beyond bias, the MSE is significantly reduced for all ORB-adjusted estimates, particularly in high heterogeneity settings. Coverage is highest under the true DGM, with ORB-adjusted estimates closely following the bias results. Naive estimation shows inflated power, while correct DGM specification aligns power appropriately. There is no substantial difference in empirical SE between naive and ORB-adjusted estimates, although the true DGM has the lowest SE.

Our parameter of interest and the focus of our investigation was treatment effect, hence in the random effects model when carrying out MLE, the primary parameter was μ. However, we also had a secondary interest in the estimation of the heterogeneity variance τ^2 in the presence of ORB, and the effect of the ORB-adjustments on its estimation. The heterogeneity variance τ^2 is jointly estimated from the ML with the μ parameter, which constitues an exploratory way of assessing its value. From Figure XXX, showin the bias in the estimatino for the various meta-analysis sizes and for different true treatment effect values, we observre that as heteroegeniy increases, the bias in its estimation for increases, i.e., in high heterogeneoty settings, naive estimation underestimates the heterogeneity variance. For no heterogeneity or relally low, this does not seem to be an issue. The ORB-adjusted etsiamte using the correct DGM mechanism appears to reduce this bias, while the other mis-speicified functions only patrially reduce the bias, with no remarkable differences noted,

- Heterogeneity
- Considerations on heterogeneity estimation in the presence of ORB
- Adjusted ORB, effect in join ML estimation

4 Discussion

This study addresses Outcome Reporting Bias (ORB), where the reporting of study outcomes is influenced by their significance, leading to overestimation of treatment effects in meta-analyses. We approached ORB adjustment through a selection model framework, a common method in publication bias (PB) literature, incorporating contributions from unreported study outcomes based on assumed missing data mechanisms. Our proposed selection functions expand on existing methods, including those from previous works like Copas *et al.* [9], by being more flexible in the missing data mechanisms and encompassing all unreported outcomes.

Applying ORB-adjustment to real-world meta-analyses on epilepsy Copas *et al.* [9], Bresnah *et al.* [3] showed substantial shifts towards null estimates, especially in cases with numerous unreported outcomes. Simulation studies within a random effects model framework demonstrated that naive estimation significantly overestimates treatment effects, particularly in high heterogeneity settings. The selection model framework successfully eliminates this bias under correct model specification, while different misspecified selection functions still reduce bias but vary in strictness.

4.1 Limitations

The ORB-adjustment methodology proposed is flexible and broadly applicable, but several limitations exist. The framework operates on individual outcomes in meta-analyses, not accounting for correlations between outcomes. Future research could explore methods to incorporate such correlations. The current approach to imputing missing variances, as done in previous works [7, 9**?** , 28], might be extended through multiple imputation techniques. Adjustments were made for normally distributed outcomes, which might not be precise for binary data [7, 9]. Exploring a binomial likelihood for such cases could be a potential avenue [28].

Maximum likelihood estimation (MLE) was used for estimating heterogeneity variance due to its connection to ORB-adjustment, though more sophisticated methods like restricted maximum likelihood (REML) could be considered [25, 33, 8, 28]. Our focus was on beneficial outcomes, but the methodology can be extended to harmful outcomes by adjusting the selection functions for a different missing data mechanism accordingly.

The simulation study tested a limited range of DGM parameters and ORB-adjustment selection functions. Future research could involve extensive sensitivity analyses and varying sample sizes to enhance the robustness of the findings.

4.2 Conclusion

This study highlights the significant impact of ORB on treatment effect estimation and demonstrates the efficacy of a flexible ORB-adjustment framework. Despite limitations, the methodology shows promise in mitigating ORB across various settings, with potential for further refinement and broader application.

References

[1] Bay, R., Liu, X., Lin, L., Liu, Y., Kimmel, S., Chu, H., and Chen, Y. (2021). A bayesian selection model for correcting outcome reporting bias with application to a meta-analysis on heart failure interventions.

[2] Begg, C. and Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, **50**, 1088–1101.

[3] Bresnah, R., Hounsome, J., Jette, N., Hutton, J., and Marson, A. (2019). Topiramate add-on therapy for drug-resistant focal epilepsy.

