Simulation Study Protocol

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

Alessandra Saracini and Leonhard Held

Contents

1 Introduction

Outcome reporting bias (ORB) occurs when outcomes within a meta-analysis are selectively reported based on the significance or direction of results. Unlike publication bias (PB), a much better known and studied phenomenon, ORB occurs when studies are published in the literature, but they lack (sufficient) information for a certain outcome for it to be included in the meta-analysis. Standard meta-analysis methods, which only include studies that fully report the outcome of interest, implicitly assuming that unreported outcomes are missing at random, can result in biased estimation. Despite its critical implications, ORB remains an under-recognized problem area, with limited statistical adjustment methods available.

The most well-established ORB-adjustment method is that of Copas et al. [4]. This methodology requires the unreported study outcomes to be classified according to the ORBIT methodology into risk of bias categories. This classification requires expert opinion and has been argued to be potentially limiting in certain settings [13, 23]. Once the classification is done, and thus assumed to be correct, Copas et al. [4] include a contribution of unreported study outcomes classified as high risk of bias (HR), by making the assumption that unreported study outcomes at HR are not significant.

To overcome the need for an ORBIT classification, relax the missing data assumptions made by Copas et al. [4], and hence include a contribution from all unreported outcomes, not just the ones at HR, we investigate ORB-adjustment methods via a selection model perspective.

Selection models are usually used in PB and consist in associating a weight function, representing the probability of publishing, to the published studies [12, 8, 5, 1, 22, 9]. Let y_i be the observed treatment effect estimate for study i in the meta-analysis, with distribution $f_i(y)$, usually assumed to be normal. By using the following relation:

$$f_i\left(y_i | i^{th} \text{ study published}\right) = \frac{f_i(y_i;\theta) \cdot w_i(y_i)}{\int_{-\infty}^{+\infty} f_i(y;\theta) \cdot w_i(y) dy} \tag{1}$$

the PB-adjusted log-likelihood $\ell_{\text{PB-adjusted}}(\theta)$ is derived [9, 12, 8, 5]. The log-likelihood to be maximized for the parameter θ in turn has an additional contribution for the published studies, $i \in$ Published, wherein $f_i(y)$ is weighted by $w_i(y)$, the selection function representing the probability of publishing.

$$\ell_{\text{PB-adjusted}}(\theta) \propto \sum_{i \in \text{Published}} \log f_i(y_i;\theta) - \sum_{i \in \text{Published}} \log \left[\int_{-\infty}^{\infty} f_i(y;\theta) \cdot w_i(y) dy\right] \tag{2}$$

In the context of ORB, we similarly can associate a weight function for the probability of (un)reporting to the missing study outcomes. In the ORB-adjusted log-likelihood, we thus have an additional contribution from the unreported study outcomes, $i \in$ Unreported, which varies depending on the shape of the probability of (un)reporting, $1 - w_i(y)$, i.e., a mathematical representation of the missing data mechanism assumed. Details of the result obtained in (3) can be found in Chapter 2 of Saracini and Held [19].

$$\ell_{\text{ORB-adjusted}}(\theta) \propto \sum_{i \in \text{Reported}} \log f_i(y_i;\theta) + \sum_{i \in \text{Unreported}} \log \left[\int_{-\infty}^{\infty} f_i(y;\theta) \cdot (1 - w_i(y))\, dy\right] \tag{3}$$

In light of the (3) framework, one can note that the Copas et al. [4] method is a special case of the general ORB-adjustment proposed. Specifically, we obtain the Copas et al. [4] ORB-adjusted log-likelihood when considering a subset of unreported outcomes, i.e., $i \in$ HR instead of all $i \in$ Unreported and a strict missing data assumption represented through $w_i(y)$. As previously noted, we propose to adjust for ORB by including the contribution from all unreported study outcomes and making more flexible missing data assumptions, represented via the probability of reporting $w_i(y)$.

In order to assess the performance of the ORB-adjustment methodology proposed, as well as to gain a deeper insight into the impact of ORB across diverse meta-analysis settings, in particular in the presence of heterogeneity between studies, a simulation study is undertaken. This document outlines the planned steps for the ORB simulation study, including its objectives, data generation mechanism, analytical methods, and performance evaluation criteria [20, 16]. This is essential in order to ensure transparency and reproducibility of study design and execution.

2 General Information

Title. Investigating the impact of outcome reporting bias (ORB) and effectiveness of ORB-adjustment methods via a selection model perspective in meta-analysis with varying levels of heterogeneity.

Contributors. Alessandra Gaia Saracini and Leonhard Held

Description. The simulation study carried out simulates ORB in a meta-analysis in the presence of heterogeneity, i.e., under the random-effects model, and applies the ORB-adjustment method proposed. It is of interest to gain an understanding of the impact of ORB and of the effectiveness of its correction techniques, on the estimation and testing of the treatment effect and of the heterogeneity parameters under different meta-analysis settings.

3 Aims

The simulation study aims to comprehensively evaluate the impact of outcome reporting bias (ORB) on treatment effect estimation and heterogeneity parameter estimation within meta-analysis when using different estimation methods. The different estimation methods correspond to different assumed selection functions, i.e., different functions describing the probability of an outcome being reported. These assessments will be conducted across a spectrum of scenarios, varying characteristics such as the number of studies and the level of heterogeneity.

Naive estimation, involving maximum likelihood estimation with contributions solely from studies which report the outcome of interest, serves as a baseline for comparison. Naive estimation implicitly assumes that unreported study treatment effects in the meta-analysis are missing at random, i.e., that there is no selection mechanism dependent on the significance. It is of interest to assess the impact of ORB on naive estimation, under the different simulated meta-anlaysis settings.

