UNIVERSITÄTSMEDIZIN GÖTTINGEN UMG

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

Comparison of Pooling Strategies for Meta-Analysis involving Rare Events: Simulations and Practical Implications

Master of Science
in Applied Statistics

Author:	Phil Tobeck
Student No.:	21761994
Supervisors:	Supervisor: Prof. Dr. Maike Hohberg
	Second reviewer: Dr. Christian Roever
University:	Georg-August-Universität Göttingen
	Universitätsmedizin Göttingen: Institut für Medizinische Statistik
Place, Date:	Göttingen, November 17, 2025

Contents

List of Figures

1 Introduction

Meta-analysis is a foundational methodology in medical research, facilitating the synthesis of evidence from multiple clinical trials to yield more precise estimates of treatment effects. In fields such as cardiology, oncology, and epidemiology, individual studies often lack adequate statistical power or produce inconclusive outcomes. Meta-analytic approaches systematically aggregate information, thereby improving the validity of inferences. Integrating data across studies makes it possible to quantify overall treatment efficacy, identify heterogeneity sources, and assess the generalizability of results. The method is especially valuable for rare adverse events, where individual trials may report few or no cases, requiring specialised modelling techniques to achieve stable estimation. The reliability of meta-analytic inference depends on the underlying data-generating processes and the assumptions imposed on study-level parameters. Several factors significantly affect the performance of meta-analytic models, including between-study heterogeneity in true effects, imbalances in treatment-to-control randomisation ratios, and the presence of rare events (Sweeting et al., 2004; Mathes and Kuss, 2018). High heterogeneity increases estimator variance and may introduce bias if normality assumptions are not met. Imbalanced randomisation can lead to disproportionate weighting of study arms and bias estimates toward the null. Sparse data scenarios, such as single-zero or double-zero studies, further complicate inference because traditional continuity corrections or study exclusions can distort pooled estimates (Ren et al., 2019; Zabriskie et al., 2024). Clarifying how these design features influence estimator performance is crucial for methodological transparency and guiding evidence synthesis. Despite progress in modelling heterogeneity and sparse outcomes, variability in baseline risk across studies is often implicitly treated in most simulation studies. Baseline risk, commonly defined as the control-group event probability, directly affects the variance of the estimated log-odds ratio and influences bias and efficiency. However, existing literature rarely addresses baseline variability as a primary parameter. Instead, it is typically incorporated indirectly through assumptions about event probabilities or fixed study-level effects. When pronounced heterogeneity, imbalance, or sparse data are present, failing to model baseline variability explicitly can result in misleading conclusions regarding estimator robustness. This thesis systematically compares the effects of explicitly considering baseline pooling in meta-analysis on model performance. Specifically, the impact of baseline variance on bias, mean squared error, and coverage is assessed across diverse simulation scenarios. Additionally, cases in which baseline risks are correlated with study size are examined, reflecting real-world study designs such as small trials with higher-risk populations and extensive trials with broader, lower-risk cohorts. The remainder of this thesis is structured as follows. First, current research on model performance in meta-analyses with rare events is reviewed, emphasising methodological challenges and research gaps. The theoretical foundations and

statistical formulations of the models considered in this study are then introduced. These models include the Inverse Variance (IV) method, Mantel–Haenszel (MH) method, Random Effects (RE) model estimated using Restricted Maximum Likelihood (REML), and the Hartung–Knapp–Sidik–Jonkman (HKSJ) adjustment. An extensive simulation study is then conducted to evaluate the performance of these models under varying baseline-risk variances, heterogeneity levels, and correlation structures. Finally, the findings are applied to the rosiglitazone dataset (Nissen and Wolski, 2007), demonstrating the application of methodological insights to real-world meta-analytic data. All simulation scripts and data generation functions were implemented in `R`, and the complete codebase, including documentation and example workflows, is available on the GitHub repository: `https://github.com/PhiltonM/PoolingSimStudy`.

2 State of Research

Extensive research has investigated the impact of data characteristics and model assumptions on the performance of meta-analytic estimators. The literature primarily addresses factors such as heterogeneity, treatment allocation imbalance, and sparse data. These factors contribute to bias and instability in classical pooling methods. Additionally, studies have examined baseline risk variability and the effects of violating distributional assumptions on model performance. However, most investigations address baseline variability in isolation, or as a secondary consideration, rather than as a central methodological parameter. Ghidey et al. (2013) fitted structural and functional models to simulated scenarios where they varied the distribution of the baseline risk over a grid of four different k-values ($k = 10, 20, 50, 100$) and two different error variance values. They found that as long as the normality assumption for the baseline risk is violated, the estimates of the structural models were slightly less biased than the parametric estimates. Arends et al. (2000) proposed a hierarchical Bayesian modelling approach to estimate the relationship between baseline and treatment risk and applied it to three different meta-analysis datasets. Thompson and Sharp (1999) examined whether a study-level covariate can explain the observed heterogeneity in meta-analysis results by comparing a weighted normal errors regression with a logistic regression method applied to different meta-analysis datasets. Given these approaches to evaluate the influence of baseline variability and violation of distributional assumptions on the model performance, it is apparent that a comprehensive simulation study regarding the effect of baseline variation in the presence of rare events and a limited number of studies has not been conducted.

Sweeting et al. (2004) found that applying a fixed 0.5 continuity correction biases the estimates towards the null, especially in the case of imbalanced randomisation. They simulated sparse 2×2 tables with varying randomisation ratios. They fitted a MH model, an IV model with corrections, a Peto model, a logistic regression model, and a Bayesian

fixed-effect model. Among these, the logistic regression model proved to be nearly unbiased. Furthermore, Bradburn et al. (2007) showed that the IV method exhibits substantial bias in rare-event settings when analysing multiple trials with baseline event rates between 0.1% and 10%, risk ratios between 0.2 and 1, and both balanced and unbalanced treatment arms. The models applied in their study consisted of the MH model with and without CC, the DL model, the IV method, logistic regression, the Peto method, the Crude method and the exact method. They concluded that the MH and logistic regression models were more robust, while the Peto method performed best for very rare events but failed under imbalance. Spittal et al. (2015) fitted IV models with and without continuity correction and Poisson (FE and RE) models to simulated incidence rate data with varying percentages of zero-events, heterogeneity, and numbers of studies. They observed poor performance of the IV method with continuity correction, while Poisson random-effects methods were more robust, even with many zeros.

Kuss (2015) simulated data with parameters calibrated from empirical meta-analyses including single-zero and double-zero arms, fitting IV models, MH models, Peto models, and beta-binomial regression models. They showed continuity correction distorts results and recommend beta-binomial models for analysing datasets with single- and double-zero arms. Cheng et al. (2016) analysed 2500 simulated datasets with varying baseline risk, treatment effect, sample size per arm, and heterogeneity. The fitted Peto, MH, and IV models found that including double-zero studies reduced the MSE and improved coverage when the true effect was equal to zero, while exclusion reduced bias when the true effect was large. Ren et al. (2019) fitted IV models with continuity correction, MH models with CC, Peto models, Bayesian models, and exact models to 368 Cochrane meta-analyses with varying proportions of double-zero studies. They found significant disagreement between models when the proportion of double-zero studies was high and recommend the Bayesian or exact methods. Zabriskie et al. (2024) fitted MH, IV, DL, and RE models with different CC strategies to simulated datasets with varying heterogeneity and zero-event rates, finding no universally optimal CC since model performance depended heavily on heterogeneity. Tsujimoto et al. (2024) analysed 885 Cochrane reviews with single-zero studies by fitting MH models with and without CC, IV RE models with CC, Peto models, and Bayesian binomial-normal models. They reported that results changed in about 30% of cases when CC was applied. Andreano et al. (2015) simulated competing risks and censored data with varying incidence and heterogeneity, fitting IV meta-analyses of arcsine, logit, and raw incidence and GLMMs. They found that the arcsine transformation had the lowest bias among the variance-based models, while GLMMs performed best at very low incidence.

Piaget-Rossel and Taffé (2019) simulated data for rare events under a homogeneous treatment effect while varying baseline risk and randomisation ratio. They showed that the MH and binomial regression models were robust under imbalance, while the IV method

failed to provide reliable results. Rücker and Schumacher (2008) analysed real rosiglitazone trials with correlation between baseline event risk and allocation ratio. By fitting a common-effect model, a RE model, and baseline-risk adjusted models, they demonstrated that such correlations may induce Simpson's paradox and that ignoring baseline effects can distort pooled estimates.

Langan et al. (2019) simulated rare binary outcomes with equal and unequal study sizes across varying levels of heterogeneity. They fitted nine different meta-analysis estimators and concluded that the RE model provides the most reliable performance. Other studies have used simulation designs where baseline risk was modelled as a covariate and error-in-variables approaches were considered using linear models with or without distributional assumptions, concluding that baseline risk can explain a large portion of heterogeneity. At the same time, mis-specification can introduce bias (Arends et al., 2000; Walter, 1997; Ghidey et al., 2013).

Mathes and Kuss (2018) evaluated the performance of beta-binomial, IV, MH, and Peto models in scenarios with between 2 and 5 studies, varying both heterogeneity and randomisation ratios. They concluded that the beta-binomial model performed best, while the IV method was unreliable even with more than ten studies. Günhan et al. (2020) compared the performance of Bayesian BNHM with weakly informative priors (WIPs) against ML and standard Bayesian models using simulated data with few studies and rare events. They found that WIPs reduced bias and improved coverage, while ML was biased. Seide et al. (2019) applied DL, REML RE models, and likelihood-based RE models to simulated data with small numbers of studies, imbalanced study sizes, and varying heterogeneity, concluding that likelihood-based RE models with HKSJ adjustments perform better than DL. Partlett and Riley (2017) investigated confidence intervals from a standard RE model and an RE model with HKSJ adjustment, simulating data with few studies, unequal study sizes, and heterogeneity. They found that HKSJ improves coverage, while standard confidence intervals under-cover. Schulz et al. (2024) reanalysed 60,000 sparse-data Cochrane meta-analyses by fitting Peto, DL, PM, RE, GLMM, and beta-binomial models, concluding that one-stage models produce more conservative estimates while being more stable in the presence of zero-event trials.

Jackson et al. (2018) compared seven RE models with Cochrane-inspired datasets, finding that GLMMs outperformed two-stage models regarding bias and coverage but were susceptible to numerical instability. Beisemann et al. (2020) simulated 162 scenarios varying the number of studies, study sizes, heterogeneity, baseline risk, and randomisation. They fitted RE Poisson models, zero-inflated Poisson models, beta-binomial models, and IV models. The RE Poisson model was the most robust, the beta-binomial provided a competitive alternative, while the zero-inflated Poisson suffered from convergence issues. Jansen and Holling (2023) created a large grid of scenarios with varying event rates, heterogeneity, and numbers of studies using two DGMs. They compared a hypergeometric-

normal GLMM, a beta-binomial model, a MH model, and an IV model, concluding that the hypergeometric-normal GLMM performed best with moderate or large heterogeneity. In contrast, the beta-binomial was more robust across DGMs. Sangnawakij et al. (2024) performed an arm- and contrast-based nonparametric RE estimation using mixture likelihoods, Poisson, and DL models. They found that contrast-based mixture models were superior for the RR, achieving better bias and MSE.

Yao et al. (2023) investigated the integration of real-world evidence (RWE) and RCT data in Bayesian frameworks, comparing naive pooling, adjusted pooling, bias-corrected models, and hierarchical three-level models under different proportions of RWE, treatment effects, and study sizes. Their findings suggest that bias-corrected Bayesian pooling yields the most reliable estimates, with RWE improving certainty in rare-event contexts. Yao et al. (2024) simulated data with varying numbers of studies, effect sizes, and event probabilities to compare four parametrisations of Bayesian random-effects models (BNHM, beta-binomial, and GLMM variants) using both weakly informative and non-informative priors. They concluded that WIPs consistently outperformed NIPs regarding bias and coverage, while some parametrisations led to unstable inference. Furthermore, Yao et al. (2025) compared different priors for the heterogeneity parameter (half-normal, Turner, uniform, and others) on simulated rare-event datasets with varying degrees of heterogeneity. They found that half-normal and Turner priors provided the most robust performance regarding bias and coverage. In addition, Pateras et al. (2021) analysed the influence of priors for variance parameters in sparse-event Bayesian RE models, recommending priors that concentrate mass at small heterogeneity values, since vague priors tended to overestimate heterogeneity.

Pateras et al. (2018) compared three different data-generating models (DGMs) for dichotomous outcomes to investigate how methodological choices affect simulation results. By fitting IV, GLMM, and beta-binomial models to simulated datasets, they demonstrated that apparent estimator performance depends strongly on the assumed DGM, cautioning against over-interpreting single designs. Kulinskaya et al. (2021) conducted a systematic investigation varying the control-arm event probability distributions, the overall effect sizes, and the distribution of study sizes when comparing IV, GLMM, and sample-size-based weighting methods. They showed that method performance is susceptible to simulation set-up and that conclusions about estimator superiority may reverse under alternative DGMs. Finally, van den Heuvel et al. (2024) contrasted simulations based on aggregated summary statistics with those using individual participant data (IPD), fitting DL, PET-PEESE, and Trim and Fill methods to explore small-study effects. They concluded that aggregated simulation shortcuts may misrepresent true estimator properties and that IPD-based designs are preferable for evaluating pooling strategies.

In summary, prior research has significantly advanced the understanding of estimator performance in meta-analyses involving heterogeneity, imbalance, and sparse data. Neverthe-

less, most studies address baseline-risk variability only implicitly, either as a consequence of the data-generating process or as a covariate, rather than as a primary methodological parameter. Comprehensive simulation studies are lacking to examine the interaction between baseline variability and factors such as high heterogeneity, imbalanced randomisation, and small study numbers. Furthermore, the influence of correlation between baseline risk and study size remains underexplored, representing a critical gap in understanding the structural dependencies that affect the robustness of meta-analytic estimators.

3 Theoretical Background

Let us consider the case of k different studies, where the estimated effect size in each study is denoted by y_i where $i = 1, ..., k$. Each of those estimated study effect sizes has an associated standard error s_i, which describes how the estimated study effect size varies from the true study effect size θ_i. It is important to note here that the true study effect size can be the same across all studies or different across all studies, depending on which additional assumptions we make. We will talk about this more thoroughly in the following sections. With these assumptions, we can write the distribution of the estimated effect sizes y_i as follows:

$$y_i|\theta_i, s_i \sim \mathcal{N}(\theta_i, s_i^2) \tag{1}$$

Furthermore, the standard error s_i is typically treated as a known and fixed parameter. At the same time, let ζ_i denote the baseline risk probability, which can be drawn from multiple different distributions, which we will elaborate further on in section 4.1. It is important to note here that since the Odds-Ratio is defined as

$$y_i = \frac{p_{treatment}}{1 - p_{treatment}} - \frac{\zeta}{1 - \zeta} \tag{2}$$

where $p_{treatment}$ denotes the event probability in the treatment group, the variance of ζ contributes to the variance of y_i. Therefore, an increased variance of ζ could influence the reliability of estimates for the effect size in the context of meta-analyses.

3.1 Inverse-Variance Method

For our first considered model, the Inverse Variance Method (IV), we assume an equal true effect size $\theta_i = ... = \theta_k = \theta$ so that the study-specific effect size is drawn from a distribution $y_i \sim \mathcal{N}(\theta, s_i^2)$. Thus, this model only considers the within-study variance s_i^2 as a source of uncertainty. This leads to a closed-form estimator for the true effect size $\hat{y}$

of the form:

$$\hat{y}_{(IV)} = \frac{\sum_{i=1}^{k} w_i y_i}{\sum_{j=1}^{k} w_j} \text{ with } w_i = \frac{1}{s_i^2} \tag{3}$$

and (4)

$$Var(\hat{y}_{(IV)}) = (\sum_{i=1}^{k} w_i)^{-1} \tag{5}$$

This closed-form estimate is unbiased as long as the normality assumption and the no-heterogeneity assumption about the true study effect size is fulfilled since:

$$Bias(\hat{y}_{(IV)}) = E(\hat{y}_{(IV)}) - \theta \tag{6}$$

$$= \frac{\sum_{i=1}^{k} w_i E(y_i)}{\sum_{j=1}^{k} w_j} - \theta \tag{7}$$

$$\overset{\text{for } y_i \sim \mathcal{N}(\theta, \int_{\rangle}^{\in})}{=} \frac{\sum_{i=1}^{k} w_i}{\sum_{j=1}^{k} w_j}\theta - \theta = 0 \tag{8}$$

However, the assumption of no-heterogeneity is frequently questioned (Wolfgang Viechtbauer, 2005). Additionally, the IV method faces challenges when confronted with double-zero and single-zero studies since the logORs are not defined in case of zero studies, leading to the omission of these studies (Spittal et al., 2015; Jansen and Holling, 2023). This leads to an increased bias of the estimates (Spittal et al., 2015). Besides the problems with heterogeneity assumptions and the treatment of zero studies regarding the IV method, simulation studies have also shown that this method shows a very high bias in the presence of imbalanced randomisation compared to other standard models (Sweeting et al., 2004). Therefore, the Inverse Variance Method is only considered here to highlight the potential problems of inappropriate pooling strategies compared with the other models. To address the IV Method's issues with single-zero or double-zero studies, we have decided to implement a variant of the IV method with continuity correction. Hereby, 0.5 gets added to each zero cell in the 2×2 contingency table. This enables the model not to exclude studies where one cell is zero while still keeping the low event rate of the study. However, studies have shown that continuity correction biases the estimate towards zero (Jackson et al., 2018)). Also Sweeting et al. (2004) have demonstrated that the IV method with continuity correction exhibits a very high bias in the presence of imbalanced randomisation. Also, the coverage of the IV method with continuity correction has been shown to be excessively high (Sweeting et al., 2004).

