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

Alessandra Gaia Saracini[1] and Leonhard Held[2]

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 to investigate ORB-adjustment techniques through a selection model lens, 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.

Keywords: Outcome Reporting Bias (ORB); ORB-adjustment; Selection Model; Selection Function; Meta-analysis.

1 Introduction

Meta-analysis is a powerful statistical tool used to combine evidence from multiple studies investigating the same research question [14, 9]. 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 [9, 41, 16]. 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 [9, 41, 16, 29]. 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 [27, 10, 12, 41, 28, 16, 46]

[27, 10, 12, 41, 28, 16, 46, 39, 45]. Therefore, unlike PB, studies affected by ORB may still be published, but certain outcomes, especially those with weaker results, 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 [7, 27, 39][7, 27, 39, 42, 45]. An investigation on a cohort of Cochrane systematic reviews by Kirkham *et al.* [27] 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 [27]. An investigation by Saini *et al.* [39], 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 [39]. Even prospective trial registration, intended to discourage selective reporting and enhance transparency, does not always succeed in preventing ORB. Trial registries often lack statistical analysis plans and are not externally audited, allowing researchers to adapt analyses or selectively report outcomes based on significance [45, 42]. This limitation is reflected in the findings

---

[1]Corresponding author: Alessandra Gaia Saracini
University of Zurich
Epidemiology, Biostatistics and Prevention Institute
ETH Zurich Department of Mathematics
alessandragaia.saracini@gmail.com
[2]Leonhard Held, Professor
University of Zurich
Epidemiology, Biostatistics and Prevention Institute
leonhard.held@uzh.ch

of Silva et al. [42], who examined 84 prospectively registered trials and found that nearly half showed evidence of selective outcome reporting, with over 20% omitting a registered primary outcome entirely. A study by Chan et al. [7], 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 [7]. Similar investigations on RTCs estimated that statistically significant outcomes were 2.2–4.7 times more likely to be fully reported compared with outcomes that were not statistically significant [45, 15]. ORB poses a substantial threat to the integrity of meta-analyses, emphasizing the need for increased awareness and methods to mitigate its impact. ~~Currently available statistical methodology to adjust for ORBdiffers in nature and underlying assumptions, see the works of Kirkham et al. [28], Copas et al. [10, 12], Bay et al. [2], including a~~

Omitting study outcomes based on their effect size or significance, leading to ORB, can be equivalently understood from a missing data perspective. Through this lens, ORB reflects a missing not at random (MNAR) scenario, meaning that the probability of an outcome being (un)reported depends on its value, or its statistical significance. This contrasts with the assumption of missing completely at random (MCAR), where missingness is unrelated to the data. Standard meta-analytic methods, which only include reported outcomes in an analysis, implicitly assume a MCAR setting, leading to biased estimates when the true mechanism is MNAR [22, 30, 46].

A related but distinct setting is missing at random (MAR), where the reporting of one outcome depends on other observed outcomes within the same investigation, which can also lead to ORB [46, 22, 30]. This is particularly relevant in ORB-adjustments in multivariate meta-analysis, where methods have been developed to leverage correlations between outcomes and borrow strength across them in case of missing information [28, 2, 46, 22, 30]. Key examples include the bivariate meta-analysis ~~adjustment of two correlated outcomes [28], a Bayesian extension of it [2], and a meta-regression approach [46]. The most established~~ of Kirkham et al. [28] and its Bayesian extension by Bay et al. [2], as well as expansions to indirect treatment comparisons and network meta-analysis by Hwang and DeSantis [22] and Liu et al. [30]. These approaches have shown to

reduce bias under MAR, especially when outcomes are strongly correlated. However, estimating the between and within study correlation structures is often difficult. Some methods address this challenge by assuming a global correlation coefficient [28, 2, 46, 22, 38, 25], estimated either via Pearson's correlation [28, 46], or, in the case of Liu et al. [30] using copula models, e.g., the Clayton copula, or via non-informative priors in a Bayesian setting [2, 30, 22]. Although these strategies offer partial improvements under MNAR, they generally do not model the MNAR mechanism directly and may inadvertently introduce bias into the estimated effects of outcomes which are fully observed due to the joint modeling assumptions.

In contrast, the most established MNAR-based ORB-adjustment ~~method, i.e., that of Copas et al. [12], relies on~~ is the method proposed by Copas et al. [12, 10], which adjusts each outcome separately and explicitly models the missing data mechanism. The method relies on first categorizing ~~van Aert and Wicherts [46]~~ 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. [12] developed a~~ classification system. Assuming these categories are correctly assigned, the likelihood-based ~~ORB-adjustment method by adding a contribution from unreported study outcomes classified as HR of bias to the likelihood function~~model includes contributions from unreported HR outcomes, under the assumption that these ~~were originally~~ would have been non-significant. ~~In the Copas et al. [12] 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~~Of note, the ORBIT classification is specific to ORB as it evaluates the risk of bias at the level of each unreported outcome within a study, rather than at the study level. In contrast, study level risk of bias assessments, typical of PB [31, 21, 3, 33], are less suitable for ORB, since published studies may contain multiple unreported outcomes, each with potentially different levels of risks of ORB [12, 10, 28].

Our work ~~can be seen as an extension of the Copas et al. [12] method by presenting~~ builds on this foundation by reinterpreting ORB-adjustment through ~~a selection model perspective~~the lens of selection models, a framework ~~typically~~ commonly used for PB adjustment [13, 18, 44]. ~~The proposed approach for ORB adjustment offers a~~

Our proposed method removes the need for ORBIT classification, allows flexible modeling of the missing data mechanism, and incorporates all unreported outcomes in the analysis, not just the ones classified as HR of bias by the ORBIT methodology. We also investigate the influence of between-study heterogeneity a novel aspect in the ORB literature - and assess the effectiveness of our method through simulations. The focus of the simulations is on treatment effect estimation, with a secondary emphasis on how heterogeneity interacts with ORB and its correction.

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

As a motivating example of ORB in meta-analysis, we consider the data used by Copas *et al.* [12], 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.* [4], 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 our research 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 the treatment and control arm sample sizes; however, some studies do not report the event frequencies, from which the log risk ratio (RR) is computed and used as the normally distributed treatment effect in (1), using a continuity correction in case of empty cell counts [12].

Table 1: Results for the outcomes "50% Seizure Reduction" and "Seizure Freedom" from 12 studies investigating the effect of Topiramate as an add-on treatment (T) vs. control (C) for drug-resistant focal epilepsy [12, 4].

Study ID	Sample Size		50% Seizure Reduction		Seizure Freedom	
(Author, date of publication)	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 manuscript is organized as follows: Section 2 introduces the selection model framework typically used for PB and illustrates the derivations done to adapt this framework to address ORB. In subsection 2.1 we elaborate on the various missing data mechanisms considered in our investigations and used in our simulation study, inspired by the PB literature and adapted to ORB. 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. Subsection 3.1, 3.2, and 3.3 describe the simulation setting, key results and the application to the epilepsy example from Table 1, respectively. Finally, Section 4 summarizes the proposed methodology and simulation study findings in a discussion, including limitations and conclusions.

2   Selection Models for ORB

Selection models have gained popularity in the publication bias (PB) adjustment literature [14, 18, 13, 44, 9], 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 weighting them with a selection function $w(y_i)$ which describes the probability that study i is published/selected based on its significance [18, 13, 44]. By using the following relation:

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

the PB-adjusted log-likelihood $\ell_{\text{Adj}}^{\text{PB}}(\theta)$ is derived [24, 19, 18, 13] as

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

The selection function can take various forms in the context of PB, generally guided by the intuition that in a meta-analysis of a beneficial outcome, for larger p-values, the probability of publication/selection decreases [24, 19, 18, 13, 6, 37, 44]. Of note, in the case of a meta-analysis of a harmful outcome, we expect the opposite: small, significant p-values are less likely to be reported, as they would indicate harm [39, 12]. In the following sections, we define the selection functions assuming beneficial outcomes and thus a positive treatment direction.

The selection function $w(y_i)$ in (2) is therefore often defined as a function of the p-value p_i, providing an intuitive way of understanding the relationship between significance and the probability of selection [18, 13, 44]. Given that the p-value is simply a transformation of the observed treatment effect y_i and the standard error σ_i, we will use the notation $w(p_i)$ when providing definitions of the selection function in terms of the p-value.