A key objective is to then evaluate the effectiveness of the ORB-adjustment method (3) in reducing the possible bias stemming from naive estimation. This methodology incorporates a contribution from unreported/missing study outcomes into the maximum likelihood function. The shape of the ORB-adjusted likelihood differs depending on the different selection function assumed for the probability of reporting, i.e., $w_i(y)$ in (3). Taking inspiration from selection functions typically used in the PB literature, along with qualitative assessment of reasons behind unreporting of outcomes in published studies, we test different selection functions.

Comparative analysis of the ORB-adjusted estimation methods, obtained from the different selection functions, is of paramount importance, striving to align their accuracy as closely as possible with the true value of the parameters of interest, known via the simulation study, and reducing the bias stemming from naive estimation.

4 Data-generating mechanism (DGM)

The DGM comprises two components. The first involves simulating a meta-analysis study, while the second involves simulating Outcome Reporting Bias (ORB) by strategically excluding certain studies from the meta-analysis, based on the direction and significance of the treatment effects.

4.1 Meta-analysis DGM

Each simulated meta-analysis dataset comprises K studies. Each study, indexed by i, includes a treatment and control arm. We assume equal sample sizes for both arms [15], for all studies, i.e., $n_i = n = 50$. This choice was motivated by a typical study size in the clinical trials include in meta-analysis [11, 15, 7] and the intention to reduce, in this simulation study, confounding stemming from differences in study trials. Every study reports an estimated treatment effect, denoted by y_i, along with its standard error σ_i. Furthermore, it is assumed that each study reports the size of the control and treatment arms n_i. We assume a positive direction of treatment, i.e., a positive value of y_i indicates a beneficial effect of the intervention relative to placebo. To obtain simulated values of the treatment effects and standard errors, we follow the simulation process outlined by IntHout et al. [11]. Of note, the values are generated independently for each study i, assuming no correlation between studies.

In order to generate the treatment effect values y_i, we first simulate the true study-specific treatment effects from a normal distribution centered around a global treatment effect μ, with between-study heterogeneity variance τ^2:

$$\theta_i \sim \mathcal{N}(\mu, \tau^2) \tag{4}$$

These true study-specific treatment effects θ_i are then utilized to derive observed treatment effects y_i for each study i. The observed treatment effects y_i are sampled from a normal distribution centered around θ_i, with variance σ^2:

$$y_i \sim \mathcal{N}(\theta_i, \sigma^2) \tag{5}$$

Given that in our setting the sample sizes in the control and treatment arms n_i are fixed and equal for each study $i \in 1, ..., K$, i.e., $n_i = n$, the within-study variance used in (5) is $\sigma^2 = 2\epsilon^2/n$, where ϵ^2 is the measurement error variance for the outcome. As per IntHout et al. [11], we set $\epsilon^2 = 1$ in the simulations, and thus $\sigma^2 = 2/n$.

After having obtained the treatment effect estimates y_i, the individual within-study variances σ_i^2 are simulated from the following scaled χ^2 distribution with $2n_i - 2$ degrees of freedom [11], such that $\text{E}(\sigma_i^2) = 2/n_i = 2/n = \sigma^2$:

$$\sigma_i^2 \sim \frac{\chi^2_{2n_i-2}}{(n_i - 1)n_i} \tag{6}$$

In the simulation study, we are interested in assessing the impact of ORB and the effectiveness of its adjustment techniques under different meta-analysis settings. To this end, we use the following parameter combinations:

- Number of studies in the meta-analysis, $K \in \{5, 15, 30\}$, following similar choices by Moreno et al. [15], IntHout et al. [11] for small, medium, and large meta-analysis studies of clinical trial.

- True global treatment effect $\mu \in \{0, ..., 0.8\}$ with 0.1 increments, representing, e.g., standardized mean differences or log odds ratios, similar to choices in Moreno et al. [15], Fernández-Castilla et al. [7] who note that, e.g., effect sizes of $0, 0.2, 0.5, 0.8$ could correspond to null, low, medium and high effects [7].
- Heterogeneity quantified by Higgins' $\mathrm{I}^2 \in \{0\%, 25\%, 50\%, 75\%, 90\%\}$ as done in IntHout et al. [11].

Higgins' I^2 is a quantity which is often used in meta-analysis to quantify heterogeneity [10]. Compared to the τ^2 heterogeneity variance, I^2 is deemed more interpretable, as it estimates the proportion of the variance in study estimates that is due to heterogeneity. Furthermore, it is preferred as it ranges from 0 to 1, which makes its magnitude easier to contextualize. However, in practice we need to use τ^2, as this is our core parameter for estimation and analysis. We thus use the following equivalence, which defines τ^2 in terms of I^2 and of the expected/typical within-study variance $\mathrm{E}(\sigma_i^2) = \sigma^2$, as follows:

$$\tau^2 = \frac{\mathrm{I}^2}{1 - \mathrm{I}^2} \cdot \sigma^2 \tag{7}$$

Given that in our setting $n_i = n = 50$, we use $\sigma^2 = 2/n$ and (7) and establish that the values of Higgins' I^2 which we wish to vary in the simulation correspond to τ^2 values of $\tau^2 \in \{0, 0.013, 0.04, 0.12, 0.36\}$.

4.2 ORB DGM

After simulating K studies, each providing a treatment effect y_i obtained from (5), standard error σ_i obtained from (6), and sample sizes n_i for the control and treatment arms, the next step is to simulate Outcome Reporting Bias (ORB). This involves selectively removing certain treatment effects and standard errors based on their significance and/or direction, thus introducing missing values not at random.