3.2 Mantel-Haenszel Estimator

The Mantel-Haenszel (MH) estimator is one of the most frequently applied methods for pooling binary outcomes across studies, particularly in the presence of sparse data. In

contrast to the inverse-variance method, which relies on normal approximations of the log odds ratios and their variances, the MH method directly operates on each study's 2×2 contingency tables. It provides an approximately unbiased estimate of the common odds ratio even when individual studies contain few events (Bradburn et al., 2007; Sweeting et al., 2004). Let a_i and c_i denote the number of events in the treatment and control groups, and b_i and d_i are the corresponding non-events in the study i. The total number of participants in study i is $n_i = a_i + b_i + c_i + d_i$. The MH estimator for the common odds ratio $\hat{\theta}_{MH}$ is then given by:

$$\hat{\theta}_{MH} = \frac{\sum_{i=1}^{k} \frac{a_i d_i}{n_i}}{\sum_{i=1}^{k} \frac{b_i c_i}{n_i}}. \tag{9}$$

In practice, working on the logarithmic scale is often more convenient. Thus, the log odds ratio is obtained as $\hat{y}_{MH} = \log(\hat{\theta}_{MH})$ with corresponding variance

$$Var(\hat{y}_{MH}) = \frac{1}{2}\left(\frac{1}{\sum_{i=1}^{k} \frac{a_i d_i}{n_i}} + \frac{1}{\sum_{i=1}^{k} \frac{b_i c_i}{n_i}}\right). \tag{10}$$

The MH approach has been shown to provide robust estimates in balanced designs and under moderate event sparsity. However, it may exhibit bias when a strong imbalance between treatment and control group sizes occurs or when event rates are extremely low (Bradburn et al., 2007; Piaget-Rossel and Taffé, 2019). Moreover, while the method implicitly down-weights small studies with no events, it does not account for between-study heterogeneity. It is therefore best interpreted as a fixed- or common-effect estimator (Sweeting et al., 2004).

3.3 Random-Effects Model

To solve the problem of heterogeneity between studies we decided also to assess the performance of the Random-Effects Model (RE) which assumes that the different true study effect sizes themselves are random realizations from a distribution $\theta_i \sim \mathcal{N}(\mu, \tau^2)$ where μ describes the true effects size and τ the in-between study heterogeneity. This changes the distributional assumption about our observed study-specific effect size y_i as follows:

$$y_i | \theta_i, s_i \sim \mathcal{N}(\mu, \tau^2 + s_i^2) \tag{11}$$

The estimator for τ must be found by an iterative likelihood-based procedure. In our case, we have decided on the REML approach since it has been shown that REML estimates show more desirable statistical properties (e.g. a reduced bias) than the classical Maximum Likelihood (ML) approach (Wolfgang Viechtbauer, 2005; Jackson et al., 2018).

The modified log-likelihood

$$l = -\frac{k}{2}ln(2\pi) - \frac{1}{2}\sum_{i=k}^{k} ln(s_i^2 + \tau^2) - \frac{1}{2}\sum_{i=k}^{k} \frac{(y_i - \hat{\mu})}{s_i^2 + \tau^2} - \frac{1}{2}ln\left(\sum_{i=k}^{k} \frac{1}{s_i^2 + \tau^2}\right) \tag{12}$$

with

$$\hat{\mu} = \frac{\sum_{i=1}^{k} w_i y_i}{\sum_{j=1}^{k} w_j} \text{ and } w_i = \frac{1}{s_i^2}$$

yields

$$\hat{\tau}^2_{REML} = max\left(0, \frac{\sum_{i=1}^{k} a_i^2((y_i - \hat{\mu})^2 - \hat{s_i^2})}{\sum_{i=1}^{k} a_i^2} + \frac{1}{\sum_{i=1}^{k} a_i}\right) \tag{13}$$

These estimates have shown desirable statistical properties (Wolfgang Viechtbauer, 2005; Langan et al., 2019) but do struggle in the presence of zero studies and small sample sizes. In case of zero studies the standard procedure is similar to the IV method since zero studies are also excluded from the analysis leading to estimates that are biased away from zero (Jackson et al., 2018). For small sample sizes ($k < 10$), the estimates are often imprecise (Langan et al., 2019) or can struggle to converge at all, as Iaquinto et al. (2025) have shown for $k < 20$. Additionally, Jackson et al. (2018) have noted that in case of a violation of the normality assumption, the estimate for τ can exhibit positive bias, especially for a moderate or no in-between study heterogeneity. Similar to the IV method, we have also decided to implement the RE model with a continuity correction of 0.5 to include the previously omitted studies with zero cells, leading to the same difficulties as in the IV method (Sweeting et al., 2004).

3.4 Hartung-Knapp-Sidik-Jonkman (HKSJ) Adjustment

Hartung and Knapp (2001) and later Sidik and Jonkman (2002) proposed an alternative variance adjustment to improve inference in random-effects meta-analysis, especially when the number of studies is small. The Hartung-Knapp-Sidik-Jonkman (HKSJ) method modifies the conventional normal-approximation confidence interval by incorporating uncertainty in the estimation of the between-study variance τ^2 (Partlett and Riley, 2017; Seide et al., 2019). Given a random-effects model with study-specific estimates $y_i \sim \mathcal{N}(\mu, \tau^2 + s_i^2)$, the pooled estimator $\hat{\mu}$ is obtained as

$$\hat{\mu} = \frac{\sum_{i=1}^{k} w_i y_i}{\sum_{i=1}^{k} w_i}, \quad \text{where } w_i = \frac{1}{s_i^2 + \hat{\tau}^2}. \tag{14}$$

The conventional DerSimonian-Laird or REML confidence interval for $\hat{\mu}$ assumes a normal distribution. The HKSJ adjustment, in contrast, estimates the variance of $\hat{\mu}$ as

$$\widehat{Var}_{HKSJ}(\hat{\mu}) = \frac{1}{k-1}\sum_{i=1}^{k} w_i(y_i - \hat{\mu})^2 \Big/ \left(\sum_{i=1}^{k} w_i\right)^2, \tag{15}$$

and constructs confidence intervals based on the t-distribution with $k-1$ degrees of freedom:

$$\hat{\mu} \pm t_{k-1,\,1-\alpha/2}\sqrt{\widehat{Var}_{HKSJ}(\hat{\mu})}. \tag{16}$$

This approach provides more conservative confidence intervals with improved coverage, particularly in small-sample settings or when heterogeneity is substantial. Simulation studies have consistently shown that the HKSJ method outperforms conventional normal-based intervals in maintaining nominal coverage, while retaining comparable bias properties ((Partlett and Riley, 2017; Seide et al., 2019; Mathes and Kuss, 2018).

4 Simulation Study

This simulation study aims to assess the performance of different meta-analysis models in scenarios facing different variability or distributions for the baseline risk, combined with a sparse data setting where single-zero and double-zero cells are common, and the number of studies is small. Additionally, the models must deal with imbalanced randomisation and correlation between the baseline risk and the study size.

4.1 Simulation Framework

For the data-generating mechanism we simulated data $10,000$ times over a design grid with four different numbers of studies ($k \in \{3, 5, 10, 20\}$), four different baseline risk variances on a logit scale ($\sigma_{logit} \in \{0.0, 0.1, 0.2, 0.5, 1.0, 2.0\}$) (except for the *Many Zeros scenario* see section 4.1.2) and four different between study heterogeneity parameters ($\tau_{logit} \in \{0, 0.1, 0.5, 1)$ while ensuring that for each simulated scenario $\tau \leq \sigma$, the true overall effect in the treatment arm is kept constant at $\theta_{logit} = log(2)$. Firstly, we draw the total study sizes, where we assume that the study sizes are correlated with each other (except for the correlation between baseline risk and study size scenarios, see section 4.1.4) to emulate the situation that meta-analyses from one field usually have similar trial sizes, so that

$$\log(N_i) = X_{\text{common}} + Y_i, \quad X_{\text{common}} \sim \mathcal{N}(\mu_{\log N},\, \sigma^2_{\text{between}}),\; Y_i \sim \mathcal{N}(0,\, \sigma^2_{\text{within}}) \tag{17}$$

where X_{common} describes the random effect for the study size between different meta-analyses datasets and Y_i the study-level deviation of the study size within a meta-analyses dataset. We can then decompose the in-between and within variance of the study sizes into a variance and correlation component:

$$\sigma_{\text{between}} = \sqrt{\rho_{\text{size}}}\,\sigma_{\log N}, \quad \sigma_{\text{within}} = \sqrt{1-\rho_{\text{size}}}\,\sigma_{\log N} \tag{18}$$

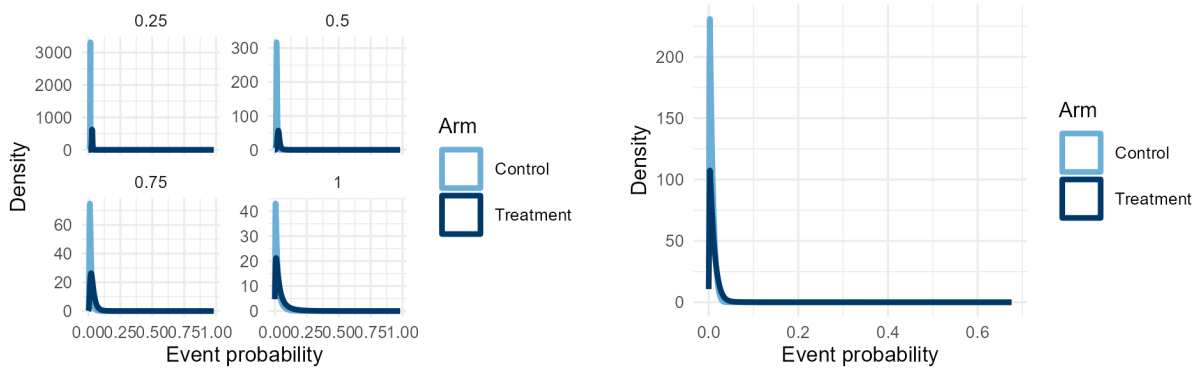
For all scenarios we have decided to use $\sigma_{\log N} = 1, \quad \rho_{\text{size}} = 0.75, \quad \mu_{\log N} = 5$. Afterwards, we split the study sizes into a treatment (n_{1i}) and control arm (n_{0i}) based on the randomisation ratio. The baseline logits are then sampled from either a log-normal distribution with parameters $\mu_{logit} = -4.184591$ and variance σ^2_{logit} or the baseline risk is sampled from a uniform distribution with parameters `baseline_min` and `baseline_max`. Afterwards we compute the true study level treatment effects δ_i by drawing values from a uniform distribution so that $\delta_i \sim \mathcal{U}(\theta_{logit} - \tau_{logit}, \theta_{logit} + \tau_{logit})$. These values are then transformed to probability scale treatment and control effects to sample the numbers of events in each arm from a binomial distribution.

4.1.1 Scenario: Standard

The *standard scenario* represents an idealised situation in which the data-generating process follows a balanced and moderately heterogeneous design. The true treatment effect is set to $\theta_{logit} = \log(2)$, corresponding to an odds ratio of 2, with control-arm event probabilities drawn from a logit normal distribution with Lognormal$(\mu_{logit} = -4.184591, \sigma^2_{logit})$ so that we have a mean control event probability of $\simeq 1.5\%$ and our four different values for the variance and between study heterogenity. Randomisation between treatment and control arms is balanced (1:1), and no correlation between baseline event risk and study size is assumed. This setting provides insights into the effect that varying variance of the baseline can have under an otherwise very ideal scenario.
For the *standard scenario*, the simulated data show the intended characteristics of a realistic meta-analytic setting. The distribution of control and treatment risk is left-skewed, producing many smaller probabilities and only a few larger ones (see figure 1a)). We can also see that the increasing variance for the baseline directly influences the variability of the treatment risk.

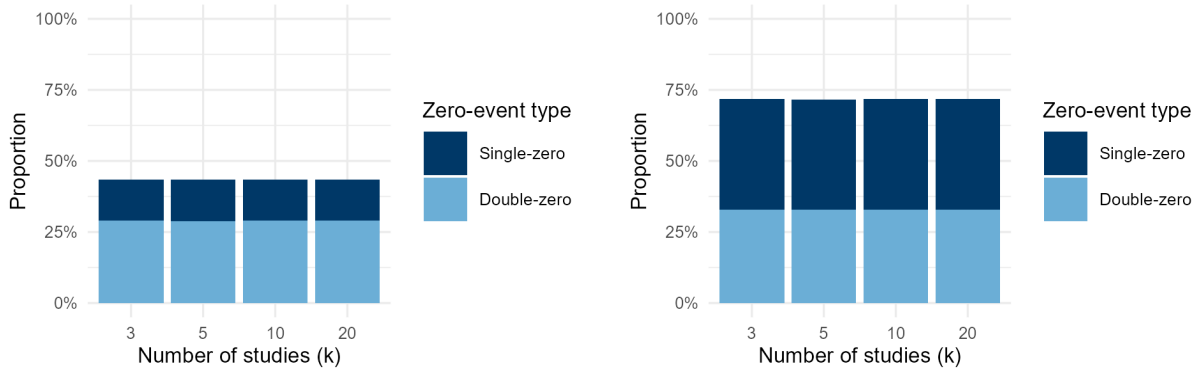
4.1.2 Scenario: Many Zeros

The *many zeros* scenario represents an extreme sparse-data situation designed to emulate very low event probabilities and violated distributional assumptions of the baseline risk. Here, the baseline event rate is drawn from a beta distribution with $\zeta_i \sim Beta(\alpha = 1, \beta = 7)$ to achieve a highly right skewed distribution (see figure 1b) producing an increased number of single and double zero cells in our 2×2 contingency tables while keeping the variance of the baseline constant. Under this configuration, both single-zero and double-zero studies occur frequently (see figure 2b)), challenging standard meta-analytic estimators that rely on continuity corrections or that exclude studies with zero cells. This scenario is therefore used to examine estimator robustness under severe data sparsity.



(a) Standard Scenario (b) Many Zeros Scenario

Figure 1: Control risk probability and treatment risk probability distribution for the Standard and Many Zeros Scenario

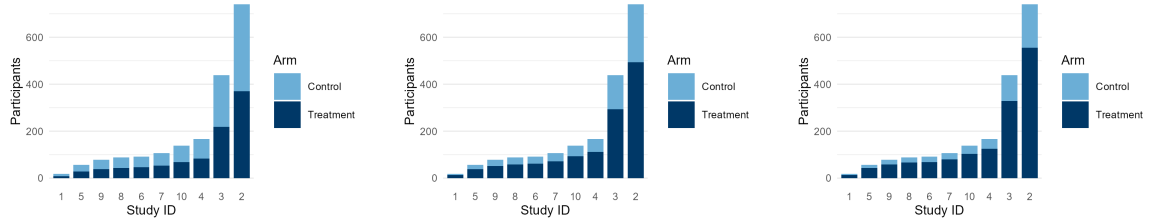


(a) Standard Scenario (b) Many Zeros Scenario

Figure 2: Percentage and composition of zero studies for the standard and many zeros scenario

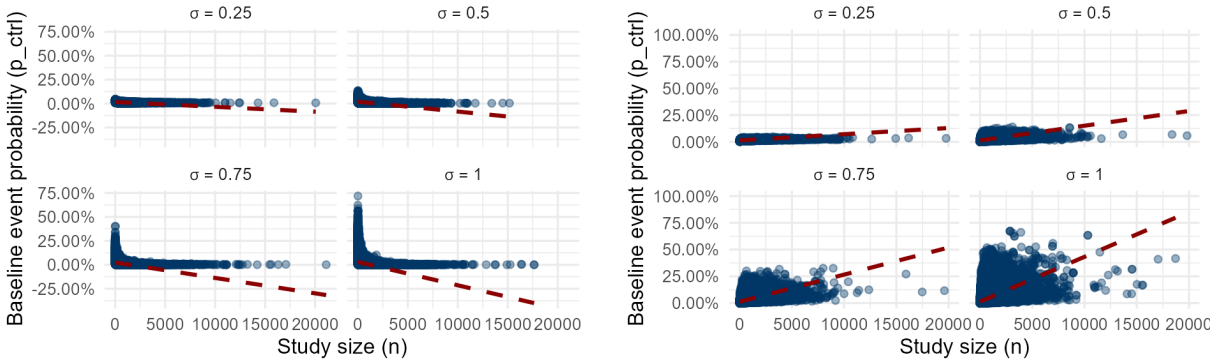
4.1.3 Scenario: Imbalanced Randomisation

The *randomisation 1:2* and *randomisation 1:3* scenarios investigate the influence of unequal group sizes on model performance. The treatment-to-control ratio is set to 1:2 or 1:3, respectively, meaning the control arm receives twice or thrice as many participants as the treatment arm (see figure 3). All other parameters are held constant at moderate effect ($\theta_{logit} = \log(2)$) and baseline risk drawn from a logit normal distribution with Lognormal($\mu_{logit} = -4.184591, \sigma^2_{logit}$). Imbalanced randomisation ratios are common in practice, especially in rare-event settings where ethical or logistical constraints limit treatment allocation, and are known to induce bias in inverse-variance and continuity-corrected estimators (Sweeting et al., 2004; Mathes and Kuss, 2018). Therefore, this scenario tests the estimators' sensitivity to allocation imbalance.



(a) Standard Scenario (b) Randomisation 1:2 (c) Randomisation 1:3

Figure 3: Control and treatment arm partition for a random dataset from the simulated datasets for the standard, randomisation 1:2 and 1:3 scenario



(a) Negative Correlation Scenario (b) Positive Correlation Scenario

Figure 4: Correlation between baseline risk and study size for the negative and positive correlation scenario

4.1.4 Scenario: Correlation between baseline risk and study size

The *positive correlation* and *negative correlation* scenarios explore the effect of an association ($\rho = \pm 0.6$) between baseline risk and study size. In this setup, the log study sizes and logit baselines are drawn jointly from a multivariate normal distribution so that we have

$$\begin{bmatrix} \alpha_i \\ \log(N_i) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_{logit} \\ \mu_{\log N} \end{bmatrix}, \begin{bmatrix} \sigma^2_{logit} & \rho\, \sigma_{logit}\, \sigma_{\log N} \\ \rho\, \sigma_{logit}\, \sigma_{\log N} & \sigma^2_{\log N} \end{bmatrix} \right), \quad \alpha_i = \text{logit}(p_{\text{ctrl},i}) \tag{19}$$

μ_{logit} is the same as in the previous scenarios with normally distributed baselines and σ_{logit} is varied using our four different values. Larger/smaller studies tend to have higher baseline event probabilities in this configuration. Smaller studies are likelier to have low/high or zero events (see figure 4). The scenario thus assesses how sensitive standard pooling methods are to correlated study characteristics.