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~~ outcome reporting bias (ORB) adjustment methods, according to the framework developed by Copas *et al.* [12], 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

$$\begin{aligned} \ell_{\text{Adj}}^{\text{ORB}} &= \sum_{i=1}^{K} \ell(\theta) \\ &= \sum_{i \in \{\text{Rep}\}} \ell(\theta) + \sum_{i \in \{\text{Unrep}\}} \ell(\theta) \\ &= \sum_{i \in \{\text{Rep}\}} \log f(y_i; \theta) + \sum_{i \in \{\text{Unrep}\}} \log f(y_i; \theta). \end{aligned}$$
$$(3)$$

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:

$$f(y_i; \theta \mid i \in \{\text{Rep}\}) = \frac{f(y_i; \theta) \cdot w(y_i)}{\int_{-\infty}^{\infty} f(y; \theta) \cdot w(y) dy}. \quad (4)$$

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

$$f(y_i; \theta \mid i \in \{\text{Unrep}\}) = \frac{f(y_i; \theta) \cdot (1 - w(y_i))}{\int_{-\infty}^{\infty} f(y; \theta) \cdot (1 - w(y)) dy}. \quad (5)$$

Using (4) and (5), and solving for $f(y_i; \theta)$, we can re-write the ORB-adjusted log-likelihood (3) as

$$\begin{aligned} \ell_{\text{Adj}}^{\text{ORB}}(\theta) &= \sum_{i \in \{\text{Rep}\}} \log f(y_i; \theta) \\ &\quad - \sum_{i \in \{\text{Rep}\}} \log \left[\int_{-\infty}^{\infty} f(y; \theta) \cdot w(y) dy \right] \\ &\quad + \sum_{i \in \{\text{Unrep}\}} \log \left[\int_{-\infty}^{\infty} f(y; \theta) \cdot (1 - w(y)) dy \right]. \end{aligned}$$
$$(6)$$

The log-likelihood (6) 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.* [12] model formulation, specific assumptions were made regarding the missing data mechanism, which result in a simplification of (6). For the reported outcomes, Copas *et al.* [12] 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, (6) can be further simplified to

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

This is thus the generic form for ~~ORB adjustment~~ORB-adjustment, which has different shapes depending on the selection function $w(y_i)$ used, representative of the missing ~~data~~ not at random (MNAR) mechanism assumed for unreported study outcomes. Given the alignment of our ~~ORB adjustment~~ ORB-adjustment with the PB framework of selection models, one can use similar selection functions which are typically found in the PB literature.

2.0.1 ~~Selection Functions~~

2.1 Selection Functions

We present a series of selection functions implemented in our methodology, defined as functions of the one-sided p-value, $p = \Phi(-y/\sigma)$, where α is the threshold for significance, e.g., $\alpha = 0.05$, and the study index i is omitted for ease of notation. As previously noted, given that the selection mechanism in the following section is defined as a function of the p-value, we use the notation $w(p)$ instead of $w(y)$.

2.1.1 Piecewise constant

We choose a one-sided p-value to model the probability of selection, in alignment with selection models of beneficial outcomes in PB [19, 44, 37]. One of the simplest selection functions used for PB is:

$$
w_A(p) = \begin{cases} 1 & \text{if } p \le \alpha \\ 0 & \text{if } p > \alpha, \end{cases}
\tag{8}
$$

While this selection function can be found in the PB literature [24, 18, 44], we note that it is also the one implicitly used in the Copas *et al.* [12] adjustment, although the authors do not explicitly frame the ~~ORB adjustment~~ QRB-adjustment via a selection model framework. Of note, in Copas *et al.* [12], ~~ORB adjustment~~ QRB-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 at LR of bias and regard them as missing at random. Furthermore, the authors use the two-sided p-value $p = 2 \cdot (1 - \Phi(\left| \frac{y}{\sigma} \right|))$ 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 [19, 44, 37].

Using the log-likelihood (7) and the selection function (8) for a subset of the unreported studies, i.e., those classified as HR of bias, along with a two-sided p-value instead of a one-sided one, we can easily see how we obtain the simplified ORB-adjusted log-likelihood presented for the random effects model in Copas *et al.* [12], namely:

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

2.1.2 Constant-decreasing

The selection function (8) 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 [18, 44]. One example is the function $w_B(p; \beta)$ with tuning parameter $\beta > 0$:

$$
w_B(p; \beta) = \begin{cases} 1 & \text{if } p \le \alpha \\ \frac{p^{-\beta}}{\alpha^{-\beta}} & \text{if } p > \alpha \end{cases}
\tag{9}
$$

The idea of this selection function in the context of PB is that the associated probability of publishing, which weighs observations, is greater than 0 for non-significant outcomes. Specifically, when applied to ORB, for the unreported study outcomes which were originally non-significant, the underlying probability of reporting is a decreasing function of the p-value, while significant study outcomes have an associated probability of reporting equal to 1.

2.1.3 Decreasing-constant

In the context of ORB we further propose a different selection function, $w_C(p; \gamma)$ with tuning parameter $\gamma > 0$ presented in (10), for which the rationale is inverted compared to $w_B(p; \beta)$ in (9). With selection function $w_C(p; \gamma)$ we assume that non-significant study outcomes have an associated probability of reporting equal to 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 [8, 43], 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 closer to missing completely at random (MCAR) and less bias [40]. 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.

$$w_C(p; \gamma) = \begin{cases} 1 - \frac{p^\gamma}{\alpha^\gamma} & \text{if } p \leq \alpha \\ 0 & \text{if } p > \alpha \end{cases} \quad (10)$$

2.1.4 Piecewise decreasing

Based on the selection functions $w_B(p; \beta)$ in (9) and $w_C(p; \gamma)$ in (10) ~~one could~~ we envisage a combination of these by using e.g., selection function $w_D(p; \beta, \gamma)$ in (11). 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 (9), ω_α was implicitly 1 and in case of (10) this was set to 0. Here, we set $\omega_\alpha = 0.5$, as a middle value between (9)

and (10). The selection function $w_D(p; \beta, \gamma)$ has the potential of being used to conduct extensive sensitivity analyses when adjusting for ORB.

$$w_D(p; \beta, \gamma) = \begin{cases} 1 - (1 - \omega_\alpha)\left(\frac{p^\gamma}{\alpha^\gamma}\right) & \text{if } p \leq \alpha \\ \omega_\alpha \left(\frac{p^{-\beta}}{\alpha^{-\beta}}\right) & \text{if } p > \alpha \end{cases} \quad (11)$$

The selection functions proposed above, namely $w_A(p)$ in (8), $w_B(p; \beta, \gamma)$ in (9), $w_C(p; \gamma)$ in (10) and $w_D(p; \beta, \gamma)$ in (11) 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.

2.1.5 ~~Imputation of Missing Variances~~

~~When utilizing any of the selection functions presented in the ORB-adjusted log-likelihood (7), we require knowledge of the standard error of the unreported study outcome, which is generally missing. This value hence needs to be imputed; we follow the methodology of Copas et al. [12] and Bay et al. [2] and impute the missing standard error of an unreported study i as~~

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

~~where n_i is the sample size of study i and~~

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

With the selection model framework for ~~ORB adjustment~~ 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 as the heterogeneity variance.

2.2 Imputation of Missing Variances

When utilizing any of the selection functions presented in the ORB-adjusted log-likelihood (7), we require knowledge of the standard error of

the unreported study outcome, which is generally missing. This value hence needs to be imputed; we follow the methodology of Copas *et al.* [12], Kirkham *et al.* [28], and Bay *et al.* [2] and impute the missing standard error of an unreported study i as

$$\sigma_i^2 = \frac{1}{\hat{k} n_i},\qquad(12)$$

where n_i is the sample size of study i and

$$\hat{k} = \frac{\sum_{i \in \{\text{Rep}\}} \sigma_i^{-2}}{\sum_{i \in \{\text{Rep}\}} n_i}.\qquad(13)$$

The rationale behind this imputation is grounded in the relationship between effect size variances and sample size. For commonly used measures such as the log risk ratio (log RR), the variance is inversely proportional to the sample size. However, this relationship is also influenced by the design of the trial, e.g., the balance between arms and the event rates, captured by the trial-specific design factor $k_i = \sigma_i^{-2}/n_i$. While this factor varies across trials, it is assumed that the average design factor for the unreported studies is similar to that of the reported studies. Hence, we estimate this average using (13) and apply it in (12) to impute missing variances [12, 10]. This imputation strategy is particularly suitable in our setting because the total sample size n_i is typically known even when the outcome is not reported, due to the nature of ORB compared to, e.g., PB, where the whole study is missing from the literature.

Application of ORB-adjustment to example epilepsy data [12, 4], using the different selection functions showcased in Figure 1. In addition, the naive estimate, without ORB-adjustment, is shown for comparison. The plot shows the risk ratio (RR) estimates and 95% profile likelihood (PL) confidence intervals (CI).