Since we are implementing an ORB-adjustment via a selection model framework inspired by the publication bias (PB) literature, we simulate ORB similarly to how PB is often generated in its simulations studies. Namely, we remove study outcomes with a probability which is a decreasing function of the p-value, as done in, e.g., Dear and Begg [5], Begg and Mazumdar [1], Macaskill et al. [14], Preston et al. [18]

$$P(\text{Reporting } \{y_i, \sigma_i\}) = e^{-4 \cdot p_i^{\gamma}} \tag{8}$$

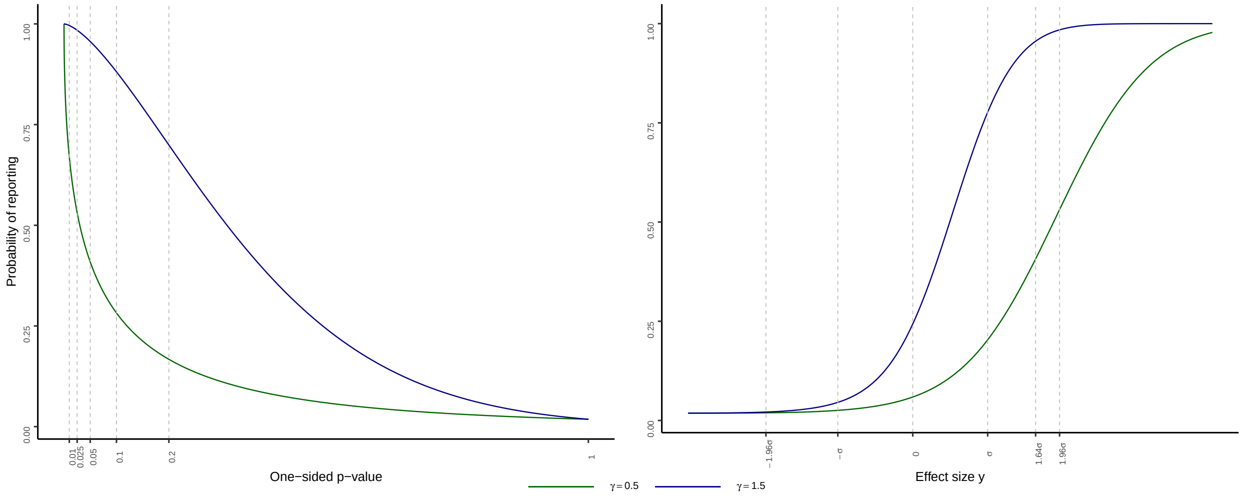
Here, we use the one-sided p-value $p_i = \Phi(-y_i/\sigma_i)$. We use a one-sided p-value to simulate ORB as we assume that a significant value but in the negative direction, i.e., indicating a negative effect of the treatment relative to control, is unlikely to be reported [9, 18, 22].

Of note, in the PB simulation study literature, the number of selected/published studies included in each meta-analysis is typically fixed; this is obtained by making the decision to select a study for inclusion based on a simulated biased coin toss with probability weight (8) and repeating the procedure until the required number of studies is selected [5, 1, 14]. In the case of the ORB simulation, the total number studies (i.e., which report or do not report the outcome of interest) is fixed, and the selection is done using (8) as a probability for selecting the reported outcomes. Hence, in each simulation, the number of reported outcomes may vary. This approach was used by Copas et al. [4] in the context of ORB simulation, as well as in Fernández-Castilla et al. [7] for both ORB and PB simulations. In our setting, as we require at least two study outcomes to be reported for anive estimation of the treatment effect and heterogeneity variance; hence, we repeat the simulation process if we obtain a dataset with less than 2 reported study outcomes. For each parameter combination, we will record the average number of reported study outcomes as well as the number of simulated datasets/trials needed to obtain the desired behavior.

The γ value in (8) indicates the strength of bias. In PB, it has been argued that a value of $\gamma = 1.5$ can be representative of strong PB. In this case, e.g., the probability of publishing a study with significant one-sided

p-values $0.01,\ 0.05$ is $0.996,\ 0.962$ respectively and drop to $0.881,\ 0.699$ for non-significant values of $0.1,\ 0.2$ respectively. However, when considering ORB, given that studies are already published, the threshold for reporting could be less stringent. Factors such as clinical relevance or prioritizing more impactful findings [21, 2] may lead to unreported outcomes with smaller p-values, compared to what one may expect in the context of PB. With this rationale, in the simulation study of ORB we use $\gamma = 1.5$, for consistency with PB; we additionally use $\gamma = 0.5$. With this parameter, the probabilities of reporting are high only for very small p-values, e.g., the probability of reporting is 0.961 for a p-value of 0.0001, and drops substantially, e.g., the probability of reporting is 0.670 for a significant p-value of 0.01 and decreases to 0.282 and 0.167 for non-significant p-values of 0.1 and 0.2. The ORB DGM with $\gamma = 0.5$ thus possibly results in unreported study outcomes with original p-values below the 0.05 threshold, as opposed to the setting of $\gamma = 1.5$, where the probability of reporting is almost one for significant p-values. It is of interest to investigate the impact of ORB in these different settings, as well as the extent to which different ORB-adjustment methods, with varying missing data assumptions, are successful in correcting for the potential bias in the parameter estimation.

Figure 1: Function (8) with $\gamma = 0.5$ and $\gamma = 1.5$ used to simulate ORB: study outcomes are removed from a simulated meta-analysis according to the probability of reporting [5], based on one-sided p-value (left) and corresponding treatment effect size (right).