4.2 Results

4.2.1 Standard Scenario

In the standard scenario, we observe the performance of the different pooling strategies under an idealized but realistic setup with balanced randomisation and moderate heterogeneity. For the IV method, the estimated bias (see figure 5(a)) ranges between approximately -0.1 for lower between-study heterogeneity values (τ) and 0.1 for higher τ values. This pattern indicates that the IV estimator tends to underestimate the true treatment effect when τ is small, while it overestimates it as heterogeneity increases. With higher baseline variance (σ_{logit}), the bias moves closer to zero, suggesting a slightly better model fit as baseline variability increases. Introducing a continuity correction (CC) modifies the bias predictably: when the uncorrected model underestimates the effect, adding a CC increases the bias, whereas in cases of overestimation, it pulls the bias closer to zero. The number of studies (k) has only a minor effect but tends to increase bias slightly in larger meta-analyses. Regarding the MSE (see figure 6(a)), the IV model exhibits values between 0 and 0.6, increasing with higher τ but decreasing with both larger σ_{logit} and greater k. Models with continuity correction generally yield lower MSE values than the uncorrected IV model. Coverage (see figure 7(a)) remains satisfactory overall but decreases modestly with increasing k and more sharply under higher σ_{logit} or τ.
The MH estimator consistently produces positive bias across all conditions (see figure 5(b)), thus overestimating the true effect size. This bias increases with higher τ, but, in contrast to the IV method, it tends to decrease as σ_{logit} grows—except in very small meta-analyses ($k = 3$), where a slight bias increase is observed. The MSE (see figure 6(b)) lies between 0 and 0.5 and decreases with larger k and σ_{logit}, while it increases with τ, following a similar trend as the IV model. Coverage (see figure 7(b)) is comparable to the IV estimator and remains generally acceptable, though it declines for larger heterogeneity and variance levels.
For the RE estimator, bias (see figure 5(c)) increases with the baseline variance but remains within a narrow range of approximately ± 0.1. The continuity-corrected version ($CC = 0.5$) exhibits slightly higher bias when the uncorrected model underestimates the effect and slightly lower bias when it overestimates it—mirroring the behaviour seen in the IV model. With increasing τ, bias tends to rise, and for $\tau = 1$, the estimator shifts from underestimation to overestimation, indicating a sensitivity to considerable between-study heterogeneity. The number of studies has a negligible influence on the bias magnitude. The MSE (see figure 6(c))follows the same general trend as IV and MH but is systematically lower, suggesting a more stable estimator. Regarding coverage (see figure 7(c)), the RE model outperforms the IV and MH estimators, showing stable coverage across all settings. Applying the Hartung-Knapp-Sidik-Jonkman (HKSJ) adjustment further improves coverage (see figure 7(d)) substantially while preserving the same bias and MSE

patterns as the classical RE model.

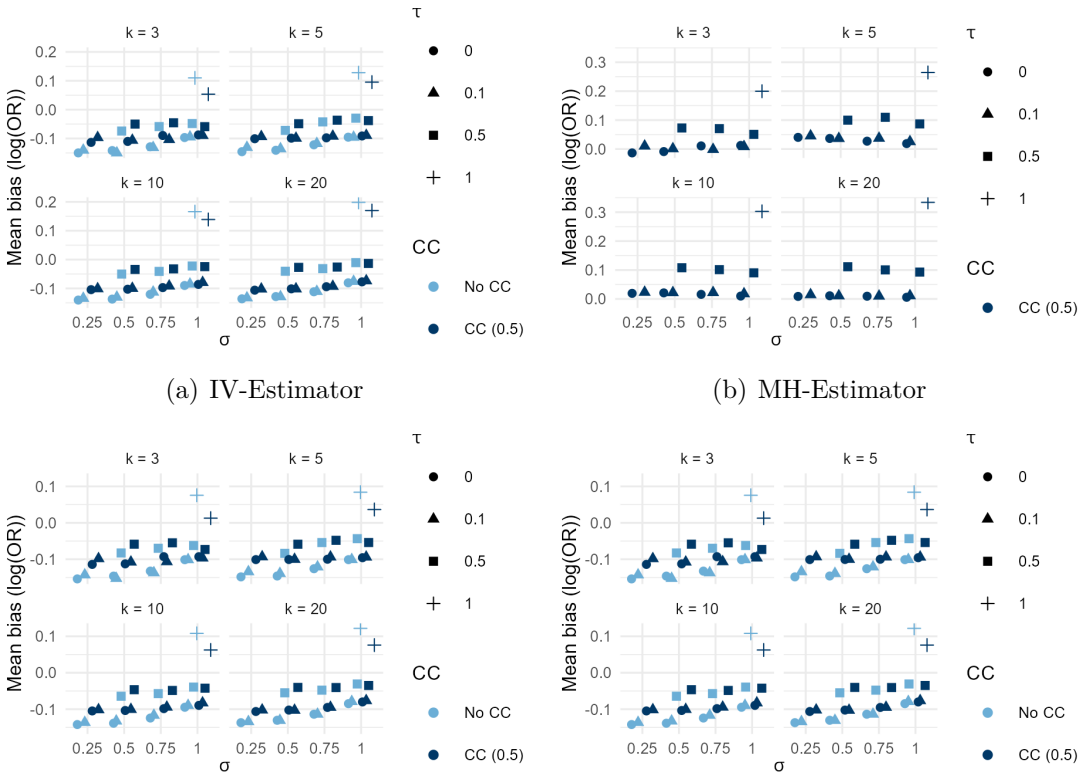
The observed results align closely with the theoretical properties discussed in Section 3. As expected, the IV method exhibits bias under heterogeneity and sparse-data conditions due to its assumption of a common effect and exclusion or distortion of zero-event studies. This behaviour is consistent with previous findings by Sweeting et al. (2004), who demonstrated that continuity corrections can induce bias towards the null, and by Bradburn et al. (2007), who observed poor IV performance in rare-event contexts. Similarly, the MH estimator's systematic overestimation of the effect agrees with earlier studies reporting robustness under moderate sparsity but bias in the presence of imbalance or considerable heterogeneity (Piaget-Rossel and Taffé, 2019). The superior performance of the RE approach in terms of both bias and MSE confirms its theoretical advantage as outlined in the random-effects formulation (Section 3.3) and corroborates simulation-based evidence by Langan et al. (2019) and Jackson et al. (2018), who showed that the RE model yields more accurate estimates of τ^2 and improved inference compared with DerSimonian–Laird or IV estimators. Finally, the enhanced coverage of the HKSJ-adjusted RE estimator aligns with the theoretical expectations from Partlett and Riley (2017) and Mathes and Kuss (2018), who demonstrated that this adjustment corrects undercoverage in small-sample settings while maintaining low bias.

In summary, the standard scenario confirms the theoretical predictions that (i) fixed-effect estimators such as IV and MH are sensitive to heterogeneity and baseline variability, (ii) the RE model provides the most balanced trade-off between bias and efficiency, and (iii) the HKSJ adjustment yields the most reliable coverage among the evaluated methods.

4.2.2 Many Zeros Scenario

The many zeros scenario represents an extreme sparse-data situation, characterised by a high proportion of single- and double-zero studies and a highly right-skewed baseline risk distribution. This setting poses a particular challenge to classical pooling estimators, as continuity corrections and the treatment of zero cells can substantially influence bias and coverage.

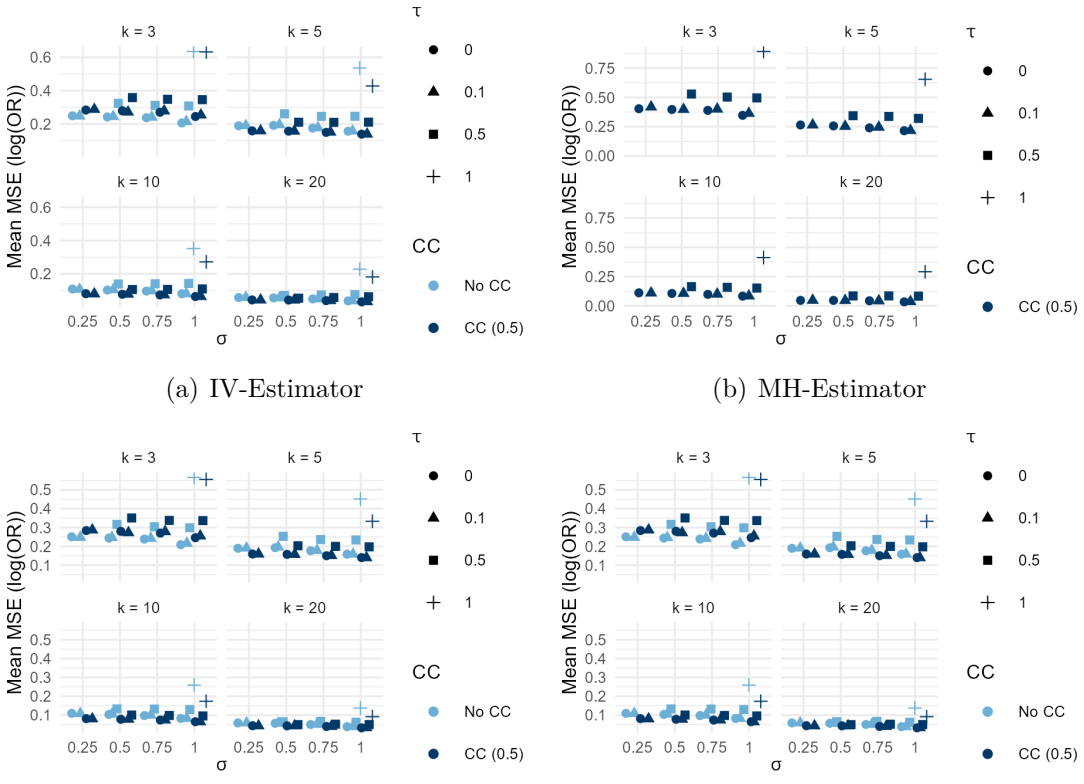
For the IV method, the estimated bias (see figure 8(a)) ranges from approximately -0.2 for small τ values to 0.2 for larger τ values. As in the standard scenario, a continuity correction (CC = 0.5) consistently pulls the bias towards zero, reducing under- and overestimation. The magnitude of this correction effect becomes more pronounced with increasing τ, particularly beyond $\tau = 0.1$. Moreover, larger study numbers (k) slightly increase the bias, indicating that additional studies with zero events may not necessarily stabilise the estimator in sparse-data contexts. The MSE (see figure 9(a))is markedly higher than in the standard scenario but decreases substantially with increasing study size. Higher τ values lead to higher MSE, reflecting the increased variability of the true effects. Applying a continuity correction reduces the MSE and bias, particularly in sce-



(a) IV-Estimator (b) MH-Estimator

(c) RE Model (d) HKSJ-Adjustment

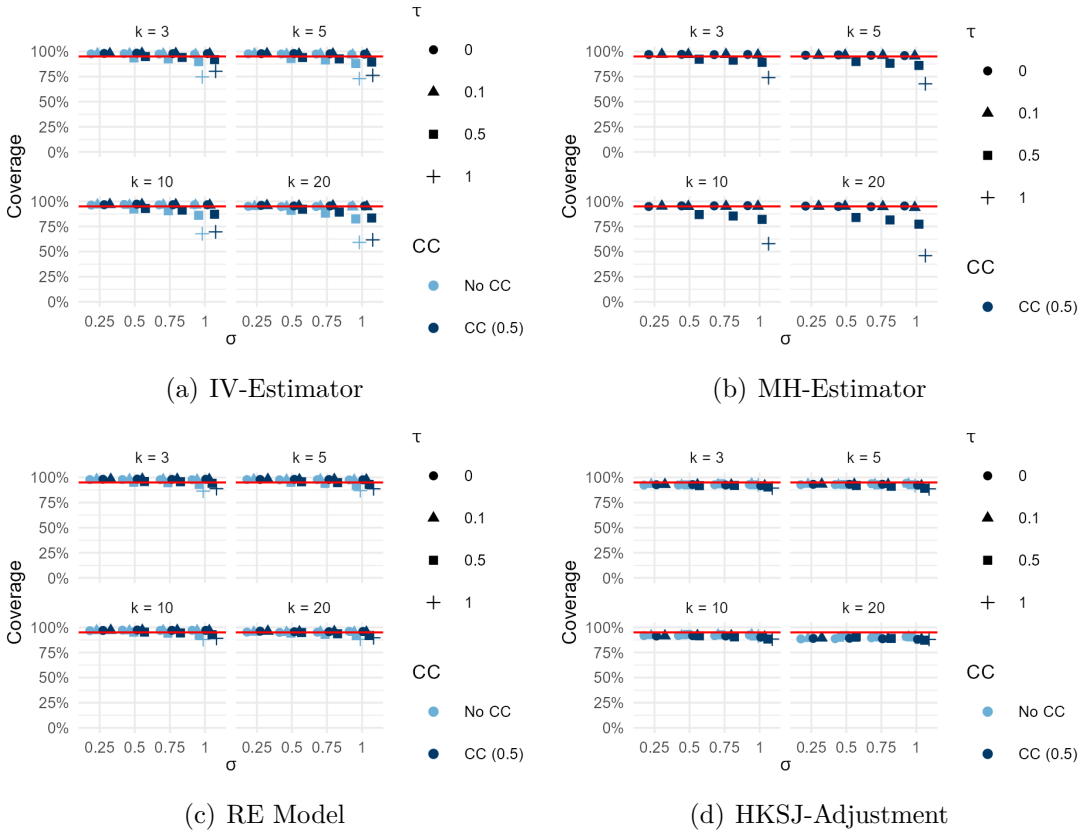
Figure 5: Mean Bias (log-OR) vs. σ_{logit} for the standard scenario



(a) IV-Estimator (b) MH-Estimator

(c) RE Model (d) HKSJ-Adjustment

Figure 6: Mean MSE (log-OR) vs. σ_{logit} for the standard scenario



(a) IV-Estimator

(b) MH-Estimator

(c) RE Model

(d) HKSJ-Adjustment

Figure 7: Mean Coverage vs. σ_{logit} for the standard scenario

narios with large heterogeneity. Coverage (see figure 10(a)) remains high overall (mostly above 0.95) but decreases with increasing numbers of studies and higher τ values. This decline is less severe when CC = 0.5 is applied, suggesting that the correction stabilises inference under extreme sparsity.

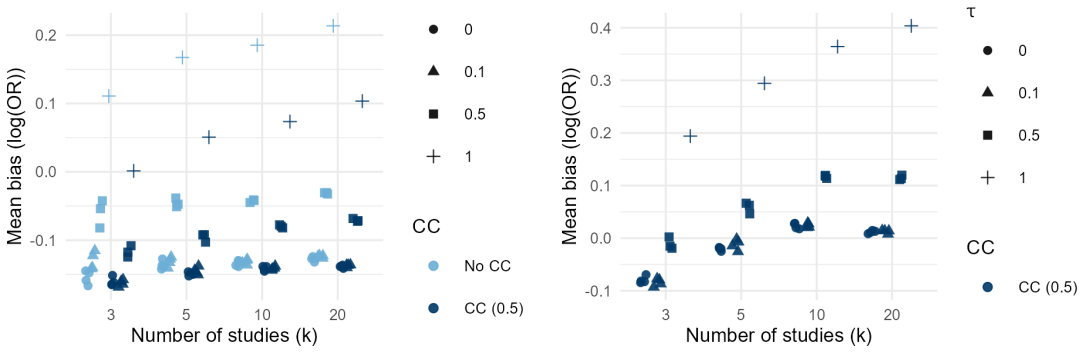
For the MH estimator, bias (see figure 8(b)) remains close to zero for $\tau < 0.5$, but for $\tau \geq 0.5$ the model increasingly overestimates the true effect size. When the number of studies is very small ($k < 5$), the MH method underestimates the effect. The MSE (see figure 9(b)) is generally higher than for the IV model but follows a similar trend: it decreases markedly with larger k and increases with higher τ. The MH estimator's coverage (see figure 10(b)) shows the same general pattern as that of the IV method, though it appears more sensitive to changes in τ. This suggests that while MH performs robustly under moderate sparsity, it becomes unstable when heterogeneity and data sparsity are high.