2.2.1 Application to Motivating Example

We apply the ORB-adjustment framework to the epilepsy meta-analysis data affected by ORB from Bresnah *et al.* [4], Copas *et al.* [12], using the selection functions proposed in this research. These selection functions utilize a one-sided p-value for significance with a threshold of $\alpha = 0.05$, in contrast to Copas *et al.* [12], which used a two-sided p-value. Of note, while a one-sided threshold is used within the

selection function to define the underlying missing data mechanism, two-sided significance is used to construct profile likelihood (PL) confidence intervals (CIs) for treatment effect estimation.

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, i.e., that obtained with standard meta-analysis methods, thereby excluding any contribution from unreported study outcomes, shows a significant positive treatment effect compared to the control.

For the 50% seizure frequency reduction outcome, the ORB-adjusted estimates are slightly shifted towards the null value and are consistent across different selection functions. Only a minor shift is expected since only one study does not report this outcome. However, for the seizure freedom outcome, with several studies not reporting it, the ORB-adjusted estimates show a substantial shift towards the null, even altering the significance of the results by causing the 95% CI to overlap with 0. The differences between the ORB-adjusted estimates using various selection functions are more pronounced for the seizure freedom outcome.

The strictness of different ORB-adjustments, obtained by using the various selection functions, is intuitive and stems from the underlying assumptions about unreported study outcomes. The estimate obtained using selection function $w_B(\beta = 3)$ 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 $w_B(\beta = 3)$ assumes a higher probability of unreporting for larger p-values, implying greater bias and thus stricter correction. Conversely, $w_C(\gamma = 3)$ 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. The functions $w_D(\gamma = 1.5, \beta = 7)$ and $w_D(\gamma = 7, \beta = 1.5)$ produce results similar to $w_B(\beta = 3)$ and $w_C(\gamma = 3)$, respectively, suggesting that the larger parameter, whether γ or β, primarily drives the ORB-adjusted 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.

Figure 1: Possible Selection Functions for ORB-adjustment. Function $w_A(p)$ from equation (8) in (a), function $w_B(p; \beta = 3)$ from equation (9) in (b), function $w_C(p; \gamma = 3)$ from equation (10) in (c), and functions $w_D(p; \beta = 1.5, \gamma = 7)$ and $w_D(p; \beta = 7, \gamma = 1.5)$ from equation (11) shown in (d).
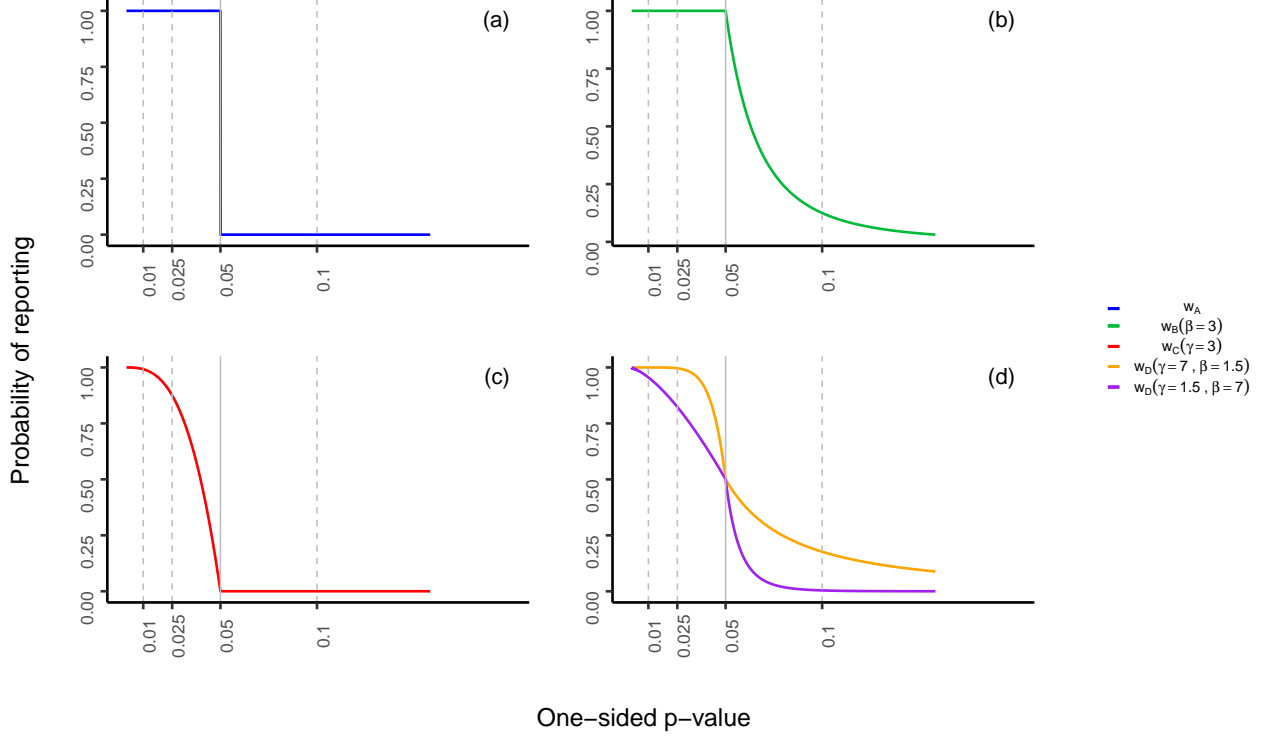
3 Simulation Study

It is of interest to assess the extent to which ~~ORB~~ outcome reporting bias (ORB) negatively impacts meta-analytic findings and the extent to which the ORB-adjustment methodology presented in the previous section of this work is effective in reducing bias. Our primary interest lies in 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. In pursuit of these objectives, 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 some observed treatment effects and standard errors from the meta-analysis dataset based on the strength and/or direction of the results, favoring the reporting of studies with small p-values. The resulting data-generating mechanism (DGM) thus follows a missing not at random (MNAR) pattern, as reporting depends on statistical significance. 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 section.

3.1 Setting

The first step of the simulation process consists of 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

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

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

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

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

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

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 (14) 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., $\gamma = 1.5$ typical in PB simulation studies [13, 3, 32, 37] and $\gamma = 0.5$, resulting in a steeper decreasing function of the p-value.

$$P(i \in \{\text{Rep}\}) = e^{-4 \cdot p_i^\gamma}. \tag{14}$$

Each meta-analysis dataset hence results in K or fewer of the original study outcomes. If for some meta-analysis dataset less than two study outcomes are reported, the simulation is repeated until at least two reported study outcomes are obtained [13, 3, 32, 17]. 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 $\mu \in \{0, 0.2, 0.4, 0.6, 0.8\}$, based on simulation studies found in the literature [34, 23, 17].

It is important to note that while our simulation uses a known missing data mechanism, i.e., selective reporting based on a continuous function of the one-sided p-value, no information about the cause of missingness is assumed to be available once the data are generated. This reflects the practical setting in which some outcomes are unreported, but their classification into, e.g., high risk (HR) or low risk (LR) of bias as per the ORBIT methodology is unknown. This differs from the simulation study in our previous work [40], where unreported outcomes were simulated mimicking the ORBIT classification into HR and LR of bias, and this information was assumed to be available prior to the ORB-adjustment. We deem our current way of simulating ORB, as a continuous function of the p-value, more realistic, as it is consistent with the selection model literature, less arbitrary than mimicking HR/LR of bias missingness, and, overall, more aligned with the goal of avoiding the need to for an ORBIT classification for ORB-adjustment. As a result however, the Copas et al. [12] method, which requires the ORBIT classification and adjusts only for outcomes deemed at HR of bias, is not applicable in this setting and thus not directly comparable to our methodology.

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) [47, 12, 20]. 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 [44, 12]. 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 (7) using the different selection functions: $w_A(p)$ from (8), $w_B(p; \beta = 3)$ from (9), $w_C(p; \gamma = 3)$ from (10), and $w_D(p; \beta = 1.5, \gamma = 7)$, $w_D(p; \beta = 7, \gamma = 1.5)$ from (11), as well as the selection function (14) used to simulate ORB, so as to include the correct model specification in the adjustment. Since the latter function used to simulate ORB can be viewed as a selection function, defining a specific missing data mechanism, we note it as $w_{DGM}(y)$ and utilize it in the ORB-adjusted log-likelihood (7). 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 $N_{\text{sim}} = 3200$ times; the simulation size N_{sim} is calculated based on the expected variance of the unknown parameter estimate [23, 35] and a desired Monte Carlo Standard Error (MCSE) of 0.005 from IntHout et al. [23], Morris et al. [35]. The per-

formance measures recorded for the unknown parameter are bias, empirical standard error (ESE), mean squared error (MSE), coverage and power, along with the MCSEs of each [23, 35].