5 Estimands and Targets

Our primary estimand is the treatment effect size, derived from evidence pooled across multiple studies in the meta-analysis. This estimation is conducted via maximum likelihood (ML), where the likelihood function is maximized. We perform both naive estimation, utilizing treatment effects solely from reported studies, and estimation using ORB-adjustment techniques. Additionally, our secondary target is the estimation of the heterogeneity variance. This is estimated concurrently with the treatment effect in both the naive and ORB-adjusted estimations. The statistical target of this simulation study is focused on estimation. Therefore, we aim to compare different estimation methods in terms of bias, mean squared error (MSE), and other relevant metrics.

6 Methods

For a given parameter of interest θ (i.e., the treatment effect, of primary interest, and the heterogeneity variance, of secondary interest), we use maximum likelihood estimation (MLE) to obtain its estimate. We thus maximize the log-likelihood function $\ell(\theta)$ with respect to θ:

$$\hat{\theta}_{ML} = \arg\max_{\mu} \ell(\theta) \tag{9}$$

Additionally, we compute the 95% confidence interval (CI) by considering the likelihood ratio (LR) CI. Specifically, we use the profile likelihood (PL) CI. The range of θ values within the PL CI are

$$\left\{\theta \mid \ell_p(\theta) - \ell_p(\hat{\theta}) + \frac{1}{2}\chi^2_{\text{df}=1,0.95} \leq 0\right\} \tag{10}$$

The ML estimate and PL CI for θ are obtained using different log-likelihoods, depending on the information and/or missing data mechanism assumed, leading to a naive, full data, and ORB-adjusted estimation methods. We further differentiate various ORB-adjusted estimates based on the selection function assumed for the probability of reporting in (3). Of note, ML estimation (accompanied by and PL CI) in random effects meta-analysis is commonly used for the treatment effect estimate μ. While this estimation method is also found for τ^2 [24], additional estimation techniques are present in the literature for the heterogeneity variance estimation [24, 3, 17]. In our case, given that μ is the primary interest of the investigation, and that the ORB-adjustment method is intrinsically based on the likelihood function, we use ML-based methods also for τ^2, as an exploratory investigation of the effect of ORB on τ^2 estimation.

6.1 Naive estimate

The naive log-likelihood includes the contribution only from reported study outcomes and disregards the unreported ones. The implicit assumption is that unreported study outcome are missing at random and are not associated with a selection mechanism [4, 22].

$$\ell_{\text{Naive}}(\theta) = \sum_{i \in \text{Reported}} \log f_i(y_i; \theta) \tag{11}$$

6.2 Complete data estimate

The complete data log-likelihood uses all studies in the meta-analysis before ORB is simulated, and is a proxy for the true treatment effect if there were no bias.

$$\ell_{\text{Complete}}(\theta) = \sum_{i}^{K} \log f_i(y_i; \theta) \tag{12}$$

6.3 ORB-adjusted estimate

Below we recall the ORB-adjusted log-likelihood of (3) as a generic form of ORB-adjustment, which includes, compared to the naive estimate, an additional term with contributions from unreported study outcomes. The term makes use of the selection function for the probability of reporting, assumed to be associated with the missing outcomes. In the simulation study, we use a variety of selection functions $w_i(y_i)$ for the probability of reporting, which, in the next subsections, we define in terms of the one sided p-value, $w_i(p_i)$ where $p_i = \Phi\left(-\frac{y_i}{\sigma_i}\right)$ is simply a transformation of y_i and σ_i.

$$\ell_{\text{ORB-adjusted}}(\theta) \propto \sum_{i \in \text{Reported}} \log f_i(y_i;\theta) + \sum_{i \in \text{Unreported}} \log\left[\int_{-\infty}^{\infty} f_i(y;\theta)\cdot(1-w_i(y))\,dy\right] \tag{13}$$

Firstly, in the ORB-adjustment, we use the two selection functions utilized for the ORB simulation, i.e., function (8) with $\gamma = 1.5$ and $\gamma = 0.5$, so as to have the correct function specification between ORB simulation and ORB adjustment. We then further test other selection functions, taking inspiration from those found in PB literature, as well as altering them to test scenarios which one may expect to occur in ORB.

6.3.1 Piece-wise constant-constant

The simplest selection function used is (14), shown in Figure 2. It is a piece-wise constant function, with probability of reporting 0 for non-significant studies and 1 for significant ones. The threshold for significance is a p-value of $p = 0.05$. While this selection function can be found in the PB literature [22, 12, 8], it is also the one implicitly used in the Copas et al. [4] adjustment. Of note, in Copas et al. [4] it is used only in the contribution of unreported study outcomes classified as high risk of bias (HR) by the ORBIT methodology. Here, however, we apply it all unreported outcomes, as we are not using the ORBIT classification system beforehand.

$$w(p) = \begin{cases} 1 & \text{if } p \leq 0.05 \\ 0 & \text{if } p > 0.05 \end{cases} \tag{14}$$

6.3.2 Piece-wise constant-continuous

One selection function also found in the PB literature [22, 8], which relaxes the somewhat strict assumptions of the previous (14) is the following, shown in Figure 3:

$$w(p) = \begin{cases} 1 & \text{if } p \leq 0.05 \\ \frac{p^{-\beta}}{0.05^{-\beta}} & \text{if } p > 0.05 \end{cases} \tag{15}$$

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 simulation study we use $\beta = 3$. With this parameter, the probability of reporting is high for p-values just above the threshold of 0.05, e.g., a probability of 0.751 for a one-sided p-value of 0.055, but rapidly drops to, e.g., 0.125 for a larger p-value of 0.1.