[4] Chan, A., Krleža-Jerić, K., Schmid, I., and Altman, D. (2004a). Outcome reporting bias in randomized trials funded by the canadian institutes of health research. *CMAJ*, **171**, 735–740.

[5] Chan, A.-W., Hróbjartsson, A., Haahr, M. T., Gøtzsche, P. C., and Altman, D. G. (2004b). Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *JAMA*, **291**, 2457–2465.

[6] Cooper, H., Hedges, L., and Valentine, J. (2009). *The handbook of research synthesis and meta-analysis*. Russell Sage Foundation.

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

Figure 3: Bias in the estimation of the treatment effect for ORB simulated with $\gamma = 1.5$

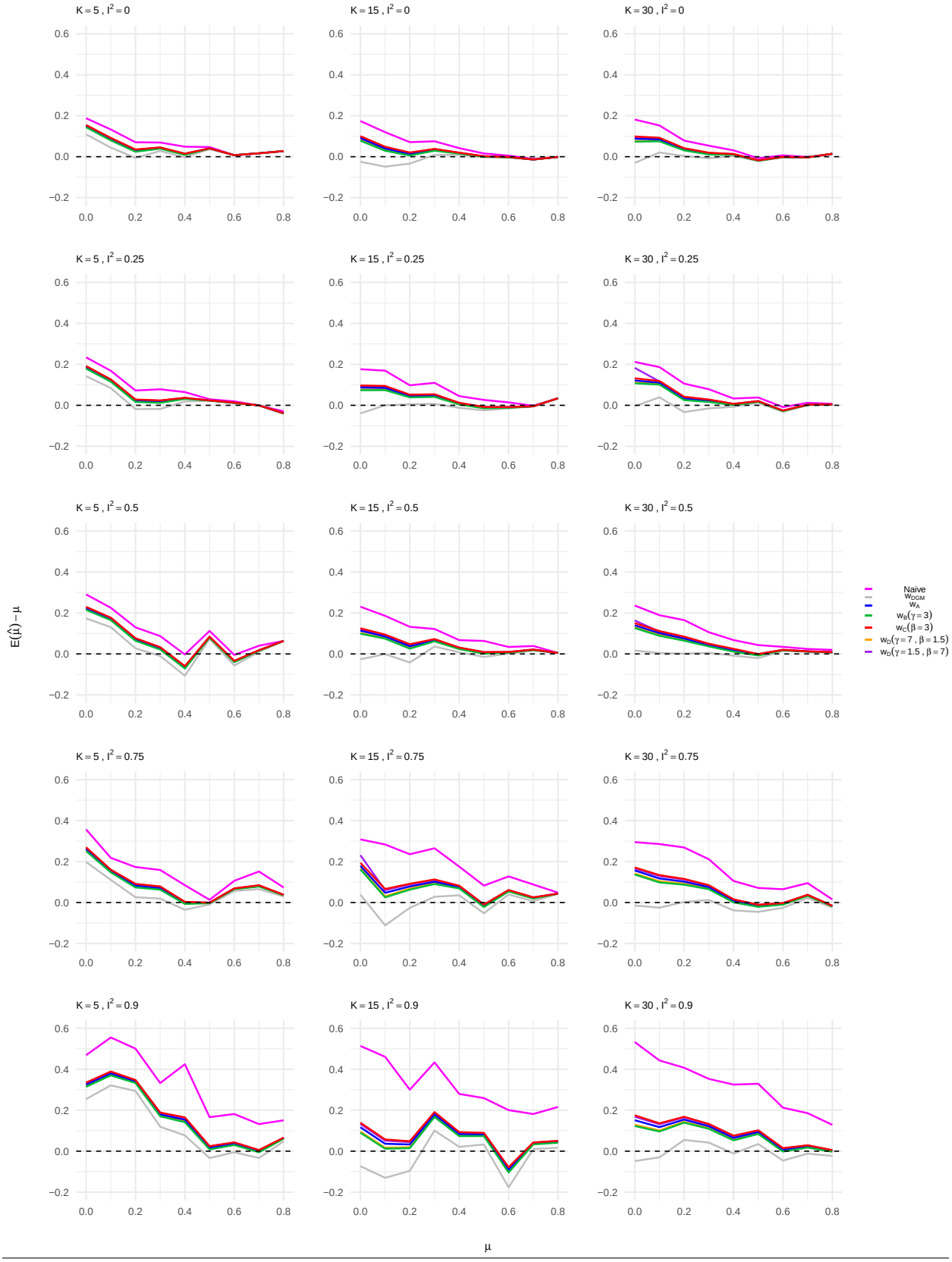


Figure 4: Bias in the estimation of the treatment effect for ORB simulated with $\gamma = 0.5$

[8] Copas, J. and Henmi, M. (2010). Confidence intervals for random effects meta-analysis and robustness to publication bias. *Statistics in Medicine*, **29**, 2969–2983.

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

[10] Dear, K. B. G. and Begg, C. B. (1992). An approach for assessing publication bias prior to performing a meta-analysis. *Statistical Science*, **7**, 237–245.

[11] DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Control Clinical Trials*, **7**, 177–88.

[12] Egger, M., Smith, G. D., and Higgins, J. (2022). *Systematic reviews in health research: Meta-analysis in context*. Wiley-Blackwell, Chichester, England, 3 edition.

[13] Fernández-Castilla, B., Declercq, L., Jamshidi, L., Beretvas, S. N., Onghena, P., and Van den Noortgate, W. (2019). Detecting selection bias in meta-analyses with multiple outcomes: A simulation study. *The Journal of Experimental Education*, , 1–20.

[14] Hedges, L. V. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science*, **7**, 246–255.

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

[16] Held, L. and Bové, D. S. (2021). *Likelihood and Bayesian Inference*. Statistics for Biology and Health. Springer, 2 edition.

[17] IntHout, J., Ioannidis, J. P. A., and Borm, G. F. (2014). The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Medical Research Methodology*, **14**, 25.