The RE model and the corresponding Hartung-Knapp-Sidik-Jonkman (HKSJ) adjustment display similar trends to the previous methods in terms of bias (see figure 8(c), 8(d)) and MSE (see figure 9(c),9(d)). However, the RE model consistently handles different τ values, producing more stable coverage (see figure 10(c)) across varying heterogeneity levels. Nevertheless, it tends over-coverage, consistent with findings from Langan et al. (2019) and Jackson et al. (2018), who showed that RE-based confidence intervals may be overly

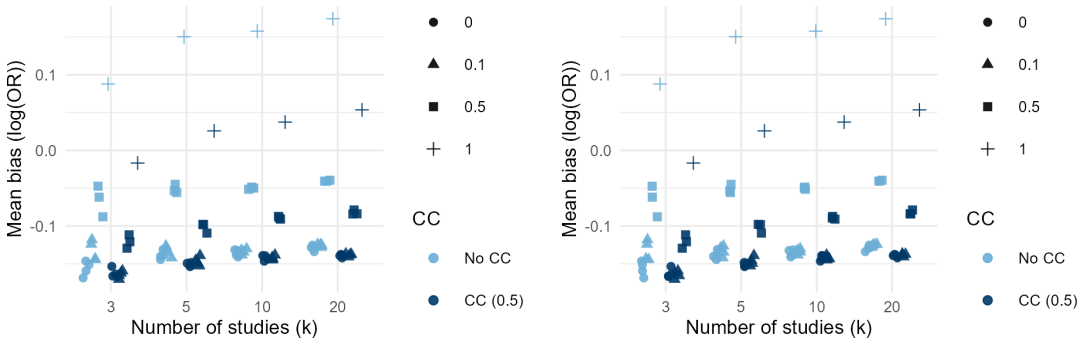
conservative in small-sample or sparse-data settings. In contrast, the HKSJ-adjusted version produces slight undercoverage, in line with Partlett and Riley (2017) and Mathes and Kuss (2018), who reported that the HKSJ adjustment can be overly liberal under extreme data sparsity, particularly when the number of studies is very small.
The results observed here align with the theoretical expectations outlined in Section 3. As anticipated, both the IV and MH estimators perform poorly in the presence of many zero-event studies due to their reliance on normal approximations and the instability of variance estimates when event counts are low (Sweeting et al., 2004; Bradburn et al., 2007). The bias reduction achieved through continuity correction is consistent with earlier work by Sweeting et al. (2004), who found that a fixed 0.5 correction can counteract numerical instability but often at the expense of interpretability. The strong dependence of bias and MSE on τ reflects the expected sensitivity of two-stage estimators to between-study heterogeneity. The RE estimator behaves largely as predicted by the random-effects theory (Section 3.3), offering better control of heterogeneity but showing inflated coverage under extreme sparsity, a phenomenon also discussed by Langan et al. (2019). Finally, the tendency of the HKSJ adjustment to yield under-coverage under such sparse conditions confirms previous simulation findings that this approach, while robust for small k, may not adequately compensate for high zero-cell proportions (Mathes and Kuss, 2018; Partlett and Riley, 2017).
Overall, the many zeros scenario confirms that sparse-data situations amplify the weaknesses of fixed-effect estimators and continuity corrections, while random-effects methods provide more reliable inference at the cost of overconservative coverage. The HKSJ adjustment mitigates some of these limitations but can itself become unstable when event rates are extremely low.

4.2.3 Randomisation 1:2 Scenario

The randomisation 1:2 scenario represents an unbalanced allocation setting where the control arm includes twice as many participants as the treatment arm. This configuration is particularly relevant in rare-event contexts, where unequal group sizes can amplify estimation bias and distort variance estimation, as shown in previous simulation studies by Sweeting et al. (2004) and Mathes and Kuss (2018).
For the IV method (see figure 11(a)), the bias is higher in absolute terms compared with the standard scenario, ranging from approximately -0.3 for lower τ values to around -0.1 for higher τ. Hence, the IV estimator consistently underestimates the true effect size and rarely crosses into overestimation. The application of a continuity correction (CC = 0.5) again pulls the bias towards zero, although this corrective effect weakens with increasing baseline variance (σ_{logit}). In general, larger baseline variance increases the bias magnitude, and higher τ values also contribute to more pronounced underestimation. The MSE (see figure 12(a)) decreases with both higher σ_{logit} and larger k, while it increases with τ. The

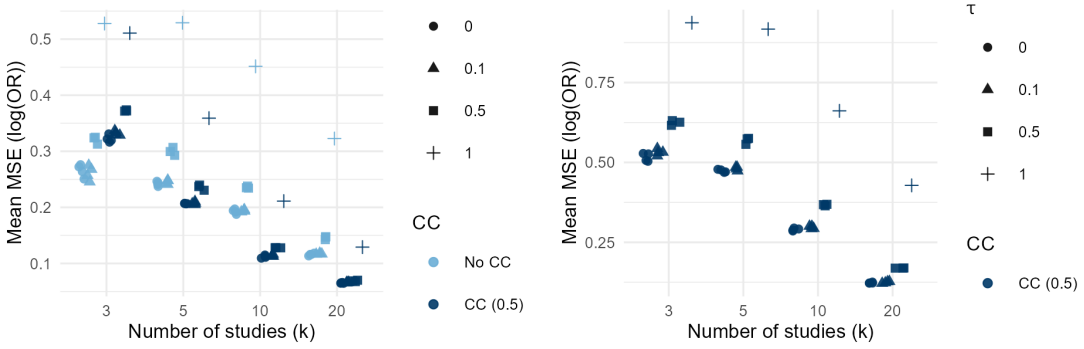


(a) IV-Estimator (b) MH-Estimator

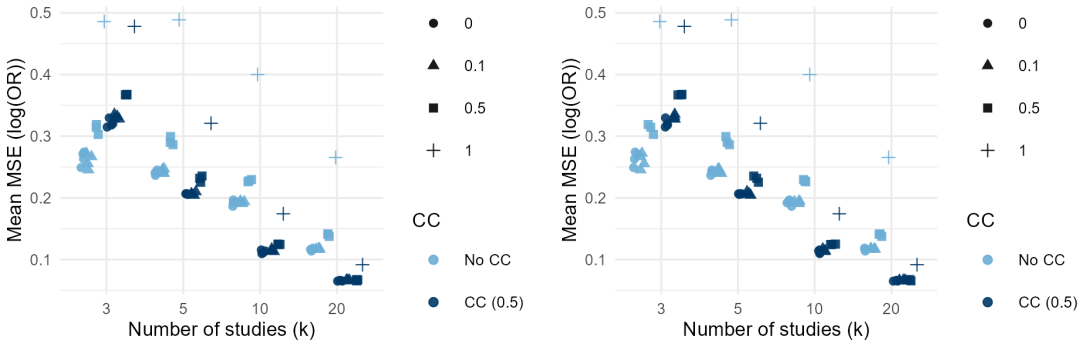


(c) RE Model (d) HKSJ-Adjustment

Figure 8: Mean Bias (log-OR) vs. number of studies for the many zeros scenario

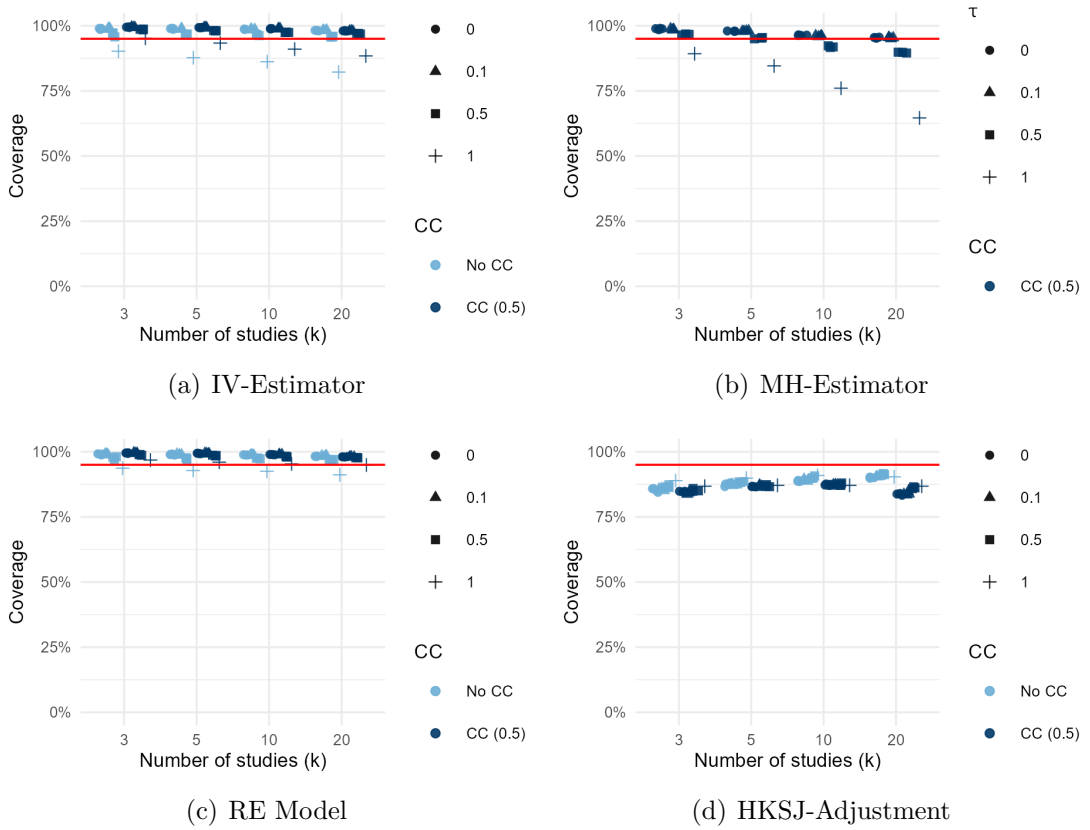


(a) IV-Estimator (b) MH-Estimator



(c) RE Model (d) HKSJ-Adjustment

Figure 9: Mean MSE (log-OR) vs. number of studies for the many zeros scenario



(a) IV-Estimator (b) MH-Estimator

(c) RE Model (d) HKSJ-Adjustment

Figure 10: Mean Coverage vs. number of studies for the many zeros scenario

continuity-corrected model (CC = 0.5) produces consistently lower MSE values than the uncorrected IV model. Coverage (see figure 13(a)) remains generally good but declines for larger τ and k, resulting in undercoverage particularly for $k = 10$ and $k = 20$.

The MH estimator shows the expected opposite behaviour, with a strictly positive bias (see figure 11(b)) across all parameter combinations, indicating systematic overestimating the true effect size. The magnitude of overestimation increases with the number of studies (k) but decreases with higher baseline variance (σ_{logit}). Larger τ values lead to stronger bias. The MSE (see figure 12(b)) of the MH estimator follows the same general pattern as the IV method—decreasing with higher σ_{logit} and larger k, but increasing with τ. The coverage performance of the MH model (see figure 13(b)) is somewhat better than that of the IV method for $\tau < 0.5$, but still declines as τ and k increase, showing undercoverage in large, heterogeneous meta-analyses.

The RE model estimator consistently underestimates the true effect size, but the bias (see figure 11(c)) magnitude diminishes (moves closer to zero) as σ_{logit} and τ increase. The influence of the continuity correction decreases with increasing σ_{logit}, causing the corrected and uncorrected models to converge in bias. This same pattern is observed for the HKSJ adjustment (see figure 11(d)), where the bias difference between CC and no-CC versions becomes negligible for higher σ_{logit} values. The MSE (see figure 12(c)) decreases with higher σ_{logit} and k, while CC = 0.5 again provides the lowest MSE. Increasing τ inflates

the MSE, but this effect becomes less pronounced as k increases. The HKSJ-adjusted model follows the same trend (see figure 12(d)), indicating that both random-effects approaches benefit from larger study numbers regarding estimation stability. Regarding coverage, the RE model (see figure 13(c)) estimator remains comparatively robust under increasing τ, but coverage collectively decreases for higher k, especially at low σ_{logit}. This decline is even more pronounced for the HKSJ adjustment (see figure 13(d)), which shows slight undercoverage under these extreme settings.

The observed results are consistent with the theoretical expectations outlined in Section 3, particularly concerning the sensitivity of two-stage estimators to imbalanced randomisation. As predicted, the IV method performs poorly when treatment and control group sizes differ substantially, leading to systematic underestimation of the pooled effect due to the biased weighting of smaller treatment arms. This finding aligns closely with the empirical results of Sweeting et al. (2004) and Bradburn et al. (2007), who reported that continuity-corrected IV estimators underestimate treatment effects in unbalanced sparse-data settings. The MH estimator, by contrast, exhibits positive bias under imbalance, consistent with Piaget-Rossel and Taffé (2019), who observed overestimation when control arms dominate the study population.

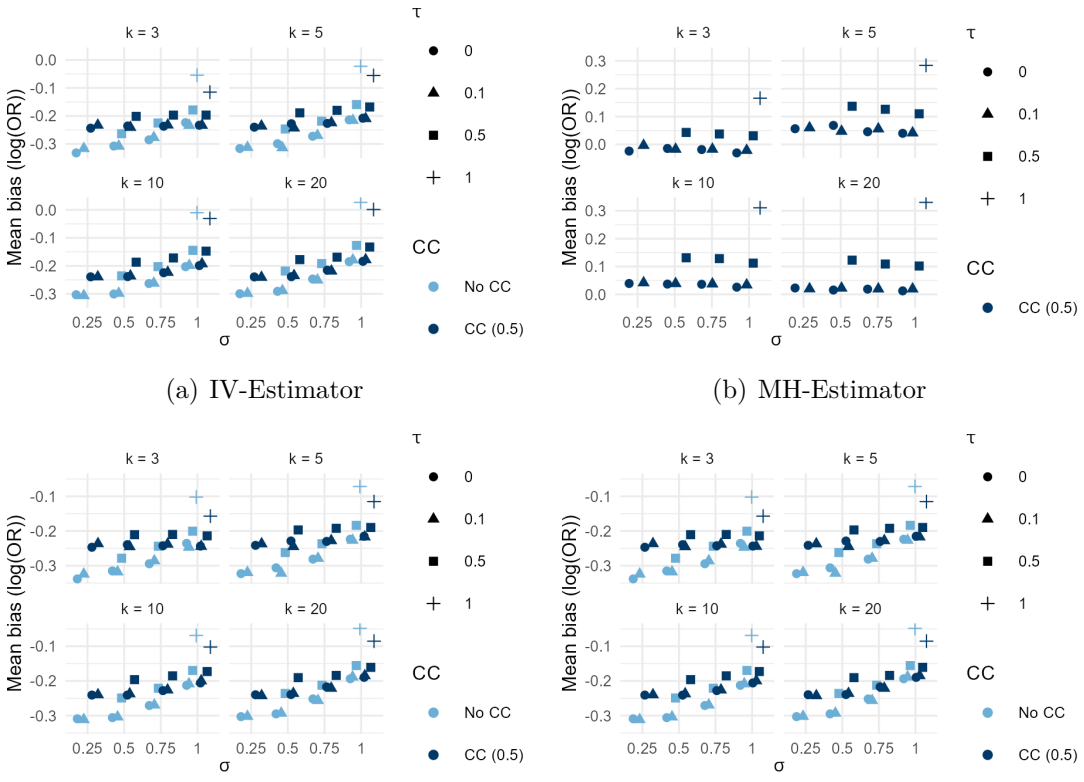
The improved stability of the RE model estimator, particularly under increased τ, reflects its theoretical advantage as a variance-based random-effects model (Section 3.3). Given the reduced influence of large imbalances under variance-based weighting, its slight underestimation of the effect size is expected. The observed over-coverage of the RE model and under-coverage of the HKSJ adjustment confirm findings from Partlett and Riley (2017) and Mathes and Kuss (2018), who showed that while HKSJ improves small-sample coverage, it may over-correct in sparse or highly imbalanced data.

Overall, the randomisation 1:2 scenario confirms that unequal group sizes exacerbate the weaknesses of fixed-effect estimators, while the RE Model provides a more balanced bias-coverage trade-off. However, both RE Model and HKSJ adjustment remain sensitive to increasing study numbers and baseline variability, underscoring the need for caution when analysing unbalanced rare-event data.

4.2.4 Randomisation 1:3 Scenario

The randomisation 1:3 scenario represents a setting with pronounced imbalance between treatment and control arms, where the control group contains three times as many participants as the treatment group. Such strong allocation asymmetry is common in clinical research with rare events, where ethical or logistical constraints lead to a preference for larger control groups. However, this imbalance is expected to increase bias and reduce coverage for classical two-stage estimators, as highlighted by Sweeting et al. (2004) and Bradburn et al. (2007).