Figure 2: Piece-wise constant-constant selection function (14) used in (13) for ORB-adjustment. The selection function represents the probability of reporting (based on one-sided p-value, left, and corresponding treatment effect size, right) which is assumed to be associated with the unreported/missing study outcomes.

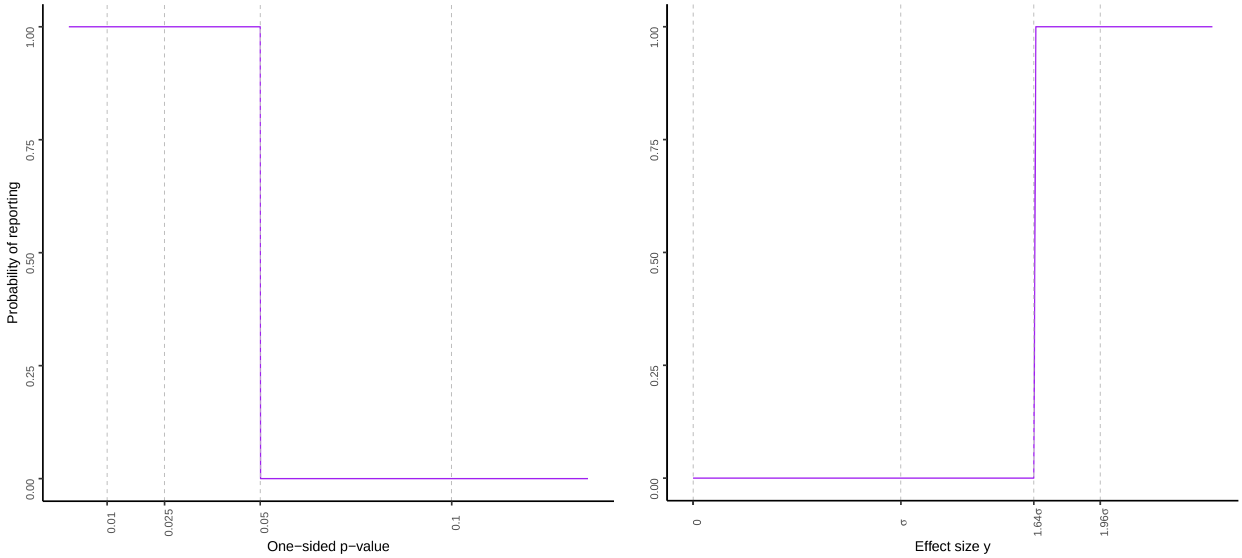
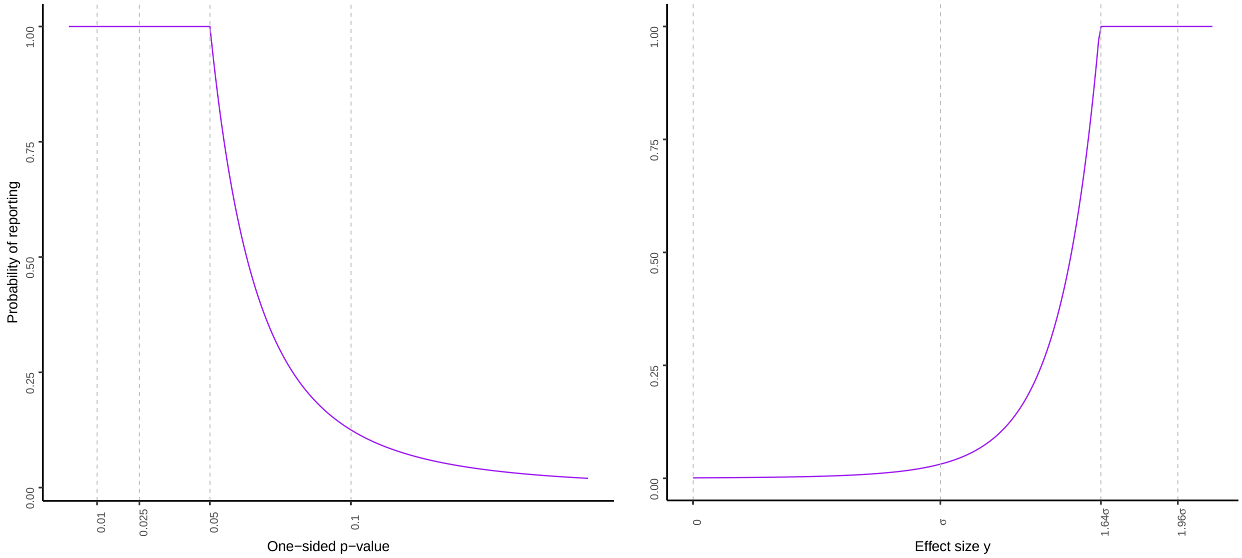


Figure 3: Piece-wise constant-continuous selection function (15) used in (13) for ORB-adjustment. The selection function represents the probability of reporting (based on one-sided p-value, left, and corresponding treatment effect size, right) which is assumed to be associated with the unreported/missing study outcomes.