[18] Iyengar, S. and Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science*, , 109–135.

[19] Kirkham, J., Dwan, K., Altman, D., Gamble, C., Dodd, S., Smyth, R., and Williamson, P. (2010). The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. *BMJ*, **340**, 365.

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

[21] Littell, J. H., Gorman, D. M., Valentine, J. C., and Pigott, T. D. (2023). Protocol: Assessment of outcome reporting bias in studies included in campbell systematic reviews. *Campbell Systematic Reviews*, **19**, e1332.

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

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

[24] Morris, T., White, I., and Crowther, M. (2019). Using simulation studies to evaluate statistical methods.

[25] Normand, S. (1999). Meta-analysis: formulating, evaluating, combining, and reporting. *Statistics in Medicine*, **18**, 241–363.

[26] Preston, C., Ashby, D., and Smyth, R. (2004). Adjusting for publication bias: modelling the selection process. *Journal of Evaluation in Clinical Practice*, **10**, 313–322.

[27] Saini, P., Loke, Y., Gamble, C., Altman, D., Williamson, P., and Kirkham, J. (2014). Selective reporting bias of harm outcomes within studies: findings from a cohort of systematic reviews.

[28] Saracini, A. G. (2023). Addressing outcome reporting bias in meta-analysis: A comprehensive review and future directions.

[29] Schmid, C. H., Stijnen, T., and White, I. R. (2022). *Handbook of meta-analysis*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, London, England.

[30] Smyth, R. M. D., Kirkham, J. J., Jacoby, A., Altman, D. G., Gamble, C., Williamson, P. R., and et al. (2011). Frequency and reasons for outcome reporting bias in clinical trials: Interviews with trialists. *BMJ*, **342**, c7153.

[31] Sutton, A. J., Song, F., Gilbody, S. M., and Adams, K. R. (2000). Modelling publication bias in meta-analysis: a review. *Statistical Methods in Medical Research*, **9**, 421–45.

[32] van Aert, R. and Wicherts, J. (2023). Correcting for outcome reporting bias in a meta-analysis: A meta-regression approach. Behav Res.

[33] Viechtbauer, W. (2006). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine*, **26**, 37–52.