To provide a methodological benchmark, we also simulate an alternative missing data mechanism under MCAR. Specifically, for each ORB setting in the above MNAR simulation process with $\gamma = 1.5$, we generate a corresponding MCAR scenario by removing the same number of studies as were dropped due to ORB, but at random and independently of their p-values. The estimation methods remain unchanged, with the naive estimation now aligned with the DGM under MCAR, while all ORB-adjusted methods are misspecified, as they assume a MNAR under MCAR. This methodological variation enables an assessment of the robustness of our main simulation process and ORB-adjustment approaches when the missingness is not due to ORB. A synthesis of the results is presented in Section 3.2.5, with additional details available in the supplementary material.

3.2 Results

3.2.1 Bias in Naive Estimation

The results demonstrate a significant bias in the estimation of the treatment effect when using naive methods that do not account for ORB, as shown in Figures 4 and 5. The bias decreases as the true treatment effect size μ increases, which aligns with existing literature [12, 2, 46] and prior exploratory analysis [40]; as treatment effect increases, results are more likely to be significant and are thus less prone to ORB. Study size variations ($K = 5, 15, 30$) do not significantly affect the bias, while heterogeneity has a substantial impact. High heterogeneity settings, particularly with $I^2 = 90$, exhibit larger biases, reinforcing findings from previous work [40]. The effect of heterogeneity on ORB is interesting and novel compared to Copas *et al.* [12], who focused primarily on a fixed effect meta-analysis framework. The patterns observed for naive estimation are consistent across both ORB simulation processes, i.e., for $\gamma = 1.5$ and $\gamma = 0.5$ in the DGM function (14).

3.2.2 Bias Reduction with ORB-adjustment

When applying the ORB-adjustment framework using selection functions, we first note that the effectiveness in bias reduction varies by meta-analysis study size. For $K = 15$ and $K = 30$, the bias obtained with naive estimation is eliminated when the selection function matches the ORB DGM, i.e., when us-

ing the correctly specified selection function w_{DGM} in Figures 4 and 5. Different selection functions (w_A, w_B, w_C, w_D) show varying degrees of bias reduction. For the DGM with $\gamma = 1.5$, shown in Figure 4, the ORB-adjusted estimates shift the bias towards the null but do not fully eliminate it unless the exact DGM function is used. The ORB-adjusted estimate using selection function w_B performs slightly better than w_A, and w_C performs the least well. However, it is important to note that the different functions (w_A, w_B, w_C, w_D) overall achieved very similar results, particularly in low heterogeneity settings. Similar patterns are observed in the DGM with $\gamma = 0.5$ setting shown in Figure 5, with selection function w_B being the least strict and w_C the most strict in bias reduction. The ORB-adjustment here tends to reduce the treatment effect size excessively, indicating potential overcorrection due to the steep p-value dependence in the ORB DGM, resulting in some unreported studies with significant p-values.

For the small meta-analysis size, $K = 5$, the ORB-adjustment reduces the bias but does not eliminate it, even with the correctly specified model. This finding holds in general for both ORB DGM settings; notably for the DGM with $\gamma = 1.5$, the correctly specified selection function shows the least bias, while for the DGM with $\gamma = 0.5$, it shows the most bias, although, overall, the ORB-adjusted estimates are similar. Based on these observations, we thus recommend to use ORB-adjustment with caution when only few studies are present in a meta-analysis affected by ORB.

3.2.3 Other Performance Measures

Beyond bias, other performance measures such as coverage, mean squared error (MSE), power, and empirical standard error (ESE) were evaluated. Coverage, shown for the ORB DGM $\gamma = 1.5$ in Figure 6, can be substantially low for naive estimation. Overall, the coverage in naive estimation decreases as heterogeneity increases. For small treatment effect sizes (e.g., $\mu = 0$), coverage is higher for small meta-analysis sizes ($K = 5$) and decreases as the meta-analysis size increases. This can be explained by larger CI for the $K = 5$ setting, which in turn cover the true underlying value. The ORB-adjusted estimates show higher coverage: the correct DGM selection function has the highest one, while others exhibit slightly lower coverage, especially when μ is small.

Other performance measures confirm the findings observed for naive and ORB-adjusted estimates, e.g., the MSE of the naive estimate of the treatment effect

is substantially reduced in high heterogeneity settings for all ORB-adjusted estimates. Furthermore, naive estimation results in severely inflated power, particularly in high heterogeneity settings and for large meta-analysis study sizes. ORB-adjusted estimates correct this inflation, with variations depending on the DGM and selection function used. The ESE of the naive estimate is generally consistent with expected SE calculations from the simulation study protocol. Naive estimates have slightly higher SE due to unreported study outcomes. For small meta-analysis sizes ($K = 5$) and high heterogeneity ($I^2 = 90$), ORB-adjusted estimates have a similar ESE, which is lower than the naive estimate. For detailed results and further plots of the additional performance measures considered and briefly mentioned in this section, please refer to the supplementary material.

3.2.4 Bias in Heterogeneity Variance

Although the primary parameter of interest was the treatment effect μ, we also investigated the bias in the estimation of the heterogeneity variance τ^2 in the presence of ORB, as shown in Figure 7, showcasing the results for the ORB DGM with $\gamma = 1.5$. Heterogeneity is generally underestimated across most estimation methods, except for the ORB-adjusted method using the correctly specified model (w_{DGM}) for $K = 15, 30$. For the small meta-analysis setting $K = 5$, the correctly specified model reduces the bias but does not fully eliminate, similarly to results observed for the main parameter of interest μ. Of note, the estimation of τ^2 is done with maximum likelihood (ML) estimation, which, overall, tends to underestimate the between study-heterogeneity [11, 36]; a more comprehensive methodological approach to heterogeneity estimation in the presence of ORB should thus be conducted to solidify and confirm these findings. The plots of the heterogeneity variance for ORB setting $\gamma = 0.5$ can be found in the supplementary material.

3.2.5 Simulation under MCAR

Simulations under a MCAR mechanism (results presented in Figures 9-10 of the supplementary material) showed that with few studies ($K = 5$) and near-null true treatment effects, the naive estimator was biased, reinforcing the caution advised in our primary simulation regarding ORB-adjustment in small meta-analyses with unreported study outcomes. In larger meta-analyses, the naive estimator was unbiased under MCAR, while ORB-adjusted methods, as expected due to model misspecification, underestimated the treatment effect, but remained reasonably robust, particularly in settings with low heterogeneity. These findings further support the use of ORB-adjustment methods when ORB is suspected and underscore the importance of conducting sensitivity analyses, given the inherent uncertainty of the true missing data mechanism.

3.3 Application to Example Data

We apply the ORB-adjustment framework to the epilepsy example from from Bresnah et al. [4] and Copas et al. [12] introduced in Table 1. We adjust for ORB in the two beneficial outcomes of the example meta-analysis, i.e., 50% reduction in seizure frequency and seizure freedom using the selection functions proposed in this research and utilized in the simulation study. Figure 2 shows the results of applying the same parameters used in the simulation study to our example data. Figure 3 shows the results of an example sensitivity analysis for the ORB-adjusted estimates for a range of such parameters.

Figure 2 presents the point estimates and 95% CI for the RR of the treatment effect for the two 50% reduction in seizure frequency and seizure freedom. For both outcomes, the naive RR estimate, i.e., that obtained with standard meta-analysis methods, thereby excluding any contribution from unreported study outcomes, shows a significant positive treatment effect compared to the control.

For the 50% seizure frequency reduction outcome, the ORB-adjusted estimates are slightly shifted towards the null value and are consistent across different selection functions. Only a minor shift is expected since only one study does not report this outcome. However, for the seizure freedom outcome, with several studies not reporting it, the ORB-adjusted estimates show a substantial shift towards the null, even altering the significance of the results by causing the 95% CI to overlap with 0. The differences between the ORB-adjusted estimates using various selection functions are more pronounced for the seizure freedom outcome.

The strictness of different ORB-adjustments varies depending on the assumptions of each selection function. The estimate obtained using selection function $w_B(\beta = 3)$ from (9) is more conservative than that obtained with w_A in (8). The selection function w_A assumes a probability of unreporting of 1 for non-significant studies, regardless of the p-value magnitude, while $w_B(\beta = 3)$ assumes a higher probability of unreporting for larger p-values, implying greater bias and thus stricter correction.

Conversely, $w_C(\gamma = 3)$ from (10) 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 ORB-adjustment. The hybrid functions $w_D(\gamma = 1.5, \beta = 7)$ and $w_D(\gamma = 7, \beta = 1.5)$ combine these behaviors, and their estimates lie between those of w_B and w_C.