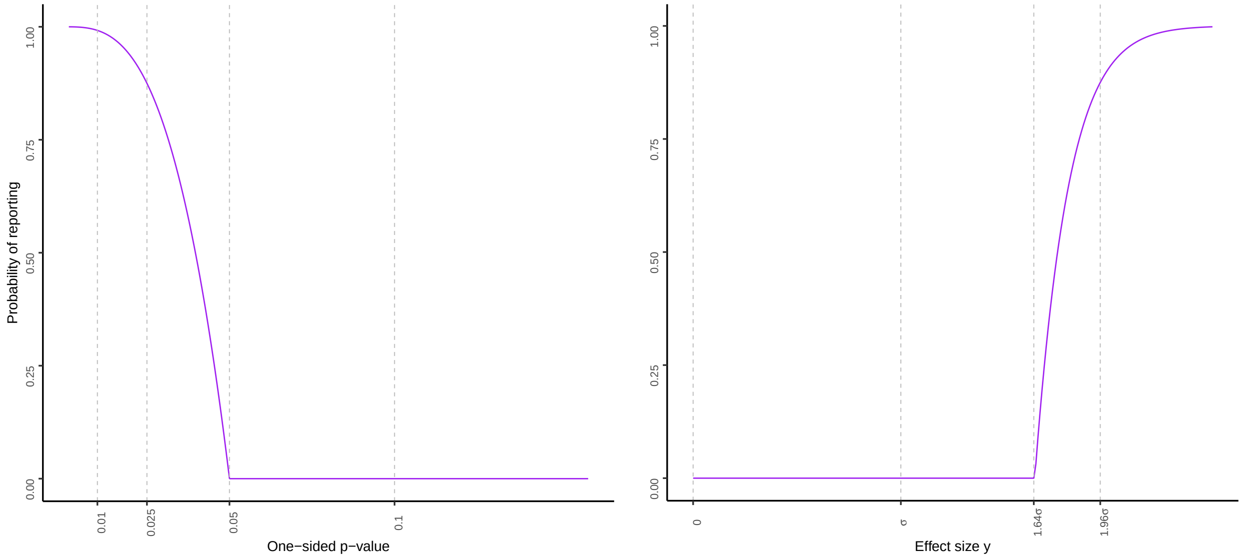
6.3.3 Piece-wise continuous-constant

Taking inspiration from the above, we additionally propose the following selection function, similar to what was done in Saracini and Held [19].

$$w(p) = \begin{cases} 1 - \frac{p^{\gamma}}{0.05^{\gamma}} & \text{if } p \leq 0.05 \\ 0 & \text{if } p > 0.05 \end{cases} \tag{16}$$

The rationale is flipped compared to (15) in that 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. In cases in which ORB is due to, e.g., prioritizing more impactful or clinically relevant findings in a published study [21, 2], even outcomes with smaller p-values could have an associated probability of reporting smaller than 1. At the same time, this setting, resulting in unreported significance outcomes, could be viewed as ORB stemming from less severe forms of bias [19]. In the simulation study, we use $\gamma = 3$. With this parameter value, the probability of reporting is close to 1 only for very small p-values, and substantially drops to low values even for p-value slightly below the 0.05 one-sided threshold, e.g., the probability is $0.992,\ 0.875$ for p-values $= 0.01,\ 0.025$ respectively and 0.271 for p-value $= 0.45$.

Figure 4: Piece-wise continuous-constant selection function (16) used in (13) for ORB-adjustment. The selection function represents the probability of reporting (based on one-sided p-value, left, and corresponding treatment effect size, right) which is assumed to be associated with the unreported/missing study outcomes.

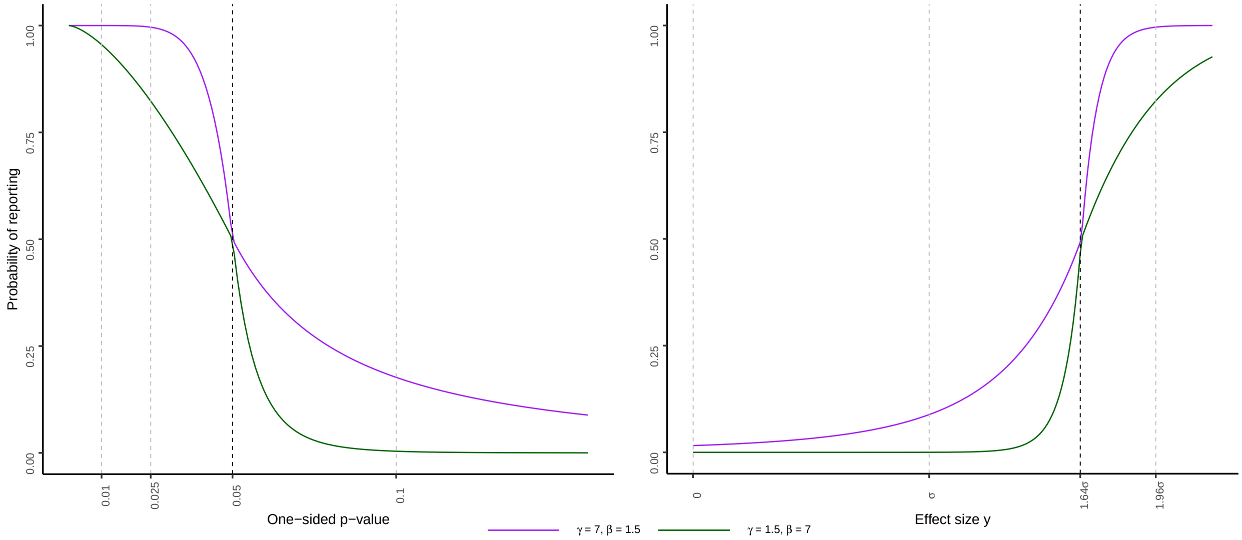


6.3.4 Piece-wise continuous-continuous

Lastly, we use the following selection function (17), which combines (15) and (16) in a flexible way, further allowing to specify the assumed probability of reporting at the threshold for significance, i.e, $w_{0.05}$. Of note, setting $w_{0.05} = 1$ in (17) results in (15), while setting $w_{0.05} = 0$ results in (16). In the simulation, when using (17), we use $w_{0.05} = 0.5$, as the in-between value of the two extreme settings of(15) and (16). Furthermore, in (17) we set $\gamma = 7$, and $\beta = 1.5$. This parameter choice corresponds to a setting in which we assume higher probabilities of reporting for p-values below 0.05, compared to the setting of (16), and lower probabilities of reporting for p-values above 0.05, compared to the setting of (15). For example, the probability of reporting is $0.996,\ 0.760$ for p-values $0.025,\ 0.045$ respectively, and $0.433,\ 0.177$ for p-values $0.55,\ 0.1$. For completeness, we also examine the opposite setting, wherein $\gamma = 1.5$, and $\beta = 7$.