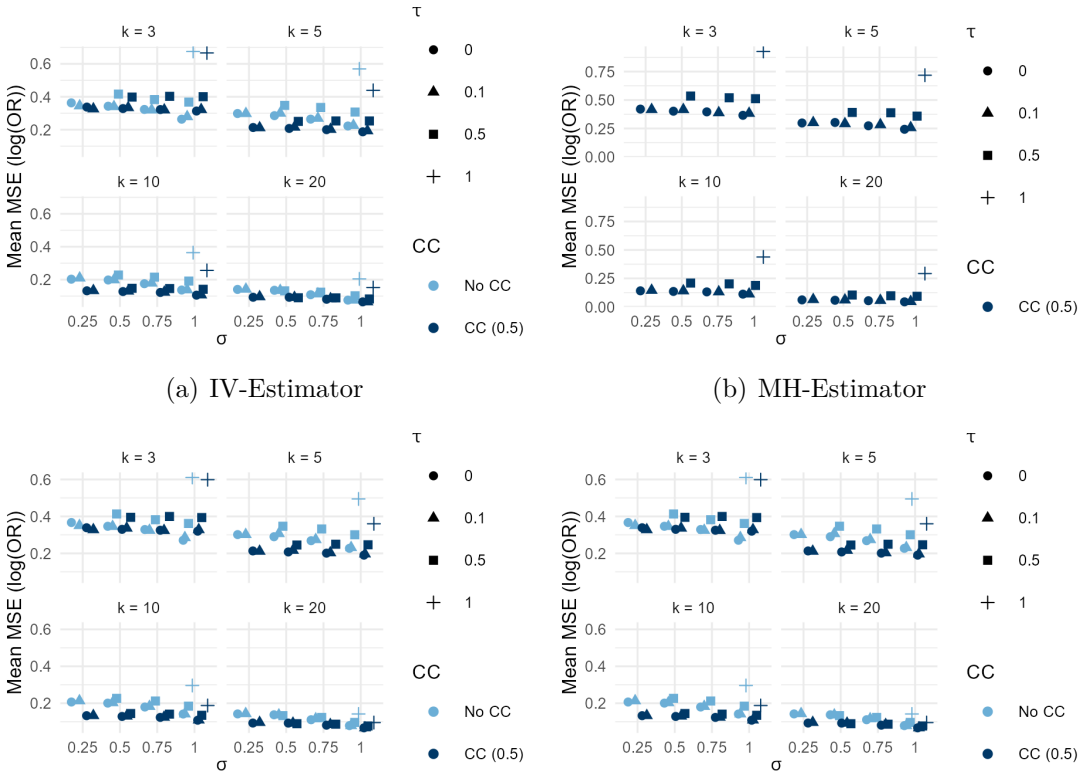
For the IV method, the bias (see figure 14(a)) is considerably larger in absolute terms than



(a) IV-Estimator (b) MH-Estimator

(c) RE Model (d) HKSJ-Adjustment

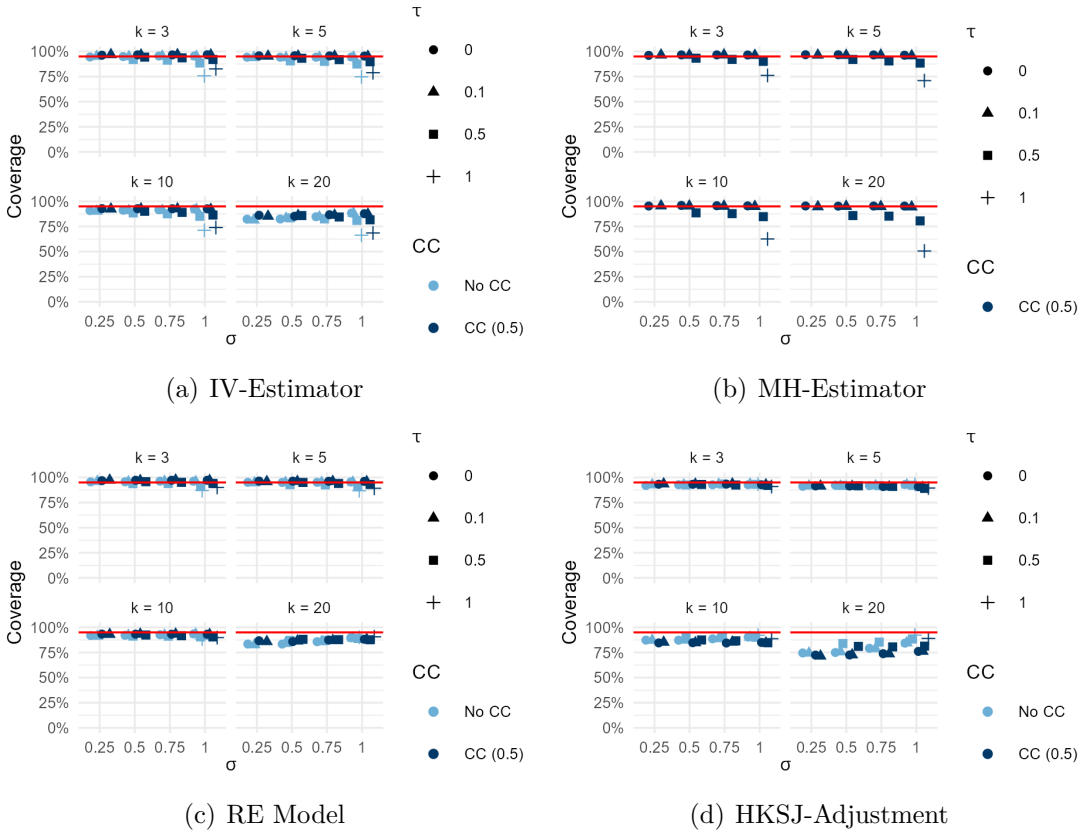
Figure 11: Mean Bias (log-OR) vs. σ_{logit} for the randomisation 1:2 scenario



(a) IV-Estimator (b) MH-Estimator

(c) RE Model (d) HKSJ-Adjustment

Figure 12: Mean MSE (log-OR) vs. σ_{logit} for the randomisation 1:2 scenario



(a) IV-Estimator

(b) MH-Estimator

(c) RE Model

(d) HKSJ-Adjustment

Figure 13: Mean Coverage vs. σ_{logit} for the randomisation 1:2 scenario

in the randomisation 1:2 scenario, with the IV estimator strictly underestimating the true treatment effect across all conditions. The bias decreases in magnitude (moves towards zero) with increasing baseline variance (σ_{logit}) and heterogeneity (τ), while the difference between the continuity-corrected (CC = 0.5) and uncorrected models becomes smaller as σ_{logit} increases. The MSE (see figure 15(a)) is also higher than in the randomisation 1:2 scenario but decreases substantially with increasing σ_{logit}. The continuity-corrected version yields consistently lower MSE than the uncorrected model. MSE increases with larger τ but decreases with higher numbers of studies (k), confirming that the estimator benefits from more data despite persistent underestimation. The coverage pattern under the IV model (see figure 16(a)) is similar to that in the previous scenario. Still, the overall under-coverage is more severe, particularly for higher k and τ, indicating that imbalanced allocation exacerbates the loss of nominal coverage in large meta-analyses.

For the MH estimator, the bias (see figure 14(b)) remains strictly positive, again reflecting systematic overestimation of the true effect size. As in the IV model, bias increases with higher τ and decreases with higher σ_{logit} and k. The MSE pattern (see figure 15(b)) mirrors that of the IV method, showing a decrease with larger σ_{logit} and k, but an increase with higher τ. The coverage behaviour (see figure 16(b)) is also consistent with the randomisation 1:2 scenario, showing adequate performance for low τ values but substantial under-coverage when τ or k increase. Overall, both fixed-effect estimators confirm the

expected sensitivity to imbalance and heterogeneity.

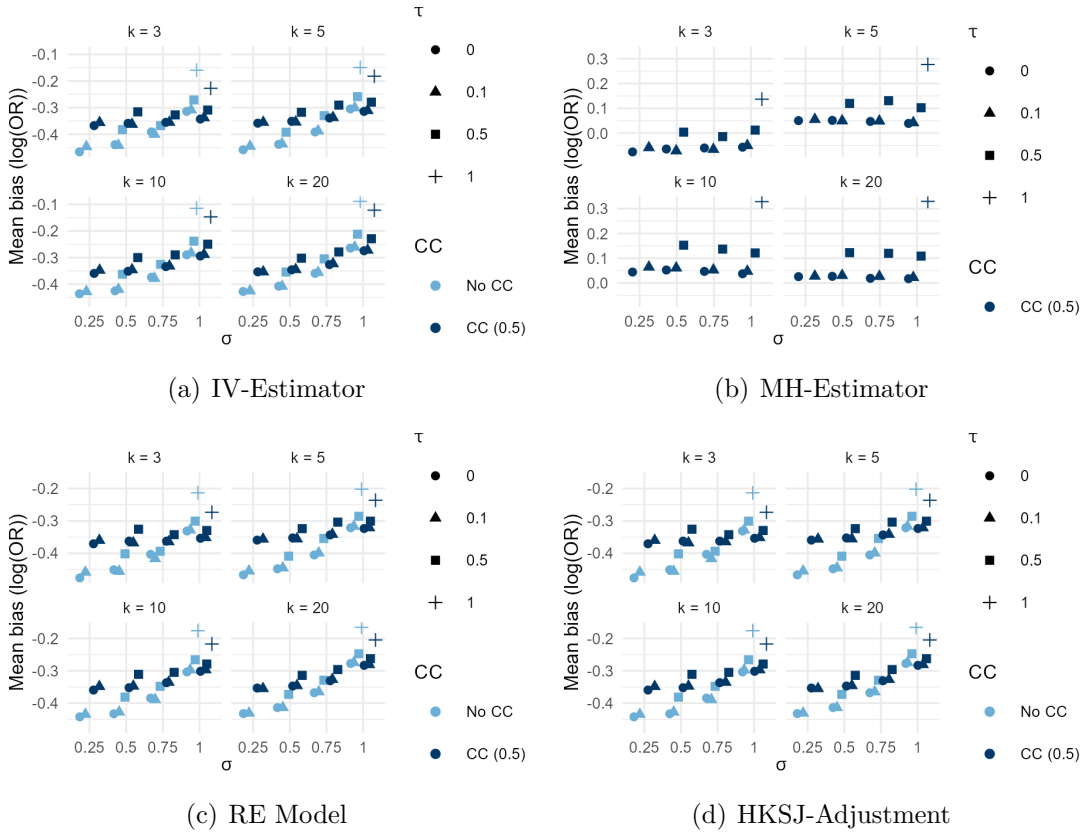
The RE model estimator displays a higher absolute bias (see figure 14(c)) than the previous randomisation scenario, indicating stronger underestimation of the true effect. The qualitative pattern of the bias remains the same: bias decreases (moves towards zero) with increasing σ_{logit} and τ, and the effect of the continuity correction diminishes as σ_{logit} increases, leading to convergence between the corrected and uncorrected versions. The MSE (see figure 15(c)) behaves similarly to the previous scenario, decreasing with increasing σ_{logit} and k, while τ leads to higher MSE, though this effect weakens as k grows. Coverage under the RE model (see figure 16(c)) estimator follows the same trend. Still, it exhibits slightly stronger under-coverage for large k, particularly when σ_{logit} is small or heterogeneity is large.

Applying the HKSJ adjustment leaves the bias (see figure 14(d)) and MSE (see figure 15(d)) virtually unchanged compared to the classical RE model, but coverage (see figure 16(d)) declines markedly for higher k. This drop in coverage is more pronounced than in the previous scenario, suggesting that the adjustment, while beneficial in small meta-analyses, becomes overly liberal when sample size and heterogeneity increase under strong randomisation imbalance.

The observed results align with the theoretical considerations presented in Section 3 and confirm empirical findings from previous simulation studies. As anticipated, the IV estimator performs poorly under severe imbalance, systematically underestimating the treatment effect due to the disproportionate weighting of smaller treatment arms and the instability of variance estimates in sparse-event settings (Sweeting et al., 2004; Bradburn et al., 2007). The continuity correction attenuates this bias but does not fully eliminate it, especially when σ_{logit} and τ are large. The MH estimator's consistent positive bias corresponds to previous findings by Piaget-Rossel and Taffé (2019), who reported overestimation in unbalanced datasets where control arms dominate.

The RE model estimator shows better control over heterogeneity and lower MSE, as expected from the random-effects formulation (Section 3.3). However, its tendency to underestimate the effect under strong imbalance highlights the limits of variance-based pooling in sparse data situations, confirming the conclusions of Langan et al. (2019) and Jackson et al. (2018). The observed over-coverage of the RE Model and pronounced under-coverage of the HKSJ-adjusted model mirror earlier findings by Partlett and Riley (2017) and Mathes and Kuss (2018), who showed that the HKSJ method, although beneficial for small k, can produce overly narrow intervals when study numbers increase or when heterogeneity is high.

Overall, the randomisation 1:3 scenario amplifies the weaknesses already visible under 1:2 imbalance: both IV and MH estimators suffer from directional bias, while the RE model offers a more stable but slightly conservative inference. The HKSJ adjustment, although effective in small-sample conditions, becomes unreliable for larger meta-analyses under



(a) IV-Estimator

(b) MH-Estimator

(c) RE Model

(d) HKSJ-Adjustment

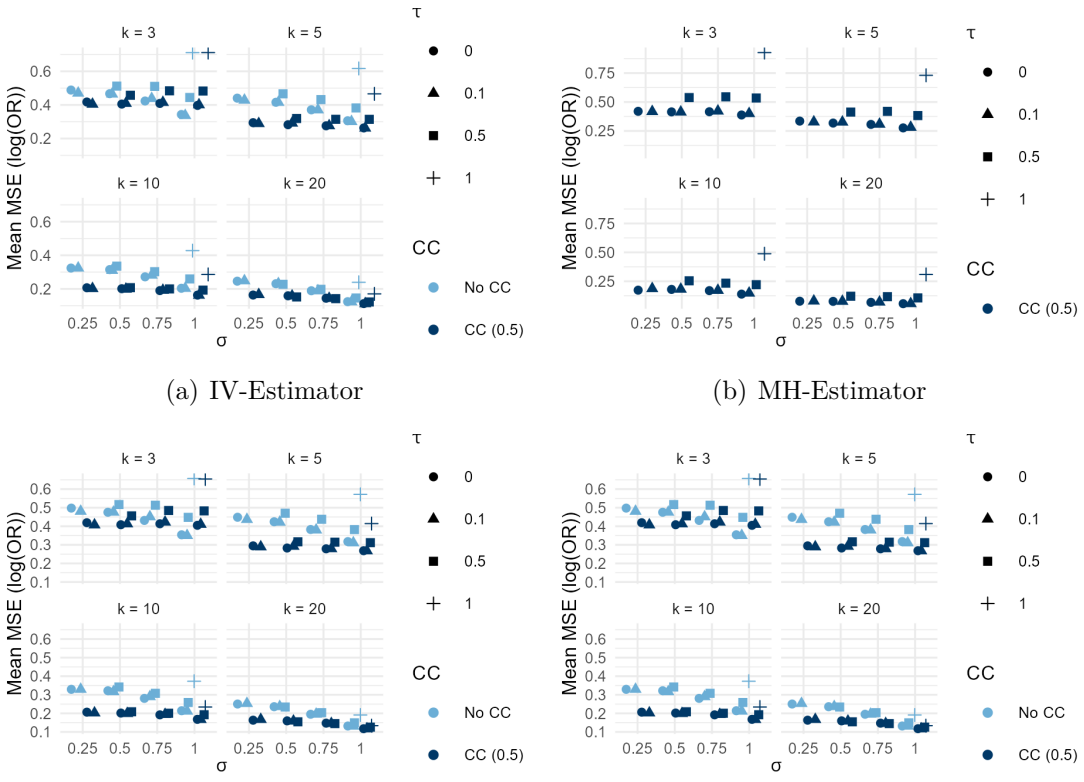
Figure 14: Mean Bias (log-OR) vs. σ_{logit} for the randomisation 1:3 scenario

extreme randomisation imbalance.

4.2.5 Negative Correlation Scenario

The negative correlation scenario explores how smaller studies exhibit higher baseline risks, while larger studies show lower baseline risks. This configuration introduces a dependence structure between study size and baseline event probability.

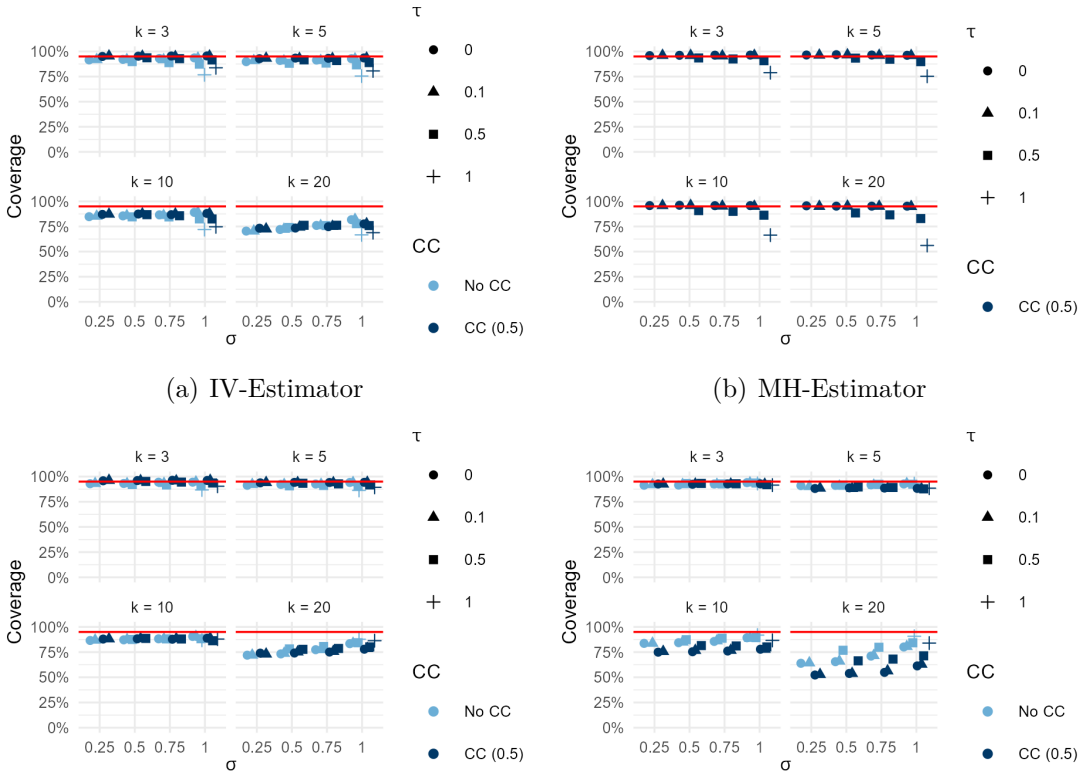
For the IV method, the bias (see figure 17(a)) ranges from approximately -0.2 for low τ to around 0.1 for high τ values. With increasing baseline variance (σ_{logit}) and between-study heterogeneity (τ), the bias moves towards zero, while for $\tau = 1$ the model begins to overestimate the true effect size. A continuity correction (CC = 0.5) effectively pulls the bias closer to zero across all parameter settings. Differences across the number of studies (k) are negligible, indicating that additional information from more studies does not improve the bias when the baselines are correlated. The MSE (see figure 18(a)) slightly increases with σ_{logit}, especially for small k, and grows with higher τ, while larger k values reduce the MSE. For small meta-analyses, the continuity-corrected estimator shows slightly higher MSE than the uncorrected version, but this relationship inverts for larger k, where the CC = 0.5 version performs better. The coverage of the IV estimator (see figure 19(a)) remains generally good but decreases with increasing τ, leading to mild under-coverage



(a) IV-Estimator (b) MH-Estimator

(c) RE Model (d) HKSJ-Adjustment

Figure 15: Mean MSE (log-OR) vs. σ_{logit} for the randomisation 1:3 scenario



(a) IV-Estimator (b) MH-Estimator

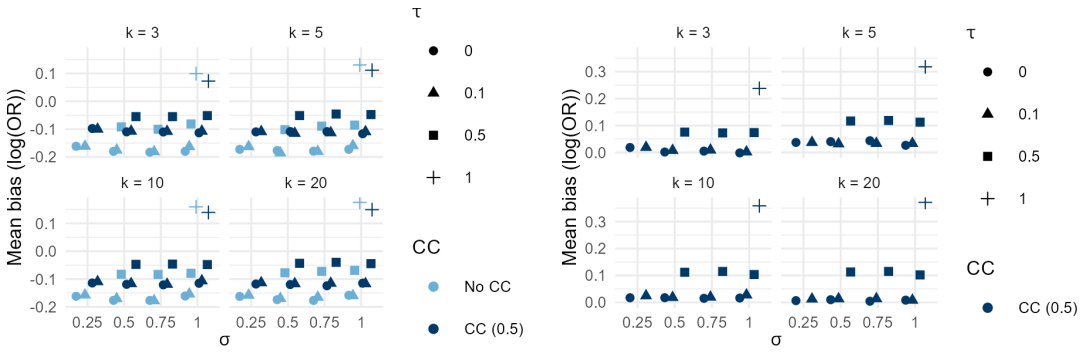
(c) RE Model (d) HKSJ-Adjustment

Figure 16: Mean Coverage vs. σ_{logit} for the randomisation 1:3 scenario

at high heterogeneity levels.
The MH estimator again shows a strictly positive bias (see figure 17(b)), indicating overestimation of the true effect size. While the bias remains near zero for smaller τ, it increases substantially for $\tau = 1$. The bias appears largely unaffected by σ_{logit} or k. The MSE for the MH estimator (see figure 18(b)) is higher than that of the IV model for smaller meta-analyses but converges to similar values as k increases. The overall MSE pattern follows the same direction as the IV model—rising with τ and decreasing with σ_{logit} and k. The coverage behaviour (see figure 19(b)) mirrors the IV method's. However, under-coverage begins to appear at $\tau = 0.5$, suggesting that the MH estimator is slightly more sensitive to heterogeneity in the presence of negative correlation between study size and baseline risk.
The RE model and HKSJ adjustment estimators display a similar bias (see figure 17(c), 17(d)) pattern to the IV model but show less variability across τ values, indicating better control of heterogeneity. Both estimators keep the bias close to zero across all σ_{logit} and k levels. The MSE of the RE Model and HKSJ adjustment (see figure 18(c), 18(d)) is nearly identical to that of the IV estimator, suggesting comparable efficiency. In terms of coverage, the RE Model (see figure 19(c)) estimator performs better than both the IV and MH models, eliminating the under-coverage at higher τ values that was observed in the previous methods. However, the HKSJ adjustment (see figure 19(d)) shows collective under-coverage for large k, consistent with earlier observations in other scenarios where the adjustment becomes overly liberal as the number of studies increases.