Given that the true missing data mechanism is unknown, we recommend a sensitivity analysis approach to ORB-adjustment, using different selection functions (e.g., $w_B(\beta)$ from (9), $w_C(\gamma)$ from (10) or $w_D(\gamma, \beta)$ from (11)) and varying parameters β and γ, which control how strongly the probability of unreporting depends on the p-value. Figure 3 showcases an example of a sensitivity analysis to ORB-adjustment using the three above-mentioned different selection functions and a range of β and γ.

From the results in Figure 3 we observe that the 50% seizure frequency reduction outcome, with only one unreported study, is minimally sensitive to parameter changes, while the seizure freedom outcome, with several missing results, exhibits greater variability. We note that smaller values of β shift the selection mechanism $w_B(\beta)$ away from the less strict piecewise constant function w_A (where all non-significant results have the same probability of unreporting of 1); therefore, smaller β lead to stricter ORB-adjustments. On the other hand, as γ increases, the function $w_C(\gamma)$ approaches the stricter behavior of w_A (where all significant results have the same probability of unreporting 0); hence, for lower γ values, we observe a less strict ORB-correction. The combined function $w_D(\gamma, \beta)$ reflects an interplay of both effects.

Overall, the parameter variations tested here ($\gamma = 0$ to 3, $\beta = 0$ to 3) do not drastically alter the results in Figure 2 but reveal how selection function shape and strictness can impact the magnitude of adjustment. Nevertheless, since the true mechanism driving outcome reporting bias is unknown in practice, we strongly recommend using a sensitivity analysis approach with multiple plausible selection functions and parameter combinations. This enhances the robustness of conclusions and helps quantify the potential impact of ORB on meta-analytic estimates.

4 Discussion

This study addresses Outcome Reporting Bias (ORB), where the significance of study outcomes influences their reporting, leading to overestimation of beneficial treatment effects in meta-analyses of clinical trials. We approached ~~ORB adjustment~~ ORB-adjustment through a selection model framework, a common method in publication bias (PB) literature. This methodology allows one to incorporate contributions from unreported study outcomes based on different assumed missing data mechanisms, specified via selection functions. Our proposed selection functions expand on existing methods, including those from previous works like Copas *et al.* [12], by being more flexible in the missing data mechanisms assumed, utilizing information from all identified unreported study outcomes, and jointly estimating both treatment effect and heterogeneity variance via a random effects model.
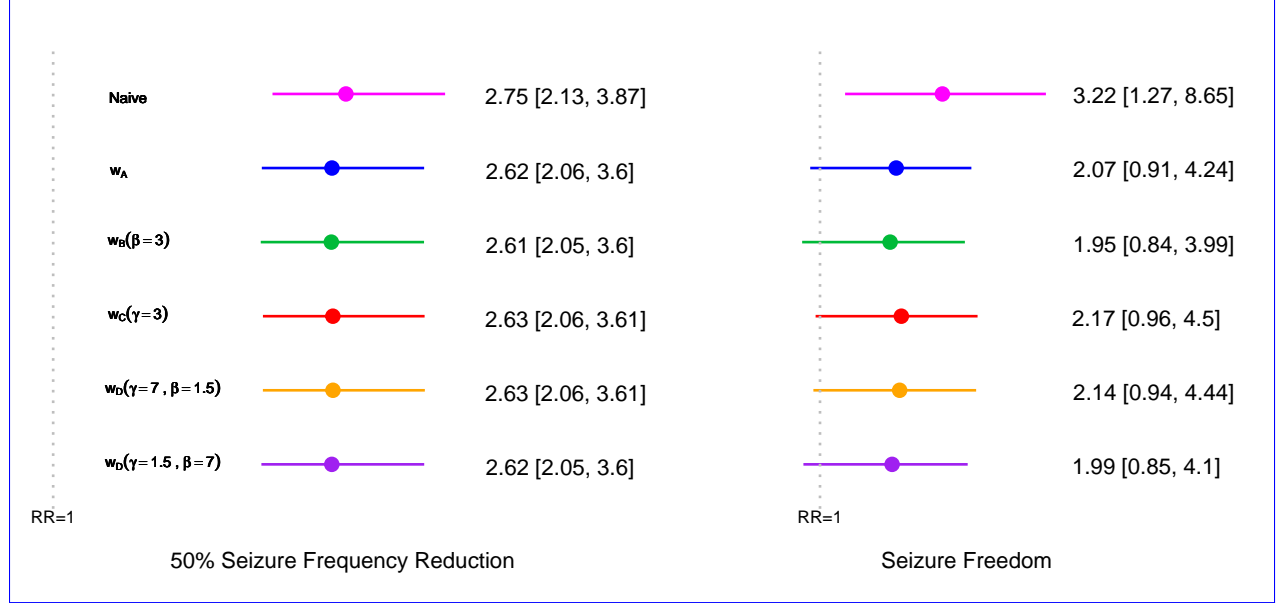
We ~~applied~~ conducted a simulation study on ORB and our ORB-adjustment methodology ~~to a real-world meta-analysis of epilepsy trials [12, 4] affected by ORB. The ORB-adjusted estimates of the treatment effect were substantially shifted towards null values, compared to the naive estimate, i.e., the standard estimation not accounting for ORB. This shift was particularly significant in the presence of numerous unreported study outcomes.~~ . The findings of our simulation study reveal several critical insights regarding the impact of ORB on the estimation of treatment effects and the efficacy of ORB-adjustment techniques. Naive estimation methods that do not account for ORB exhibit substantial bias, particularly in high heterogeneity settings, underscoring the importance of incorporating ~~ORB adjustments~~ ORB-adjustments.

Our results demonstrate that ORB-adjustment frameworks using selection functions can significantly reduce bias, although their effectiveness may vary with meta-analysis study size and the underlying method used to simulate ORB. For larger meta-analyses ($K = 15, 30$), correctly specified ORB-adjustment models effectively eliminate bias. Different misspecifications of the assumed missing data mechanism can be either slightly too lenient or slightly too strict, though their performance does not vary significantly. For smaller meta-analyses ($K = 5$), we must be cautious as bias reduction is limited even with correctly specified models.

Other measures of performance confirm these findings, demonstrating, e.g., substantial improvements in the coverage and power of the treatment effect estimates with ORB-adjustment. These findings highlight the necessity of using ORB-adjustment methods to achieve more accurate treatment effect estimates. Additionally, they suggest that heterogeneity estimation is impacted by ORB, warranting further attention to improve the robustness of meta-analyses in

Figure 2: Application of ORB-adjustment to the epilepsy example [12, 4], using the different selection functions shown in Figure 1. In addition, the naive estimate, without ORB-adjustment, is also given. The plot shows the risk ratio (RR) estimates and 95% profile likelihood (PL) confidence intervals (CI).

the presence of ORB.

We applied our ORB-adjustment methodology to a real-world meta-analysis of epilepsy trials [12, 4] affected by ORB and conducted an example sensitivity analysis approach under different strengths of ORB-adjustment. The ORB-adjusted estimates of the treatment effect were substantially shifted towards null values, compared to the naive estimate, i.e., the standard estimation not accounting for ORB. This shift was particularly significant in the presence of numerous unreported study outcomes. Varying the parameters in the sensitivity analysis did not strongly impact the ORB-adjusted estimates; however, we highlight that this was done on only one example data and for a limited set of parameters. Therefore, in practice we strongly recommend a sensitivity analysis approach to ORB-adjustment through selection models.

The ORB-adjustment methodology via selection models proposed in this research is flexible and broadly applicable. Although promising results have been observed, several limitations exist and should be noted to promote future research in this field. Firstly, our framework operates on individual outcomes in meta-analyses, not accounting for correlations between outcomes. Future research could explore methods to incorporate such correlations, although Bay et al. [2], Kirkham et al. [28] noted some the explicit modelling of the missing data mechanism through selection functions into the multivariate meta-analysis framework [30, 22, 2], as well as continue addressing the limitations in the estimation of correlations in the presence of ORB .-[2, 28, 22, 30] .˜

Another possible avenue of improvement in our current approach is the imputation of missing variances, as described in Section 2.2 and done in previous works [10, 12, 2, 40]. While this did not greatly impact our results due to equal study sizes in the simulation study setup, alternative ways to estimate the missing variances, e.g., via multiple imputation, could be considered [26, 48]. Additionally, our ORB-adjustment methodology assumes normally distributed outcomes, which might not be precise for binary data [10, 12], especially in cases of zero/low event numbers such as the epilepsy data example from Copas et al. [12]. Exploring a binomial likelihood for ORB-adjustment could be a potential avenue, as noted in Saracini [40]. In Saracini [40] we set-up the binomial likelihood contribution of reported studies, which can be extended to include a term from unreported studies with a specified probability of reporting.