$$w(p) = \begin{cases} 1 - (1 - w_{0.05}) \left(\frac{p^{\gamma}}{0.05^{\gamma}}\right) & \text{if } p \leq 0.05 \\ w_{0.05} \left(\frac{p^{-\beta}}{0.05^{-\beta}}\right) & \text{if } p > 0.05 \end{cases} \tag{17}$$

Figure 5: Piece-wise continuous-continuous selection function (17) used in (13) for ORB-adjustment. The selection function represents the probability of reporting (based on one-sided p-value, left, and corresponding treatment effect size, right) which is assumed to be associated with the unreported/missing study outcomes.



7 Performance measures

The primary focus of the simulation study is estimation. We are thus interested in repeating the simulation process several times and comparing various performance measures between the naive and ORB-adjusted estimates.

We repeat the above-described simulation process and analyses $\mathrm{N}_{\mathrm{sim}}$ times, for each parameter setting (i.e., varying study size K, heterogeneity variance τ^2, and global true treatment effect μ). The number of simulations needed is calculated in the next section, 7.1.

We record the performance measures for the parameter of interest θ, i.e., $\theta = \mu$ or $\theta = \tau^2$. Of note, our primary parameter of interest is the treatment effect μ, while the heterogeneity variance τ^2 is our secondary parameter of interest, for which we will use the same performance measures as those for μ. Table 1 includes the definitions, estimation methods, and Monte Carlo SE (MCSE) for each performance measure. The formulas for the MCSE for each of the performance measures are taken from Morris et al. [16] and will be reported via graphs, adding an error band of $\pm 1.96 \cdot \mathrm{MCSE}$.

7.1 Simulation Size

Given that our primary estimand is treatment effect μ, our key performance measure if bias. We thus establish the number of simulations needed based on that and on the MCSE required, e.g., 0.005 from IntHout et al. [11], Morris et al. [16].

$$\mathrm{N}_{\mathrm{sim}} = \frac{\mathrm{Var}(\hat{\mu})}{(\mathrm{MCSE\ Required})^2} \tag{18}$$

From the above formula, we need to establish, in expectation, the $\mathrm{Var}(\hat{\mu})$. Using the DerSimonian and Laird [6] estimate of the variance in the random effects model [11], we have:

Table 1: Performance measures: deinfition, estimate, and Monte Carlo standard errors (MCSE)

Performance Measure	**Estimate**	**MCSE**
Bias	$\frac{1}{N_{\text{sim}}}\sum_{j=1}^{N_{\text{sim}}}(\hat{\theta}_j - \theta)$	$\sqrt{\frac{1}{N_{\text{sim}}(N_{\text{sim}}-1)}\sum_{j=1}^{N_{\text{sim}}}(\hat{\theta}_j - \bar{\theta})^2}$
Empirical SE	$\sqrt{\frac{1}{N_{\text{sim}}-1}\sum_{j=1}^{N_{\text{sim}}}(\hat{\theta}_j - \bar{\theta})^2}$	$\frac{\widehat{\text{SE}}}{\sqrt{2(N_{\text{sim}}-1)}}$
MSE	$\frac{1}{N_{\text{sim}}}\sum_{j=1}^{N_{\text{sim}}}(\hat{\theta}_j - \theta)^2$	$\sqrt{\frac{\sum_{j=1}^{N_{\text{sim}}}[(\hat{\theta}_j-\theta)^2-\widehat{\text{MSE}}]}{N_{\text{sim}}(N_{\text{sim}}-1)}}$
Coverage	$\frac{1}{N_{\text{sim}}}\sum_{j=1}^{N_{\text{sim}}}(\hat{\theta}_{\text{low},j} \leq \theta \leq \hat{\theta}_{\text{up},j})$	$\sqrt{\frac{\widehat{\text{Cov}}(1-\widehat{\text{Cov}})}{N_{\text{sim}}}}$
Power	$\frac{1}{N_{\text{sim}}}\sum_{j=1}^{N_{\text{sim}}}1(p_i \leq 0.05)$	$\sqrt{\frac{\widehat{\text{Power}}(1-\widehat{\text{Power}})}{N_{\text{sim}}}}$

$$\text{Var}(\hat{\mu}) = \frac{1}{\sum_i^K w_i} \qquad \text{where} \qquad w_i = \frac{1}{\sigma_i^2 + \tau^2} \approx \frac{1}{\text{E}(\sigma_i^2) + \tau^2} \tag{19}$$

Using $\text{E}(\sigma_i^2) = \sigma^2 = 2/n_i$ and noting that $n_i = n$ is constant, this corresponds to

$$\text{Var}(\hat{\mu}) \approx \frac{1}{\sum_i^K \frac{1}{\frac{2}{n_i}+\tau^2}} = \frac{1}{K \cdot \left(\frac{1}{\frac{2}{n}+\tau^2}\right)} = \frac{\frac{2}{n}+\tau^2}{K} \tag{20}$$