4.2.6 Positive Correlation Scenario

The positive correlation scenario investigates the case where larger studies are associated with higher baseline event probabilities, while smaller studies tend to have lower baseline risks. This configuration reverses the dependency structure of the negative correlation scenario and represents situations where study size may act as a proxy for higher event exposure or broader inclusion criteria.
For the IV estimator, the bias (see figure 20(a)) ranges between approximately -0.1 for smaller τ and 0.1 for larger τ values, representing a smaller variance in the bias compared with the negative correlation scenario. The difference between the continuity-corrected (CC $= 0.5$) and uncorrected estimates is smaller, and the corrective effect of the CC pulling estimates towards zero is less pronounced. As σ_{logit} and τ increase, the bias moves closer to zero, and for $\tau = 1$, the estimator slightly overestimates the true effect size. The influence of the number of studies (k) remains negligible. The MSE (see figure 21(a)) behaves similarly to the previous scenario—slightly increasing with τ and decreasing with both σ_{logit} and k. However, in contrast to the negative correlation case, the coverage of the IV estimator (see figure 22(a)) is worse overall, as under-coverage now occurs already at $\tau = 0.5$ instead of $\tau = 1$, and this effect becomes more pronounced as k increases.



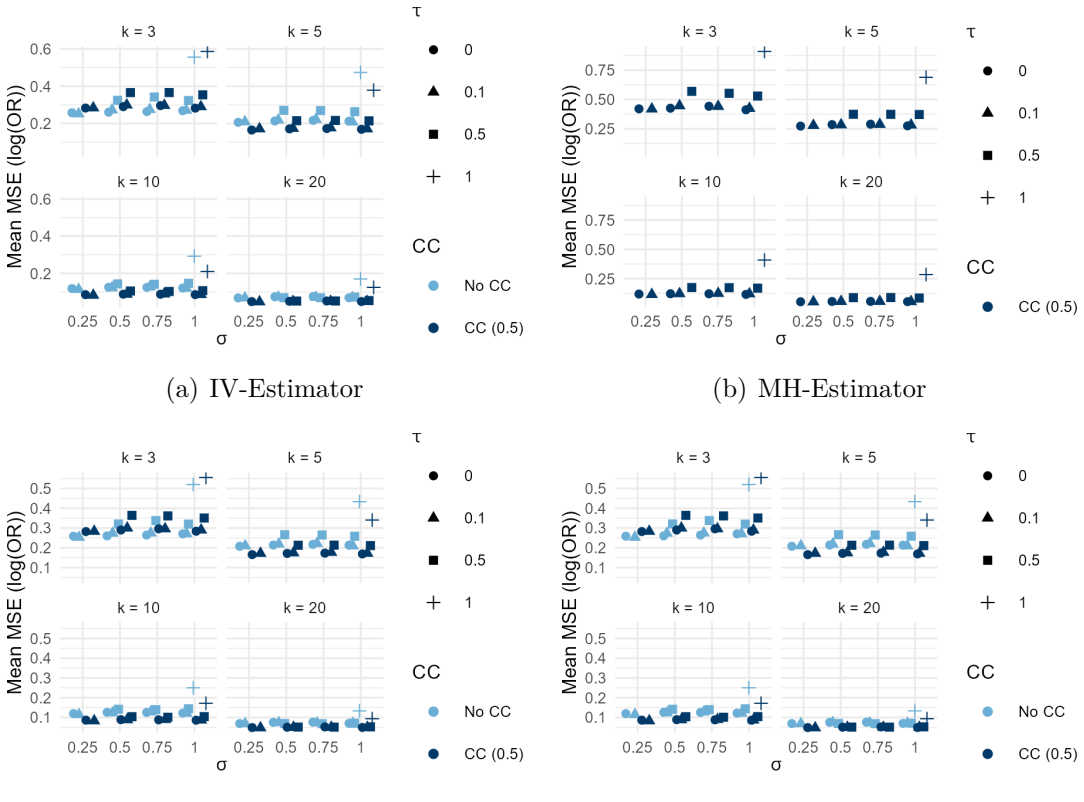
(a) IV-Estimator

(b) MH-Estimator

(c) RE Model

(d) HKSJ-Adjustment

Figure 17: Mean Bias (log-OR) vs. σ_{logit} for the negative correlation scenario



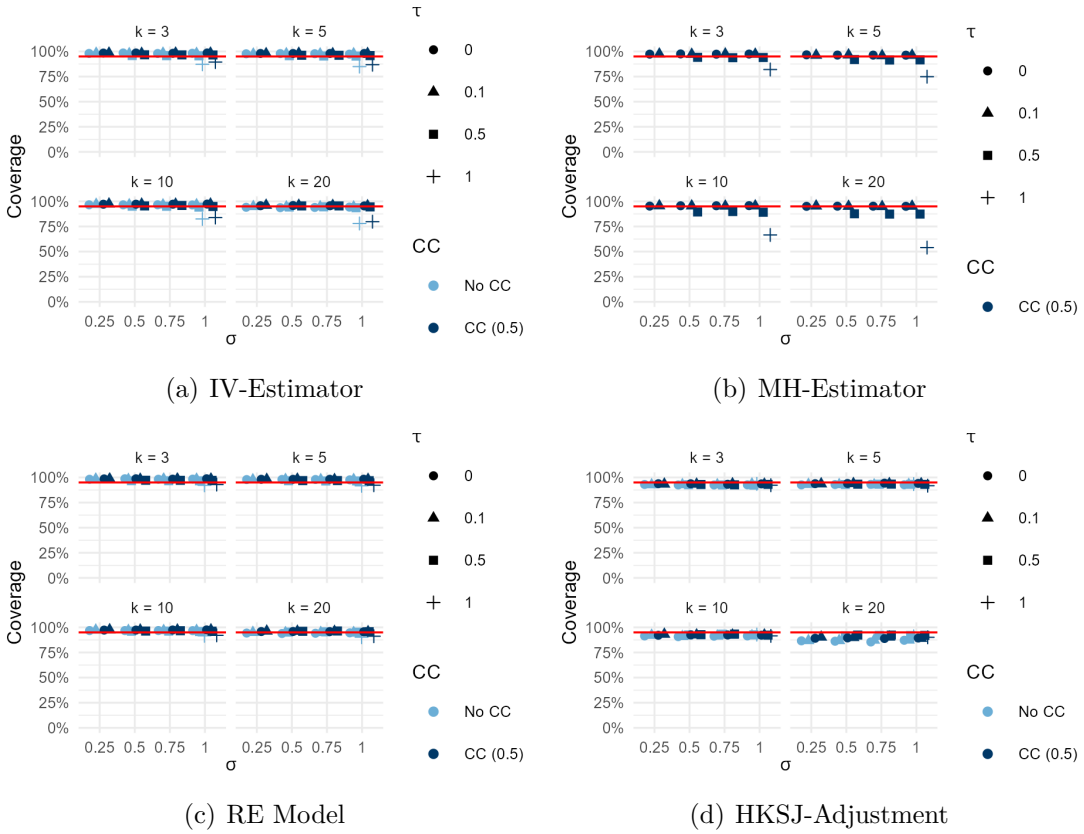
(a) IV-Estimator

(b) MH-Estimator

(c) RE Model

(d) HKSJ-Adjustment

Figure 18: Mean MSE (log-OR) vs. σ_{logit} for the negative correlation scenario



(a) IV-Estimator

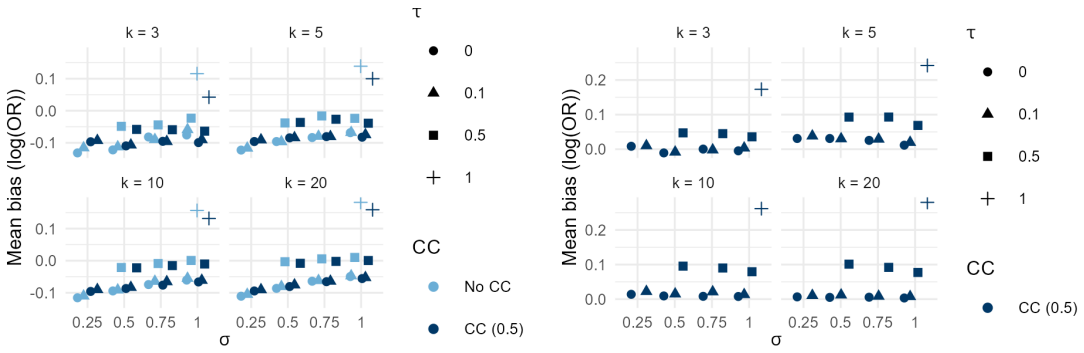
(b) MH-Estimator

(c) RE Model

(d) HKSJ-Adjustment

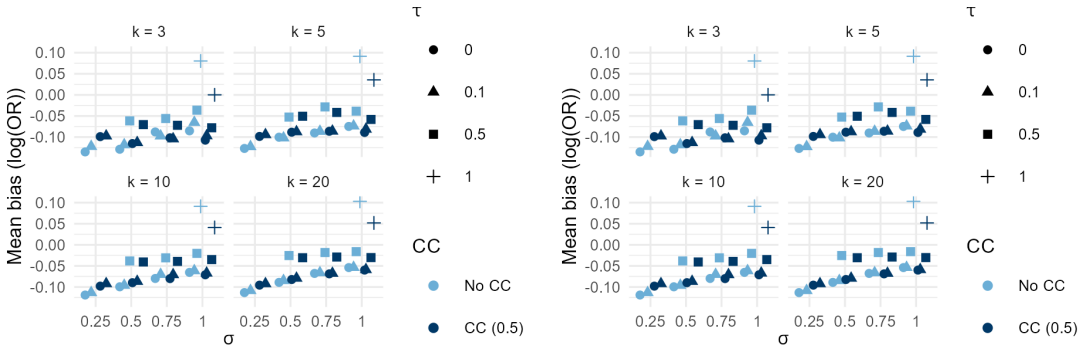
Figure 19: Mean Coverage vs. σ_{logit} for the negative correlation scenario

The MH estimator behaves analogously to the IV estimator in terms of bias (see figure 20(b)) and MSE (see figure 21(b)). The bias remains positive across all conditions and increases with higher τ, decreasing with larger σ_{logit} and k. The MSE follows the same directional pattern as in the IV model, and the coverage (see figure 22(b)) exhibits the same trend of earlier onset under-coverage with increasing τ and k. Compared with the negative correlation scenario, the MH model displays similar magnitudes of bias and MSE but overall slightly poorer coverage, indicating that positive correlation between baseline risk and study size amplifies the effect of heterogeneity on confidence interval performance.
The RE model estimator and its HKSJ adjustment exhibit less bias (see figure 20(c), 20(d)) and smaller variability in estimates than in the negative correlation scenario, suggesting better control of correlation-induced heterogeneity. The bias remains within a narrow range across σ_{logit} and k, and the overall pattern mirrors the IV estimator's. The MSE (see figure 21(c), 21(d)) also behaves similarly to the IV estimator but shows improved robustness to increasing τ, particularly for larger k, indicating that the RE model benefits more from additional studies than the fixed-effect estimators. The coverage of the RE model (see figure 22(c)) is again more stable under increasing τ compared with the IV and MH estimators. However, the improvement is less pronounced than in the negative correlation scenario. The HKSJ adjustment maintains similar bias and MSE levels but shows slight under-coverage (see figure 22(d)), especially for large k values.



(a) IV-Estimator

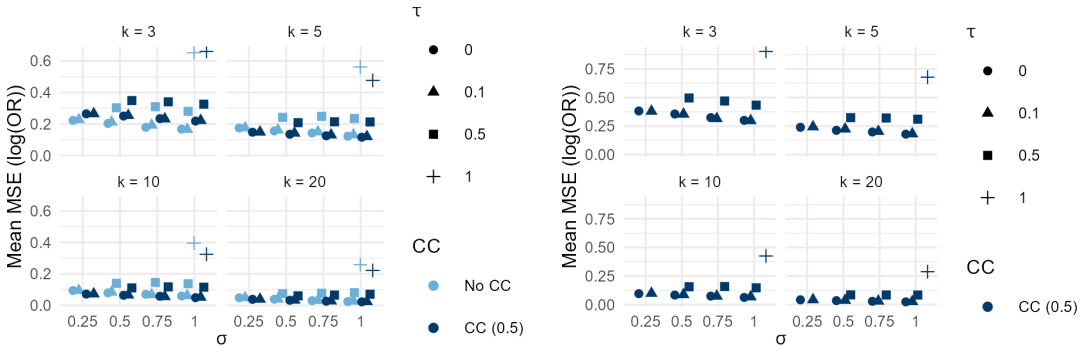
(b) MH-Estimator



(c) RE Model

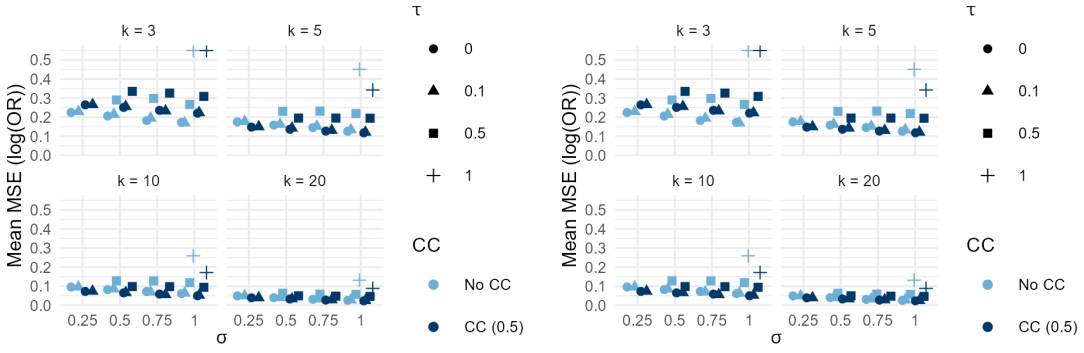
(d) HKSJ-Adjustment

Figure 20: Mean Bias (log-OR) vs. σ_{logit} for the positive correlation scenario



(a) IV-Estimator

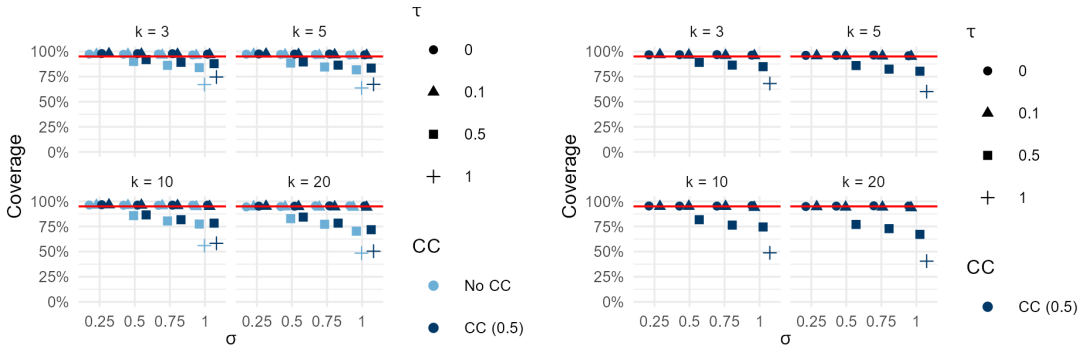
(b) MH-Estimator



(c) RE Model

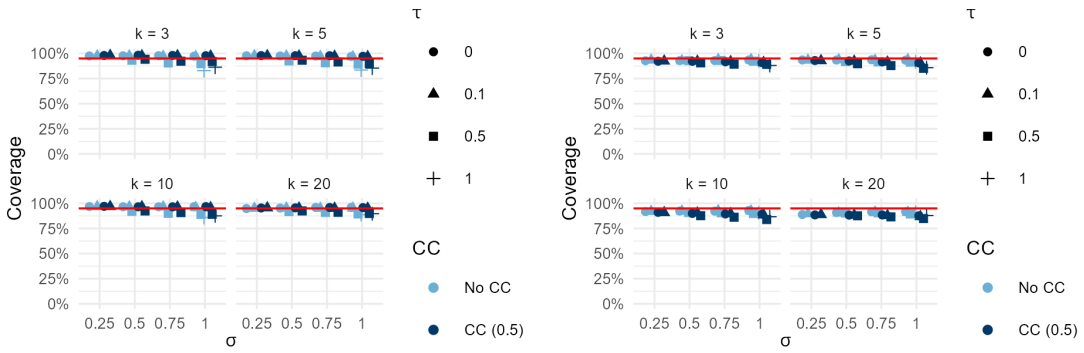
(d) HKSJ-Adjustment

Figure 21: Mean MSE (log-OR) vs. σ_{logit} for the positive correlation scenario



(a) IV-Estimator

(b) MH-Estimator



(c) RE Model

(d) HKSJ-Adjustment

Figure 22: Mean Coverage vs. σ_{logit} for the positive correlation scenario

4.3 Implications

The results across all scenarios provide several important implications for researchers conducting meta-analyses of rare events or datasets characterised by small study numbers, heterogeneity, or imbalance. Overall, the findings emphasise that the choice of pooling method and the treatment of sparse and correlated data structures can substantially influence both the magnitude and reliability of pooled estimates.