We established that heterogeneity variance estimation is affected by ORB, and, at the same time, the true underlying heterogeneity influences the bias in the treatment effect estimate due to ORB. Therefore, considering heterogeneity in ORB and ORB

13

Figure 3: Application of ORB-adjustment sensitivity analysis to the epilepsy example [12, 4], using selection function (9) in (a), selection function (11) in (b), and selection function (10) in (c).
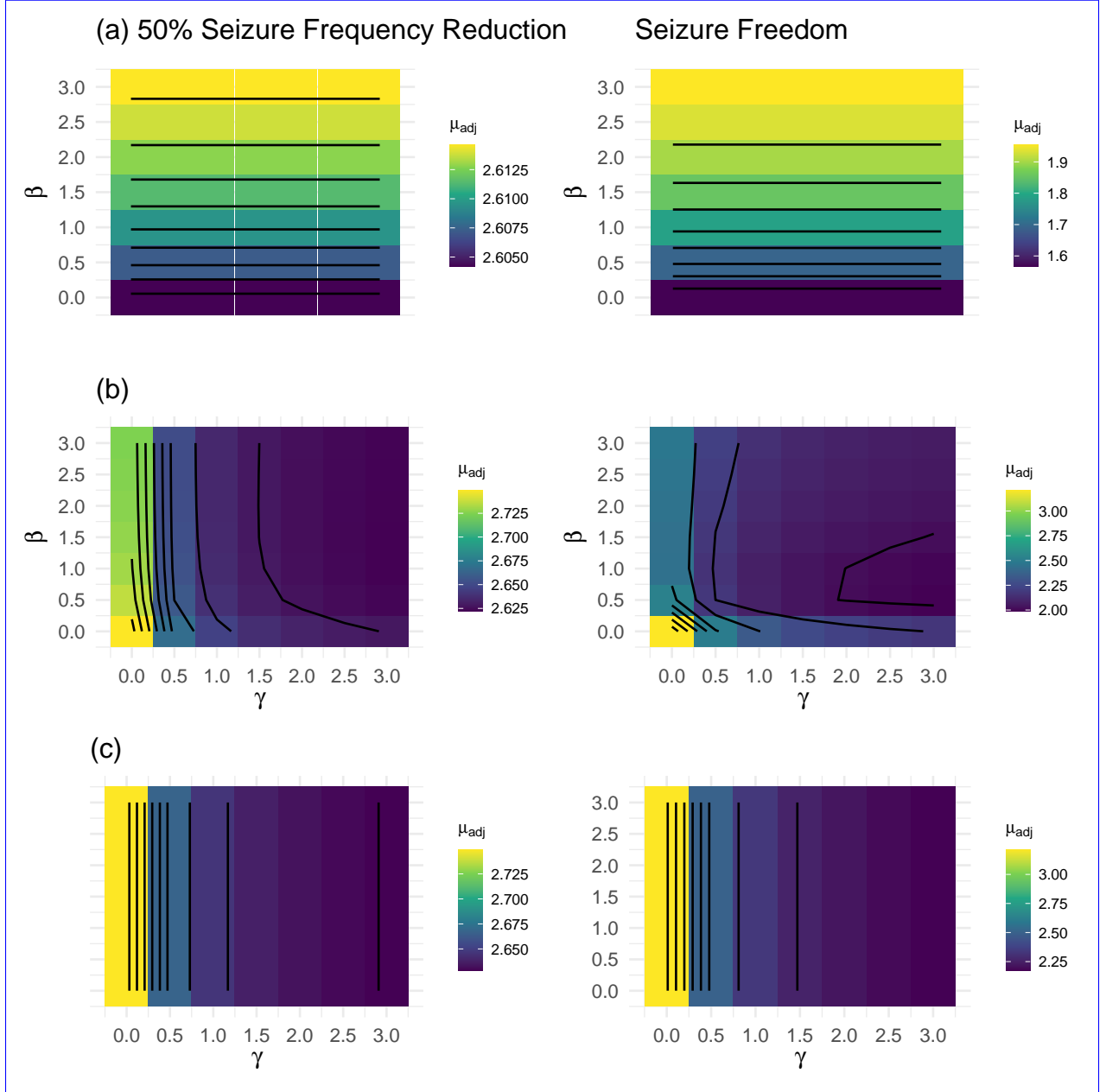
adjustments ~~adjustments~~ ORB-adjustments is of paramount importance. To address this, we focused on and conducted simulations using the random effects model, in contrast to Copas *et al.* [12], which concentrated on the fixed effects model. Maximum likelihood estimation (MLE) was used for estimating heterogeneity variance due to its connection to ORB-adjustment, i.e., the ~~ORB adjustment~~ ORB-adjustment itself is defined via a likelihood function contribution. More

sophisticated methods in the likelihood framework, such as restricted maximum likelihood (REML), could be considered [36, 47, 11, 40]. An exploratory REML approach was proposed in previous work [40], but a more robust derivation could be investigated. Obtaining accurate estimates of τ^2 is crucial, and while challenging to intertwine it with ORB-adjustment outside the likelihood framework of joint estimation with μ, novel methods could be investi-

gated [47, 36, 11]. Another potential area for future research is the effect of ORB on prediction intervals [1] and how ~~ORB adjustments~~ ORB-adjustments impact them, as mentioned in previous work [40].

One additional avenue for future research on ORB is the multiple imputation (MI) of missing study outcomes, which has also been used in publication bias (PB). In the context of PB, Carpenter *et al.* [5] fit a model to the observed study data and impute missing studies using a missing at random (MAR) assumption. They then use a re-weighting scheme that follows similar selection function assumptions to those made in this work. In the context of ORB, the possibility to impute not at random by sampling from a different distribution that directly models the selection process could be considered. Additionally, in the PB context, Carpenter *et al.* [5] had to impute the study sizes, which are known in the ORB setting proposed here. Hence, this could be a compelling avenue for future research. In this sense, MI would benefit from modelling outcomes that are correlated to borrow strength in case of missing outcomes [5], as done in previous work on ORB [2, 28]. This could be of particular interest in cases such as our motivating example from Copas *et al.* [12], Bresnah *et al.* [4], where numerous outcomes are considered in the meta-analysis. The challenge in this approach, as previously noted, lies in the estimation of the correlation coefficients [40, 2, 28].

Our focus was on ORB-adjustment of beneficial outcomes, but the methodology proposed in this work can be easily extended to harmful outcomes by adjusting the selection functions for a different missing data mechanism accordingly. This could mean changing the assumed selection mechanism for unreported outcomes to, for example, assuming that a positive value of the treatment effect for a harmful outcome, or a significant one, results in a lower probability of reporting [12, 10, 39]. Future implementations of this ORB-adjustment framework could hence investigate which missing data assumptions are reasonable to make for harmful outcomes and, e.g., conduct a simulation study similar to the one done here for various, flexible, selection functions.

For future research on ORB, we encourage the refinement and further exploration of simulation studies and strongly recommend using a pre-defined protocol for transparency and reproducibility. The simulation study conducted in this work utilized a limited range of data-generating mechanism (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, as well as comparisons with new poten-

tial approaches, such as the MI ones above-described.

Overall, this study highlights the significant impact of ORB on treatment effect estimation, as well as heterogeneity variance, and demonstrates the efficacy of a flexible ORB-adjustment framework based on selection models. This framework allows the inclusion of contributions from unreported study outcomes and the specification of the desired assumed missing data mechanism via the selection function. The methodology shows promise in mitigating ORB across various settings, with potential for further refinement and broader application.

5   Code Availability

The `R` code for the simulation study presented in this research can be accessed in the ORBproject GitHub repository. This repository includes all the scripts used in the simulation, featuring the key function `reORBadj`, which implements the ORB-adjustment following the selection model framework described in this study. This function is thus applicable to any meta-analysis dataset with unreported study outcomes that might indicate ORB. By making the simulation study code accessible in the ORBproject GitHub repository and providing a pre-defined simulation study protocol in the OSF project repository, this research underscores the importance of reproducibility and transparency in scientific investigations.

References

[1] Amer, F. A. and Lin, L. (2021). Empirical assessment of prediction intervals in cochrane meta-analyses. *Eur J Clin Invest*, **51**, 13524.

[2] 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. https://doi.org/10.48550/ARXIV.2110.08849.

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

[4] Bresnah, R., Hounsome, J., Jette, N., Hutton, J., and Marson, A. (2019). Topiramate add-on therapy for drug-resistant focal epilepsy. *Cochrane Database of Systematic reviews*, **10**, CD001417.