For the different parameter settings, i.e., varying study sizes $K \in \{5, 15, 30\}$ and varying heterogeneity levels $\tau^2 \in \{0, 0.013, 0.04, 0.12, 0.36\}$ we expect the following values of the standard errors, i.e., $\sqrt{\text{Var}(\hat{\mu})}$:

Table 2: Expected standard errors of $\hat{\mu}$ for varying K and τ^2

	$K = 5$	$K = 15$	$K = 30$
$\tau^2 = 0$	0.09	0.05	0.04
$\tau^2 = 0.01$	0.10	0.06	0.04
$\tau^2 = 0.04$	0.13	0.07	0.05
$\tau^2 = 0.12$	0.18	0.10	0.07
$\tau^2 = 0.36$	0.28	0.16	0.12

We utilize the smallest meta-analysis study size, i.e., $K = 5$, and the largest heterogeneity value, i.e., $\tau^2 = 0.36$, as a worst-case scenario for the estimation of $\text{Var}(\hat{\mu})$. From this value and (18) we determine the number of simulations required as:

$$\text{N}_{\text{sim}} = \frac{\text{Var}(\hat{\mu})}{(\text{MCSE Required})^2} = \frac{\frac{\frac{2}{50}+0.36}{5}}{.005^2} = 3200 \tag{21}$$

We thus perform, for each parameter combination, $\text{N}_{\text{sim}} = 3200$ simulations.

8 Reproducibility and Error Handling

The above-described simulation process will be implemented in the software R, version 4.2.0, and will be made available in the GitHub repository agaiasaracini/ORBproject. The repository contains both the R scripts of the simulations, as well as the Rnw files used to obtain the pdf of the reports, including this protocol. The simulations are executed using the Euler cluster for High Performance Computing (HPC) of ETH Zurich and/or additional computing services of the University of Zurich (UZH). To guarantee reproducibility, the random seed of the simulation will be fixed to `seed = 123` and the R package `doRNG`, version 1.8.6, will be used to perform reproducible parallel foreach loops. A `sessionInfo()` output with more information on R environment and code to reproduce the simulation study will be made available in the GitHub repository.

The functions and R scripts of the simulations process make use of the `TryCatch()` functionality so as to stop the simulation in case of an error, and record the results up until that point. In case of convergence issues for the MLE and/or PL CI of the parameters of interest presented in the Methods Section 6, the estimate and/or bounds of the CI will be denoted as `NA` and excluded from the simulation results, i.e., they will not be replaced with new simulations. The number of valid outputs per parameter combination will be recorded. Furthermore, as noted in Section 4.2 on ORB DGM, as at least 2 reported study outcomes are deemed necessary to conduct a meta-analysis and in particular to estimate the heterogeneity, simulated meta-analysis datasets with less than 2 reported outcomes will be excluded and the simulation will be repeated until a dataset with at least 2 reported outcomes is obtained. The number of trials needed to obtain this will be recorded per parameter combination. We do not expect severe issues in simulation process, also based on the exploratory simulations conducted in Saracini and Held [19]; however, in case of major problems and failures, the parameters of the simulation may be adjusted post hoc. This would be indicated in the discussion.

References

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

[2] Chan, A.-W., A. Hróbjartsson, M. T. Haahr, P. C. Gøtzsche, and D. G. Altman (2004). Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *JAMA 291*(20), 2457–2465.

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

[4] Copas, J., A. Marson, P. Williamson, and J. Kirkham (2019). Model-based sensitivity analysis for outcome reporting bias in the meta analysis of benefit and harm outcomes. *Statistical Methods in Medical Research 28*(3), 889–903. PMID: 29134855.

[5] Dear, K. B. G. and C. B. Begg (1992). An approach for assessing publication bias prior to performing a meta-analysis. *Statistical Science 7*(2), 237–245.

[6] DerSimonian, R. and N. Laird (1986). Meta-analysis in clinical trials. *Control Clinical Trials 7*(3), 177–88.

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

[8] Hedges, L. V. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science 7*(2), 246–255.

[9] Hedges, L. V. and J. L. Vevea (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*(4), 299–332.

[10] Higgins, J. and S. Thompson (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine 21*, 1539–1558.

[11] IntHout, J., J. P. A. Ioannidis, and G. F. Borm (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.

[12] Iyengar, S. and J. B. Greenhouse (1988). Selection models and the file drawer problem. *Statistical Science* (1), 109–135.

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

[14] Macaskill, P., S. D. Walter, and L. Irwig (2001). A comparison of methods to detect publication bias in meta-analysis. *Statistics in Medicine 20*(4), 641–654.

[15] Moreno, S. G., A. J. Sutton, A. Ades, T. D. Stanley, K. R. Abrams, J. L. Peters, and N. J. Cooper (2009). Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Medical Research Methodology 9*(1), 2.

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

[17] Normand, S. (1999). Meta-analysis: formulating, evaluating, combining, and reporting. *Statistics in Medicine 18*(3), 241–363.

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

[19] Saracini, A. G. and L. Held (2023, August). Addressing outcome reporting bias in meta-analysis: A comprehensive review and future directions. Master's thesis.

[20] Siepe, B. S., F. Bartoš, T. P. Morris, A.-L. Boulesteix, D. W. Heck, and S. Pawel (2023, October). Simulation studies for methodological research in psychology: A standardized template for planning, preregistration, and reporting. *PsyArXiv*.

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

[22] Sutton, A. J., F. Song, S. M. Gilbody, and K. R. Adams (2000). Modelling publication bias in meta-analysis: a review. *Statistical Methods in Medical Research 9*(5), 421–45.

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

[24] Viechtbauer, W. (2006). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine 26*(1), 37–52.