The RE model estimator generally demonstrated the most stable performance across all investigated settings. It consistently provided a balanced trade-off between bias, efficiency, and coverage, representing the most reliable default option for rare-event meta-analyses. The IV and MH estimators exhibited systematic, directionally opposing biases. The IV method underestimated the pooled effect under most conditions, particularly when data were unbalanced or sparse. At the same time, the MH method tended to overestimate the effect in the same situations. These results confirm earlier findings by Sweeting et al. (2004) and Bradburn et al. (2007), who reported that two-stage estimators relying on large-sample approximations perform poorly when event counts are low. Applying a fixed continuity correction, such as adding 0.5 to each cell, reduced bias and mean squared error in many situations, particularly when zero-event studies were present. However, the correction also introduced distortions in several scenarios and its influence varied

with heterogeneity and allocation imbalance. Therefore, continuity corrections should be applied with caution.

The simulations also confirmed that increasing heterogeneity (τ) amplifies both bias and under-coverage for fixed-effect estimators. At the same time, baseline variability (σ_{logit}) only marginally improves the bias of IV models without fully resolving coverage problems. Practitioners should thus quantify heterogeneity carefully and avoid relying on fixed-effect estimators when between-study variation is substantial. In such cases, random-effects estimators provide more credible inference, though they still require careful interpretation when k is large.

The number of included studies (k) is central in determining estimator performance. In this simulation study, we have seen that a higher heterogeneity paired with more studies ($k = 20$) can lead to overly conservative confidence intervals, which can be corrected by using a random effects model for inference. Also, it is important to note that in the case of small k, e.g. $k = 3$, even the random-effects models can have issues mitigating the effect a high heterogeneity has on the confidence intervals (see figure 22(c)). Sparse-data situations, such as the many-zeros scenario, presented a particular challenge. IV and MH estimators displayed strong bias and unstable variance estimations when many studies contained no events. In contrast, the RE model provided more reliable inference, but with slightly inflated coverage. The HKSJ-adjusted variant partially mitigated this overcoverage but tended to produce overly narrow intervals when sparsity was extreme. Researchers confronted with many zero-event studies should rely on RE models as the main analysis method. Reporting results with and without continuity correction may help illustrate the sensitivity of estimates to sparse-data handling.

Scenarios with unbalanced randomisation, such as 1:2 or 1:3 allocation ratios, showed that unequal group sizes accentuate the weaknesses of fixed-effect estimators. The IV estimator consistently underestimated the effect, while MH consistently overestimated it. The RE model again demonstrated the most balanced behaviour, though underestimation increased slightly under severe imbalance. In such settings, RE models should be preferred as the primary analysis approach.. The HKSJ correction, though helpful for small meta-analyses, showed a tendency toward under-coverage in large, highly imbalanced datasets and should therefore be applied cautiously in these contexts.

The correlation scenarios highlighted an additional and often overlooked source of bias in meta-analysis. When baseline risk and study size were negatively correlated—meaning smaller studies had higher event rates—fixed-effect estimators showed systematic bias and early onset under-coverage in the case of the HKSJ adjustment, especially for $k = 20$. A positive correlation leads to early undercoverage in the case of small k and higher τ, even for the random effect models, showing that under positive baseline correlation, all models become less reliable in the presence of heterogeneity. The RE model was again less affected by such correlation, offering more consistent bias control and coverage stability,

though not eliminating the problem. Practitioners should therefore routinely explore the relationship between baseline risk and study size before analysis and interpret pooled estimates with caution when a strong correlation is detected.

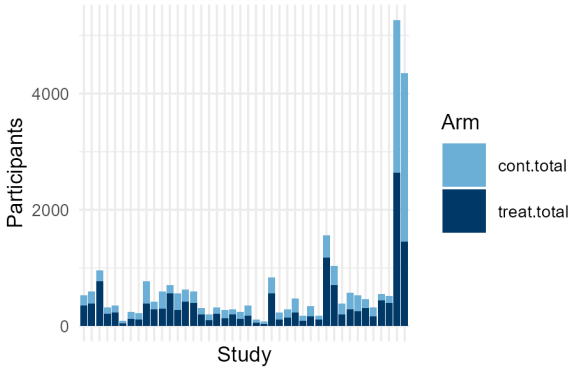
5 Application

For the application example, we are using the Rosiglitazone dataset (Nissen and Wolski, 2007), which is a frequently cited real-world example in meta-analytic research on rare adverse events. It compiles data from 42 randomised controlled trials evaluating rosiglitazone, a thiazolidinedione used for type 2 diabetes, compared with control treatments. The outcomes of interest are binary—occurrence of myocardial infarction and cardiovascular death—and the dataset is characterised by low event rates, unbalanced treatment-to-control ratios, and study-level heterogeneity. Because some trials contain few or no cardiovascular events, the dataset provides a realistic test case for assessing how different meta-analytic models handle sparse and zero-event data, making it a suitable empirical illustration for the methodological evaluation in the application section of this thesis (Rücker and Schumacher, 2008; Nissen and Wolski, 2007). For this application example, we will fit the models used in the simulation study to examine the estimated effect sizes for the risk of death and the risk of infarction separately. From the descriptive plots for both outcomes, we can see that the distribution of the study sizes reveals very heterogeneous study sizes with strong imbalance where the treatment groups are often more than double the control group (see figure 23(a)). The percentage of zero studies lies at almost 80% for the death events and at nearly 70% for the infarction events, while half of the zero studies in the death event case consist of double zero studies. The fraction of double-zero studies for the infarction events is much lower (see figure 23(b)). The risk distribution for the death events reveals a strongly right-skewed distribution of the risk in the control group with moderate variance in the baseline risk (see figure 23(c)). The treatment risk of the death events follows a similar pattern but is slightly less right-skewed and slightly more heterogeneous (see figure 23(c). Regarding the risk distribution for infarction events (see figure 23(d)), we can observe a strongly right-skewed control risk distribution with again moderate variance in the baseline. Still, the treatment risk distribution shows a considerably higher heterogeneity and less right-skewness. When looking at the relationship between the baseline risk and the study size, we cannot observe sufficient evidence for correlation for both event types (see figure 24(a), 24(b)).
Therefore, for synthesizing the finding regarding the death events we would choose the MH-estimator with 0.5 continuity correction since in cases of low heterogeneity and extreme data sparsity (see figure 8(b), 10(b)), as well as unbalanced randomisation it has proven to deliver consistent almost unbiased results while providing sufficient coverage (see figure 14(b), 16(b)). The resulting model fits for the death events (see figure 25(a)

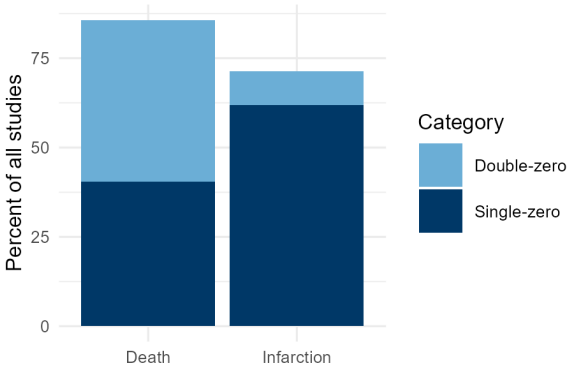
show the expected very narrow CIs for the HKSJ adjusted random effects model, especially with continuity correction. In contrast, the classic random effects model shows wider confidence intervals, indicating the expected possible over-coverage in the presence of imbalanced randomisation and sparse data. The MH estimator is the only one that is clear about the direction of the treatment effect size. All other CIs include values < 1 and > 1, indicating uncertainty about the direction of the treatment effect size. It is important to note here that we have shown that all models except the MH model tend to underestimate the true effect size, while the MH estimator has the opposite tendency. However, in case of extreme data sparsity and imbalanced randomisation, the bias was closest to zero for the MH estimator.
Thus, we would strongly recommend the MH estimator in this context since it has proven to produce reliable results in scenarios with extreme data sparsity, moderate heterogeneity and unbalanced randomisation.
In case of the infarction events, we would choose a different approach. Since we face higher heterogeneity and less data sparsity, but still unbalanced randomisation, choosing one of our models becomes quite difficult. The fixed effect models have been shown to perform worse under higher heterogeneity, leading to significant under-coverage in all scenarios. At the same time, for the imbalanced randomisation scenarios (especially in case of a 1:3 randomisation ratio, see figure 16(c)), the random effects models have shown significant under-coverage, while showing over-coverage in scenarios with extreme data sparsity (see figure 10(c)). Therefore, we recommend the random effects model with cc 0.5 since it has proven to be a robust all-around model that performs best for all scenarios. The fitted models show that the estimates and CIs are very similar across all models, while again the MH model shows its tendency to overestimate, and the HKSJ adjusted model provides overly conservative CIs.
Overall, this application example shows that modelling choice in the context of meta-analysis remains challenging, especially for scenarios with varying heterogeneity, data sparsity, and imbalanced randomisation. The combination of extreme data sparsity, high heterogeneity, and imbalanced randomisation remains an important subject for further research.

6 Discussion

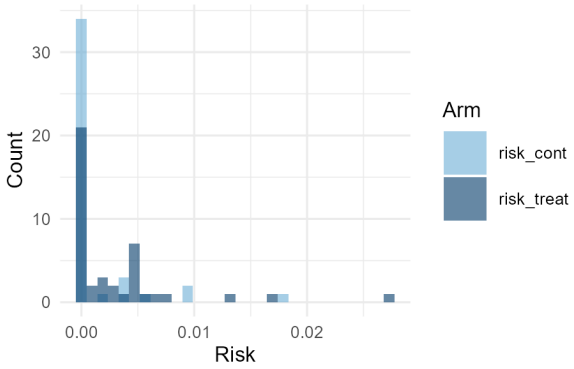
In this thesis, we examined how different variants of baseline pooling in classical meta-analysis models influence model performance under varying baseline risk variability, distributional form, and data imbalance. Specifically, we assessed how the IV method, MH estimator, RE models, and HKSJ estimators perform when confronted with different baseline-risk variances, heterogeneity levels of the true effects, imbalanced randomisation, small numbers of studies, and rare event probabilities. By incorporating both normally



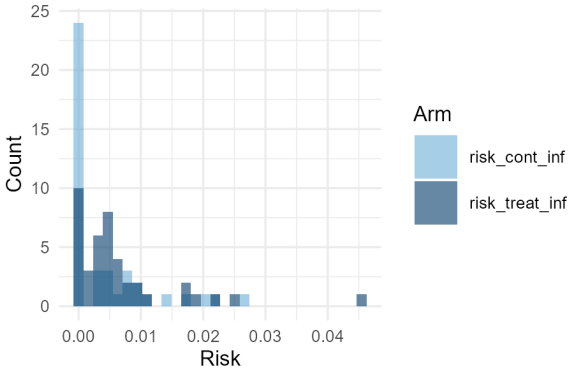
(a) Distribution and partition of the study size



(b) Percentage and partition of zero studies

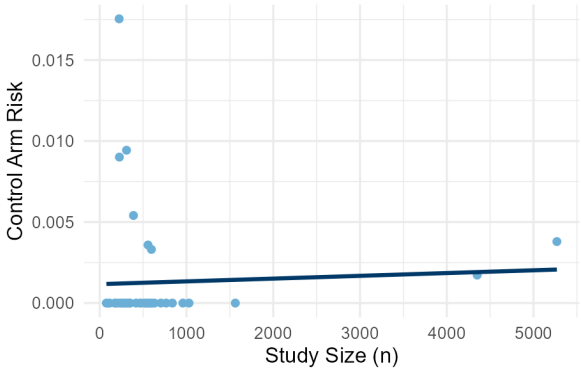


(c) Death

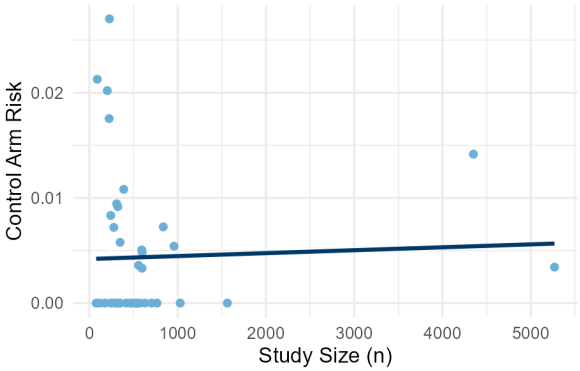


(d) Infarction

Figure 23: Risk distribution in control and treatment arm



(a) Death



(b) Infarction

Figure 24: Study size vs. baseline risk

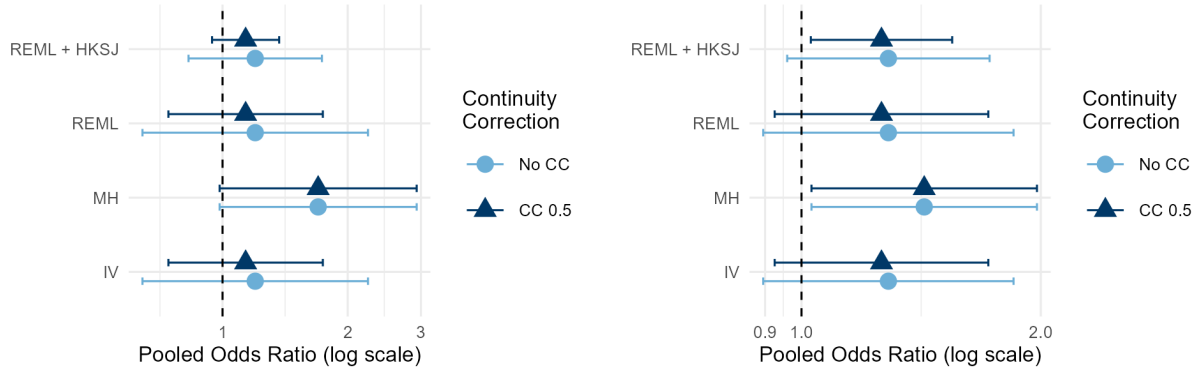
Model	CC	k	$\hat{\tau}^2$	log(OR)	SE	OR [95% CI]
IV	No CC	6	0.000	0.181	0.319	1.20 [0.64, 2.24]
MH	No CC	42	0.000	0.529	0.278	1.70 [0.98, 2.93]
RE	No CC	6	0.000	0.181	0.319	1.20 [0.64, 2.24]
RE + HKSJ	No CC	6	0.000	0.181	0.144	1.20 [0.83, 1.73]
IV	CC 0.5	42	0.000	0.128	0.218	1.14 [0.74, 1.74]
MH	CC 0.5	42	0.000	0.529	0.278	1.70 [0.98, 2.93]
RE	CC 0.5	42	0.000	0.128	0.218	1.14 [0.74, 1.74]
RE + HKSJ	CC 0.5	42	0.000	0.128	0.092	1.14 [0.94, 1.37]

Table 1: Pooled Odds Ratios (OR) for Death Outcome under Different Meta-Analytic Models

Model	CC	k	$\hat{\tau}^2$	log(OR)	SE	OR [95% CI]
IV	No CC	12	0.000	0.252	0.185	1.29 [0.90, 1.85]
MH	No CC	42	0.000	0.356	0.167	1.43 [1.03, 1.98]
RE	No CC	12	0.000	0.252	0.185	1.29 [0.90, 1.85]
RE + HKSJ	No CC	12	0.000	0.252	0.133	1.29 [0.96, 1.72]
IV	CC 0.5	42	0.000	0.232	0.158	1.26 [0.93, 1.72]
MH	CC 0.5	42	0.000	0.356	0.167	1.43 [1.03, 1.98]
RE	CC 0.5	42	0.000	0.232	0.158	1.26 [0.93, 1.72]
RE + HKSJ	CC 0.5	42	0.000	0.232	0.101	1.26 [1.03, 1.55]

Table 2: Pooled Odds Ratios (OR) for Myocardial Infarction under Different Meta-Analytic Models

distributed and beta-distributed baselines, the simulations covered a range of realistic meta-analytic conditions while enabling the comparison of common pooling strategies. Across all scenarios, a higher variance of the baseline risk showed a slight positive effect on estimator performance regarding bias reduction, particularly for the IV and RE models. However, this effect was small and not practically relevant in most situations, suggesting that moderate baseline variability does not substantially alter the statistical properties of classical meta-analysis estimators. In contrast, the correlation between baseline risk and study size—especially the positive correlation scenario—had a pronounced impact on model performance. Under high between-study heterogeneity (τ), all models exhibited a considerable loss in coverage and increased bias, including the RE model based estimators that otherwise performed most robustly. This finding is particularly interesting as it highlights that even random-effects models are sensitive to structural correlations in the data-generating process, potentially leading to biased inference when large studies are systematically associated with higher baseline risks. Despite the systematic design of the simulation study, several limitations must be acknowledged. First, only two distributions for the baseline risk were considered—normal and beta—limiting the generalisability of the findings to other plausible baseline structures. The variance of the beta distribu-



(a) Death (b) Infarction

Figure 25: Forest plots of the fitted models

tion was kept constant for simplicity, which may have constrained the observed effects of baseline heterogeneity. Second, the examined range of baseline variances ($\sigma_{\text{logit}} \in [0.1, 1]$) was restricted to avoid unrealistic event probabilities, particularly in the correlation scenarios where very high variance would have produced implausibly large baseline risks of up to 50%. Future research might address this limitation by exploring alternative data-generating mechanisms beyond the multivariate normal framework. Third, the scope of models fitted to the simulated data was limited to frequentist two-stage estimators. More advanced approaches were not included, such as generalised linear mixed models (GLMMs) or Bayesian binomial–normal models. These methods provide improved stability and inference under sparse-event or highly heterogeneous conditions (Günhan et al., 2020; Jackson et al., 2018) and could serve as valuable extensions to the present study.
The limitations of the simulation framework became evident in the application section, where recommendations could be made for models performing reliably for scenarios involving moderate heterogeneity and imbalanced randomisation. Still, recommendations for higher heterogeneity scenarios were difficult to make. This issue is apparent in the simulation findings, which showed that none of the classical estimators achieved adequate coverage or unbiased estimation when heterogeneity and data imbalance were pronounced. Consequently, the recommendations derived from this work are most applicable to meta-analyses with moderate between-study variability. In contrast, situations characterised by strong heterogeneity or correlation structures may require alternative modelling approaches.