[5] Carpenter, J., Rücker, G., and Schwarzer, G. (2011). Assessing the sensitivity of meta-analysis

to selection bias: A multiple imputation approach. *Biometrics*, **67**, 1066–1072.

[6] Carter, E. C., Schönbrodt, F. D., Gervais, W. M., and Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, **2**, 115–144.

[7] 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.

[8] 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.

[9] Cooper, H., Hedges, L., and Valentine, J. (2009). *The Handbook of Research Synthesis and Meta-Analysis*. Russell Sage Foundation.

[10] 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.

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

[12] 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.

[13] 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.

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

[15] Dwan, K., Altman, D. G., Arnaiz, J. A., Bloom, J., Chan, A.-W., Cronin, E., Decullier, E., Easterbrook, P. J., Elm, E. V., Gamble, C., Ghersi, D., Ioannidis, J. P. A., Simes, J., and Williamson, P. R. (2008). Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS ONE*, **3**, e3081.

[16] Egger, M., Smith, G. D., and Higgins, J. (2022). *Systematic Reviews in Health Research: Meta-Analysis in Context*. Wiley-Blackwell, Chichester, England, 3 edition.

[17] Fernandez-Castilla, B., Declercq, L., Jamshidi, L., Beretvas, N., Onghena, P., and Van den Noortgate, W. (2021). Detecting selection bias in meta-analyses with multiple outcomes: A simulation study. *The Journal of Experimental Education*, **89**, 125–144.

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

[19] 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.

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

[21] Higgins, J., Altman, D., Gøtzsche, P., Jüni, P., Moher, D., Oxman, A., Savović, J., Schulz, K., Weeks, L., and Sterne, J. (2011). The cochrane collaboration's tool for assessing risk of bias in randomised trials. *BMJ*, **343**, .

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

[23] 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.

[24] Iyengar, S. and Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science*, **3**, 109–135.

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

[26] Kambach, S., Bruelheide, H., Gerstner, K., Gurevitch, J., Beckmann, M., and Seppelt, R. (2020). Consequences of multiple imputation of missing standard deviations and sample sizes in meta-analysis. *Ecology and Evolution*, **10**, 11699–11712.

[27] 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.

[28] 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.

[29] 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.

[30] Liu, Y., DeSantis, S. M., and Chen, Y. (2017). Bayesian mixed treatment comparisons meta-analysis for correlated outcomes subject to reporting bias. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **67**, .

[31] Lunny, C., Veroniki, A.-A., Higgins, J. P. T., Dias, S., Hutton, B., Wright, J. M., White, I. R., Whiting, P., and Tricco, A. C. (2024). Methodological review of nma bias concepts provides groundwork for the development of a list of concepts for potential inclusion in a new risk of bias tool for network meta-analysis (rob nma tool). *Systematic Reviews*, **13**, 25.

[32] 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.

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

[34] 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.

[35] Morris, T., White, I., and Crowther, M. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, **38**, 2074–2102.

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

[37] 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.

[38] Riley, R., Thompson, J., and Abrams, K. (2007). An alternative model for bivariate random-effects meta-analysis when the within-study correlations are unknown. *Biostatistics*, **9**, 172–186.

[39] 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. *BMJ*, **349**, g6501.

[40] Saracini, A. G. (2023). Addressing outcome reporting bias in meta-analysis: A comprehensive review and future directions. Master Thesis. `https://doi.org/10.3929/ETHZ-B-000666626`.

[41] 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.

[42] Silva, S., Singh, S., Kashif, S., Ogilvie, R., Pinto, R. Z., and Hayden, J. A. (2024). Many randomized trials in a large systematic review were not registered and had evidence of selective outcome reporting: a metaepidemiological study. *Journal of Clinical Epidemiology*, **176**, 111568.

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

[44] 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.

[45] Thomas, E. T. and Heneghan, C. (2022). Catalogue of bias: selective outcome reporting bias. *BMJ Evidence-Based Medicine*, **27**, 370–372.

[46] van Aert, R. and Wicherts, J. (2024). Correcting for outcome reporting bias in a meta-analysis: A meta-regression approach. *Behavior Research Methods*, **56**, 1994–2012.

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

[48] Weir, C. J., Butcher, I., Assi, V., Lewis, S. C., Murray, G. D., Langhorne, P., and Brady, M. C. (2018). Dealing with missing standard deviation and mean values in meta-analysis of continuous outcomes: a systematic review. *BMC Med Res Methodology*, **18**, 25.

Figure 4: Bias in the estimation of the treatment effect μ for ORB simulated according to DGM function (14) with $\gamma = 1.5$, using different estimation methods, i.e., naive or ORB-adjusted according to the various selection functions indicated in the legend. The bias is shown for varying meta-analysis study sizes, heterogeneity levels, and increasing true treatment effect on the x-axis of each plot shown.
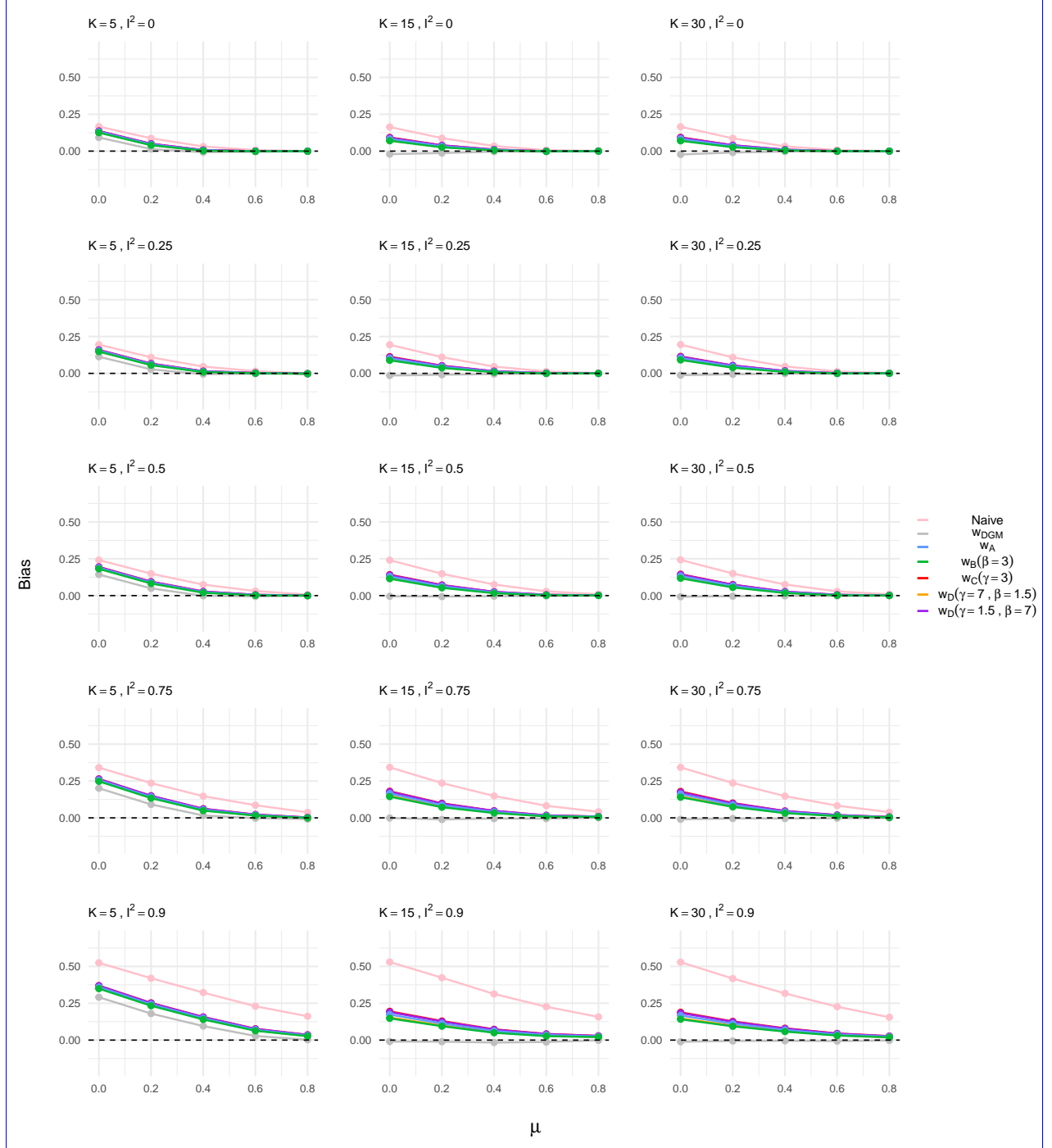
Figure 5: Bias in the estimation of the treatment effect μ for ORB simulated according to DGM function (14) with $\gamma = 0.5$, using different estimation methods, i.e., naive or ORB-adjusted according to the various selection functions indicated in the legend. The bias is shown for varying meta-analysis study sizes, heterogeneity levels, and increasing true treatment effect on the x-axis of each plot shown.
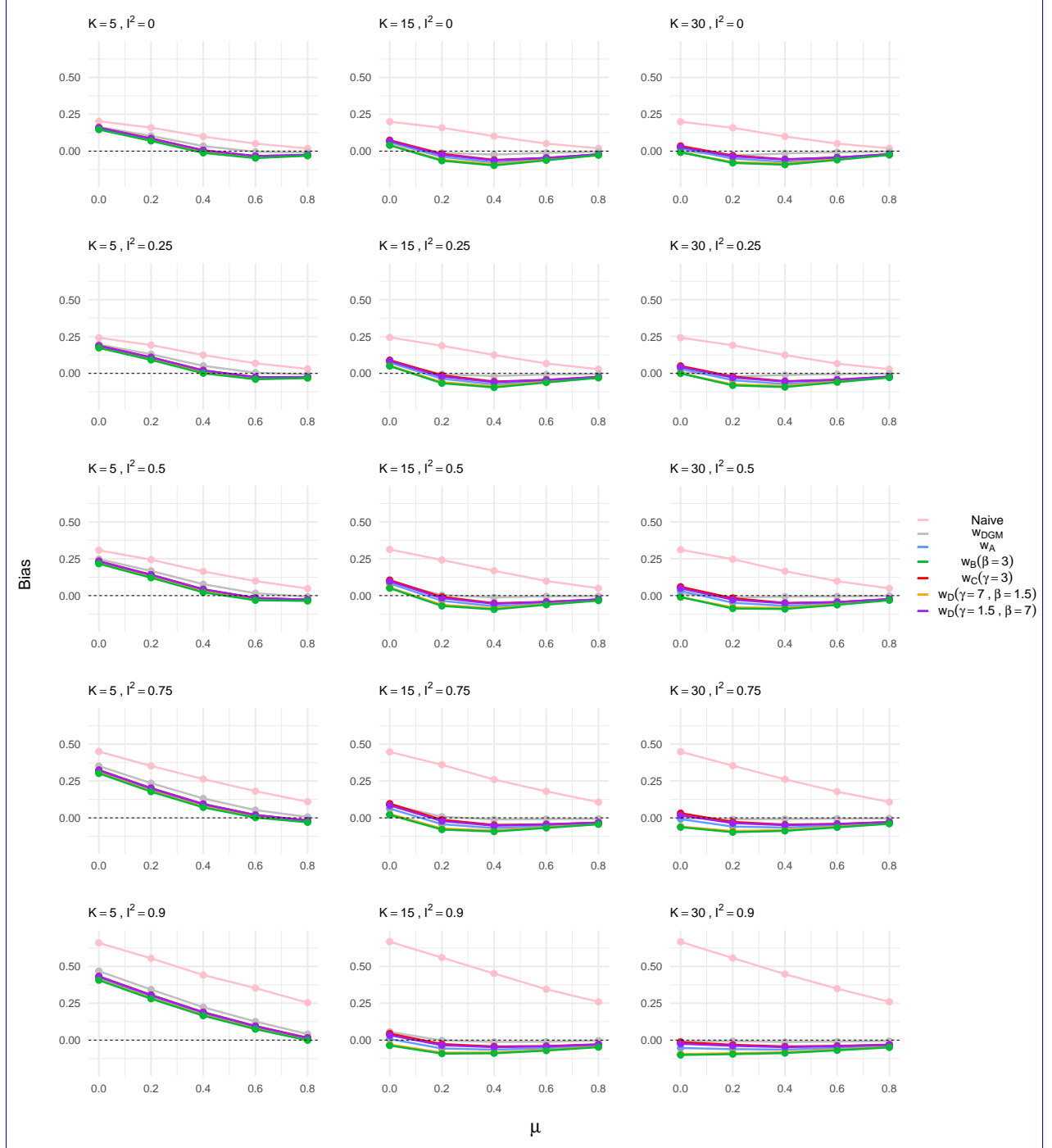
Figure 6: Coverage in the estimation of the treatment effect μ for ORB simulated according to DGM function (14) with $\gamma = 1.5$, using different estimation methods, i.e., naive or ORB-adjusted according to the various selection functions indicated in the legend. The coverage is shown for varying meta-analysis study sizes, heterogeneity levels, and increasing true treatment effect on the x-axis of each plot shown.
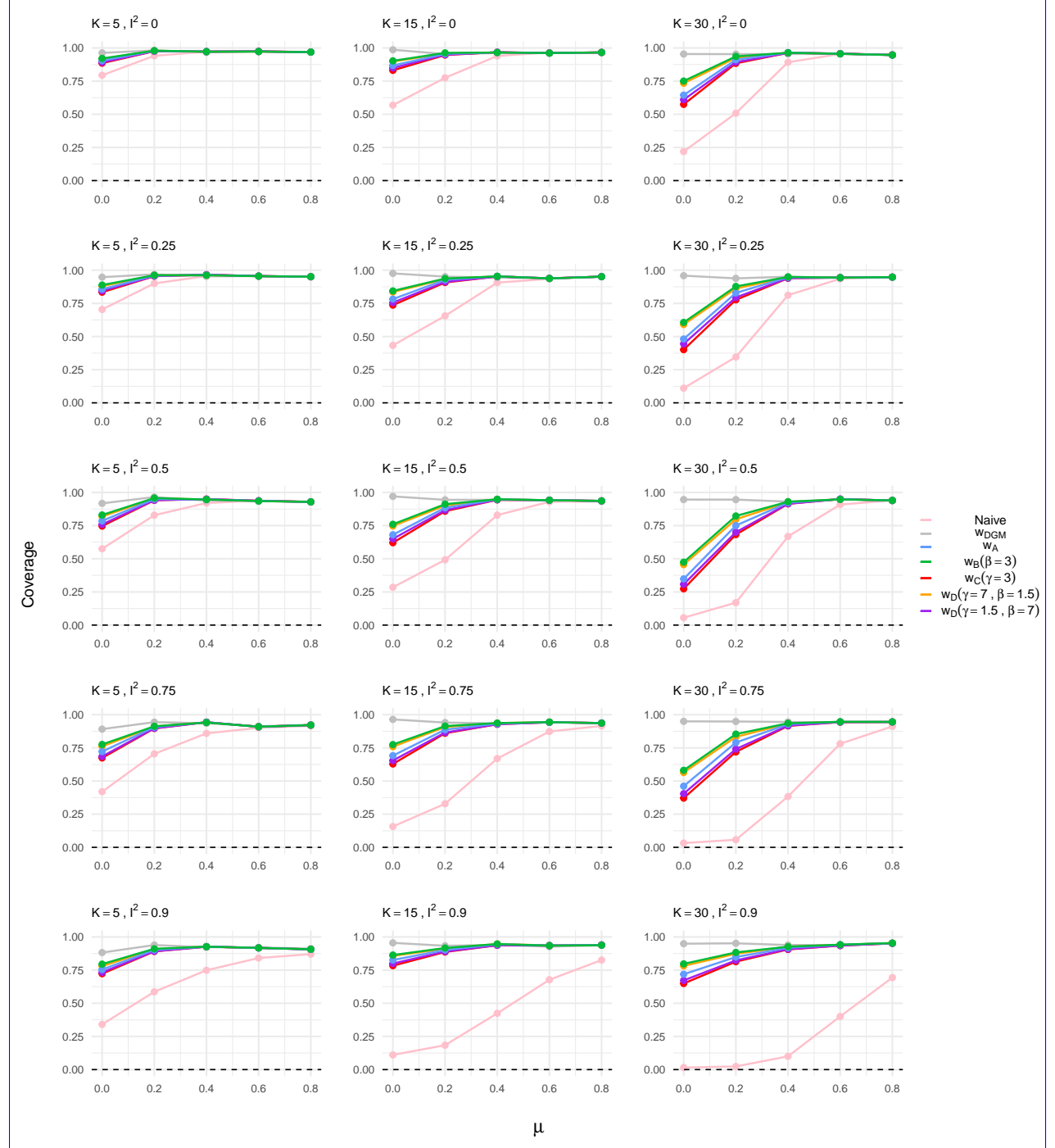
Figure 7: Bias in the estimation of the heterogeneity variance τ^2 for ORB simulated according to DGM function (14) with $\gamma = 1.5$, using different estimation methods, i.e., naive or ORB-adjusted according to the various selection functions indicated in the legend. The bias is shown for varying meta-analysis study sizes, true treatment effect values, and increasing heterogeneity on the x-axis of each plot shown.