7 Conclusion

This thesis aimed to investigate how different baseline pooling strategies influence the performance of classical meta-analytic models under varying baseline-risk variances, distributions, and structural data characteristics. Building upon the gap identified in the

literature, namely that baseline variability is often treated implicitly rather than as a parameter of interest, this work systematically evaluated the role of explicit baseline modelling within realistic simulation settings. By varying heterogeneity, randomisation imbalance, rare-event occurrence, and the correlation between baseline risk and study size, this study sought to provide a comprehensive assessment of estimator robustness across practical meta-analytic conditions.

The results demonstrated that moderate increases in baseline variance have a negligible impact on estimator performance. Across both normal and beta-distributed baselines, all models IV, MH, RE, and the Hartung–Knapp–Sidik–Jonkman (HKSJ) adjustment showed minimal bias or coverage changes when the baseline variance increased. This finding suggests that the omission of explicit baseline modelling is unlikely to distort inference within realistic parameter ranges substantially.

In contrast, the introduction of a correlation between baseline risk and study size, particularly a positive correlation, strongly affected model performance. Under these conditions, all estimators, including the RE model, showed marked decreases in coverage and increased bias at higher levels of heterogeneity. This indicates that dependencies between design-level characteristics can compromise even variance-based random-effects estimators that are otherwise robust under standard assumptions.

Together, these results clarify the contexts where baseline pooling matters for inference. While baseline variance is a minor determinant of estimator behaviour, the structural relationships between baseline risk and other study-level quantities—such as sample size or allocation ratio—play a far more critical role. This insight contributes to ongoing discussions in the meta-analytic literature (Jackson et al., 2018; Günhan et al., 2020) concerning how data-generating mechanisms interact with estimator assumptions and may help explain the divergent performance of classical and hierarchical models observed in prior simulation studies.

From a methodological perspective, the present work advances the current research landscape by explicitly incorporating baseline variability and correlation structures into the simulation design—features that have largely been ignored in previous studies focusing solely on heterogeneity, randomisation, and zero-event prevalence. The results therefore bridge the gap between theoretical considerations of baseline risk variability and practical concerns about estimator robustness in small, sparse, or imbalanced meta-analyses.

Nevertheless, the findings should be interpreted within the constraints of the study design. Only two baseline distributions were evaluated, and the variance range was restricted to avoid implausible event probabilities. Moreover, the focus on frequentist two-stage estimators excluded other important modelling paradigms, such as Bayesian hierarchical models or generalised linear mixed models (GLMMs), which may exhibit different behaviour in similar conditions. As such, the conclusions primarily apply to classical frequentist frameworks.

Future research should extend this work in three directions. First, incorporating Bayesian and GLMM-based approaches would enable direct comparison with modern hierarchical methods that explicitly model baseline parameters. Second, alternative data-generating distributions could be explored to represent skewed or multimodal baseline structures more realistically. Third, a deeper theoretical investigation into how the correlation between baseline risk and study size propagates through estimator weighting schemes could yield a more formal understanding of the performance degradation observed in this thesis.

In conclusion, this thesis provides new empirical evidence that while baseline pooling and variance have limited effects under standard conditions, structural dependencies particularly correlations between baseline risk and study size pose a substantial challenge to conventional meta-analytic estimators. Addressing these complexities requires more flexible, model-based frameworks that explicitly account for baseline variation. In doing so, future research can contribute to a more comprehensive understanding of estimator behaviour in meta-analysis and strengthen the methodological foundations of evidence synthesis in medical research.

References

Anita Andreano, Paola Rebora, and Maria Grazia Valsecchi. Measures of single arm outcome in meta-analyses of rare events in the presence of competing risks. *Biometrical journal. Biometrische Zeitschrift*, 57(4):649–660, 2015. doi: 10.1002/bimj.201400119.

Lidia R. Arends, Arno W. Hoes, Jacobus Lubsen, Diederik E. Grobbee, and Theo Stijnen. Baseline risk as predictor of treatment benefit: three clinical meta-re-analyses. *Statistics in Medicine*, 19(24):3497–3518, 2000. ISSN 0277-6715. doi: 10.1002/1097-0258(20001230)19:243497::AID-SIM8303.0.CO;2-H.

Marie Beisemann, Philipp Doebler, and Heinz Holling. Comparison of random-effects meta-analysis models for the relative risk in the case of rare events: A simulation study. *Biometrical journal. Biometrische Zeitschrift*, 62(7):1597–1630, 2020. doi: 10.1002/bimj.201900379.

Michael J. Bradburn, Jonathan J. Deeks, Jesse A. Berlin, and A. Russell Localio. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Statistics in medicine*, 26(1):53–77, 2007. doi: 10.1002/sim.2528.

Ji Cheng, Eleanor Pullenayegum, John K. Marshall, Alfonso Iorio, and Lehana Thabane. Impact of including or excluding both-armed zero-event studies on using standard meta-analysis methods for rare event outcome: a simulation study. *BMJ open*, 6(8):e010983, 2016. doi: 10.1136/bmjopen-2015-010983.

Wendimagegn Ghidey, Theo Stijnen, and Hans C. van Houwelingen. Modelling the effect of baseline risk in meta-analysis: a review from the perspective of errors-in-variables regression. *Statistical methods in medical research*, 22(3):307–323, 2013. doi: 10.1177/0962280211412244.

Burak Kürsad Günhan, Christian Röver, and Tim Friede. Random-effects meta-analysis of few studies involving rare events. *Research Synthesis Methods*, 11(1):74–90, 2020. doi: 10.1002/jrsm.1370.

Stefania Iaquinto, Lea Bührer, Maria Feldmann, Beatrice Latal, and Ulrike Held. How to quantify between-study heterogeneity in single-arm evidence synthesis?-it depends! *Systematic reviews*, 14(1):138, 2025. doi: 10.1186/s13643-025-02831-1.

Dan Jackson, Martin Law, Theo Stijnen, Wolfgang Viechtbauer, and Ian R. White. A comparison of seven random-effects models for meta-analyses that estimate the summary odds ratio. *Statistics in medicine*, 37(7):1059–1085, 2018. doi: 10.1002/sim.7588.

Katrin Jansen and Heinz Holling. Random-effects meta-analysis models for the odds ratio in the case of rare events under different data-generating models: A simulation study.

Biometrical journal. Biometrische Zeitschrift, 65(3):e2200132, 2023. doi: 10.1002/bimj.202200132.

Elena Kulinskaya, David C. Hoaglin, and Ilyas Bakbergenuly. Exploring consequences of simulation design for apparent performance of methods of meta-analysis. *Statistical methods in medical research*, 30(7):1667–1690, 2021. doi: 10.1177/09622802211013065.

O. Kuss. Statistical methods for meta-analyses including information from studies without any events-add nothing to nothing and succeed nevertheless. *Statistics in medicine*, 34 (7):1097–1116, 2015. doi: 10.1002/sim.6383.

Dean Langan, Julian P. T. Higgins, Dan Jackson, Jack Bowden, Areti Angeliki Veroniki, Evangelos Kontopantelis, Wolfgang Viechtbauer, and Mark Simmonds. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Research Synthesis Methods*, 10(1):83–98, 2019. doi: 10.1002/jrsm.1316.

Tim Mathes and Oliver Kuss. A comparison of methods for meta-analysis of a small number of studies with binary outcomes. *Research Synthesis Methods*, 9(3):366–381, 2018. doi: 10.1002/jrsm.1296.

Steven E Nissen and Kathy Wolski. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *New England Journal of Medicine*, 356(24):2457–2471, 2007.

Christopher Partlett and Richard D. Riley. Random effects meta-analysis: Coverage performance of 95% confidence and prediction intervals following reml estimation. *Statistics in medicine*, 36(2):301–317, 2017. doi: 10.1002/sim.7140.

Konstantinos Pateras, Stavros Nikolakopoulos, and Kit Roes. Data-generating models of dichotomous outcomes: Heterogeneity in simulation studies for a random-effects meta-analysis. *Statistics in medicine*, 37(7):1115–1124, 2018. doi: 10.1002/sim.7569.

Konstantinos Pateras, Stavros Nikolakopoulos, and Kit C. B. Roes. Prior distributions for variance parameters in a sparse-event meta-analysis of a few small trials. *Pharmaceutical statistics*, 20(1):39–54, 2021. doi: 10.1002/pst.2053.

Romain Piaget-Rossel and Patrick Taffé. Meta-analysis of rare events under the assumption of a homogeneous treatment effect. *Biometrical journal. Biometrische Zeitschrift*, 61(6):1557–1574, 2019. doi: 10.1002/bimj.201800381.

Yanan Ren, Lifeng Lin, Qinshu Lian, Hui Zou, and Haitao Chu. Real-world performance of meta-analysis methods for double-zero-event studies with dichotomous outcomes using the cochrane database of systematic reviews. *Journal of general internal medicine*, 34 (6):960–968, 2019. doi: 10.1007/s11606-019-04925-8.

Gerta Rücker and Martin Schumacher. Simpson's paradox visualized: the example of the rosiglitazone meta-analysis. *BMC medical research methodology*, 8:34, 2008. doi: 10.1186/1471-2288-8-34.

Patarawan Sangnawakij, Dankmar Böhning, Heinz Holling, and Katrin Jansen. Nonparametric estimation of the random effects distribution for the risk or rate ratio in rare events meta-analysis with the arm-based and contrast-based approaches. *Statistics in medicine*, 43(4):706–730, 2024. doi: 10.1002/sim.9981.

Maxi Schulz, Malte Kramer, Oliver Kuss, and Tim Mathes. A re-analysis of about 60,000 sparse data meta-analyses suggests that using an adequate method for pooling matters. *Research Synthesis Methods*, 15(6):978–987, 2024. doi: 10.1002/jrsm.1748.

Svenja E. Seide, Katrin Jensen, and Meinhard Kieser. Simulation and data-generation for random-effects network meta-analysis of binary outcome. *Statistics in medicine*, 38 (17):3288–3303, 2019. doi: 10.1002/sim.8193.

Matthew J. Spittal, Jane Pirkis, and Lyle C. Gurrin. Meta-analysis of incidence rate data in the presence of zero events. *BMC medical research methodology*, 15:42, 2015. doi: 10.1186/s12874-015-0031-0.

Michael J. Sweeting, Alexander J. Sutton, and Paul C. Lambert. What to add to nothing? use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in medicine*, 23(9):1351–1375, 2004. doi: 10.1002/sim.1761.

Simon G. Thompson and Stephen J. Sharp. Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in Medicine*, 18(20):2693–2708, 1999. ISSN 0277-6715. doi: 10.1002/(SICI)1097-0258(19991030)18:20{\textless}2693::AID-SIM235{\textgreater}3.0.CO;2-V.

Yasushi Tsujimoto, Yusuke Tsutsumi, Yuki Kataoka, Akihiro Shiroshita, Orestis Efthimiou, and Toshi A. Furukawa. The impact of continuity correction methods in cochrane reviews with single-zero trials with rare events: A meta-epidemiological study. *Research Synthesis Methods*, 15(5):769–779, 2024. doi: 10.1002/jrsm.1720.

Edwin R. van den Heuvel, Osama Almalik, and Zhuozhao Zhan. Simulation models for aggregated data meta-analysis: Evaluation of pooling effect sizes and publication biases. *Statistical methods in medical research*, 33(3):359–375, 2024. doi: 10.1177/09622802231206474.

S. D. Walter. Variation in baseline risk as an explanation of heterogeneity in meta-analysis. *Statistics in Medicine*, 16(24):2883–2900, 1997. ISSN 0277-6715. doi: 10.1002/(SICI)1097-0258(19971230)16:24{\textless}2883::AID-SIM825{\textgreater}3.0.CO;2-B.

Wolfgang Viechtbauer. Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics*, 30(3):261–293, 2005. URL `https://www.jstor.org/stable/3701379`.

Minghong Yao, Yuning Wang, Yan Ren, Yulong Jia, Kang Zou, Ling Li, and Xin Sun. Comparison of statistical methods for integrating real-world evidence in a rare events meta-analysis of randomized controlled trials. *Research Synthesis Methods*, 14(5):689–706, 2023. doi: 10.1002/jrsm.1648.

Minghong Yao, Yulong Jia, Fan Mei, Yuning Wang, Kang Zou, Ling Li, and Xin Sun. Comparing various bayesian random-effects models for pooling randomized controlled trials with rare events. *Pharmaceutical statistics*, 23(6):837–853, 2024. doi: 10.1002/pst.2392.

Minghong Yao, Fan Mei, Kang Zou, Ling Li, and Xin Sun. Comparison of prior distributions for the heterogeneity parameter in a rare events meta-analysis of a few studies. *Pharmaceutical statistics*, 24(2):e2448, 2025. doi: 10.1002/pst.2448.

Brinley N. Zabriskie, Nolan Cole, Jacob Baldauf, and Craig Decker. The impact of correction methods on rare-event meta-analysis. *Research Synthesis Methods*, 15(1):130–151, 2024. doi: 10.1002/jrsm.1677.

A Appendix

A.1 Simulated Data

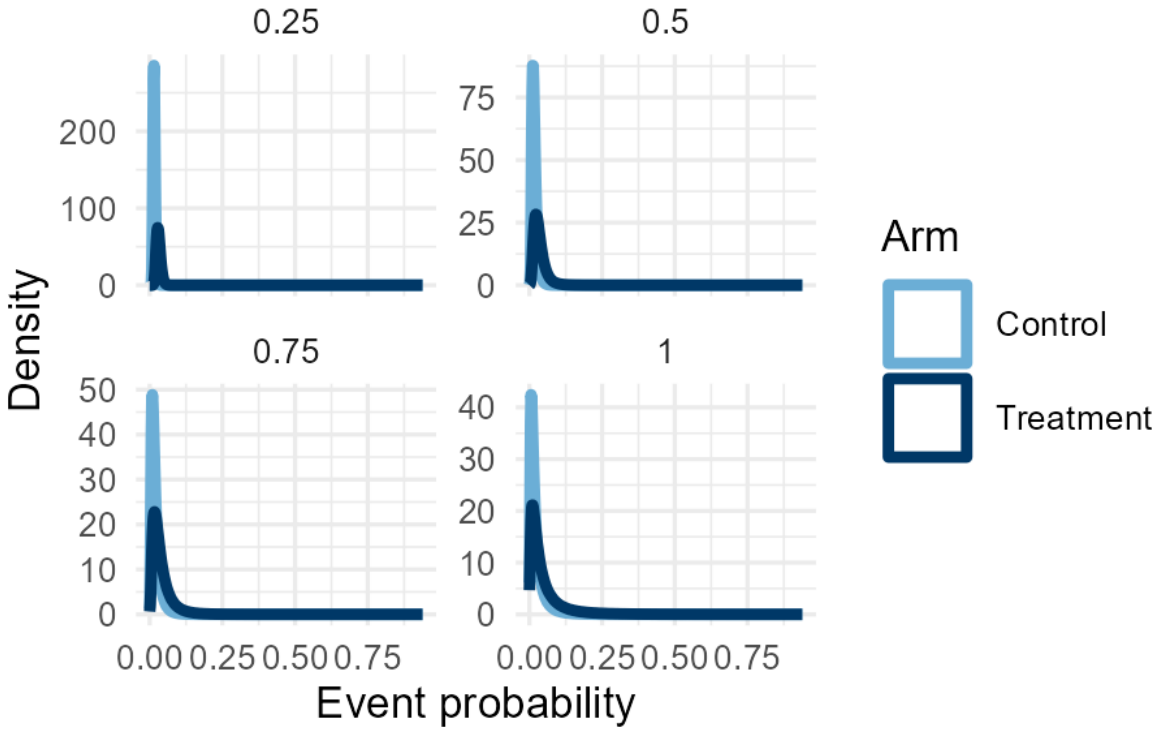


Figure 26: Risk Distribution (Negative Correlation Scenario)

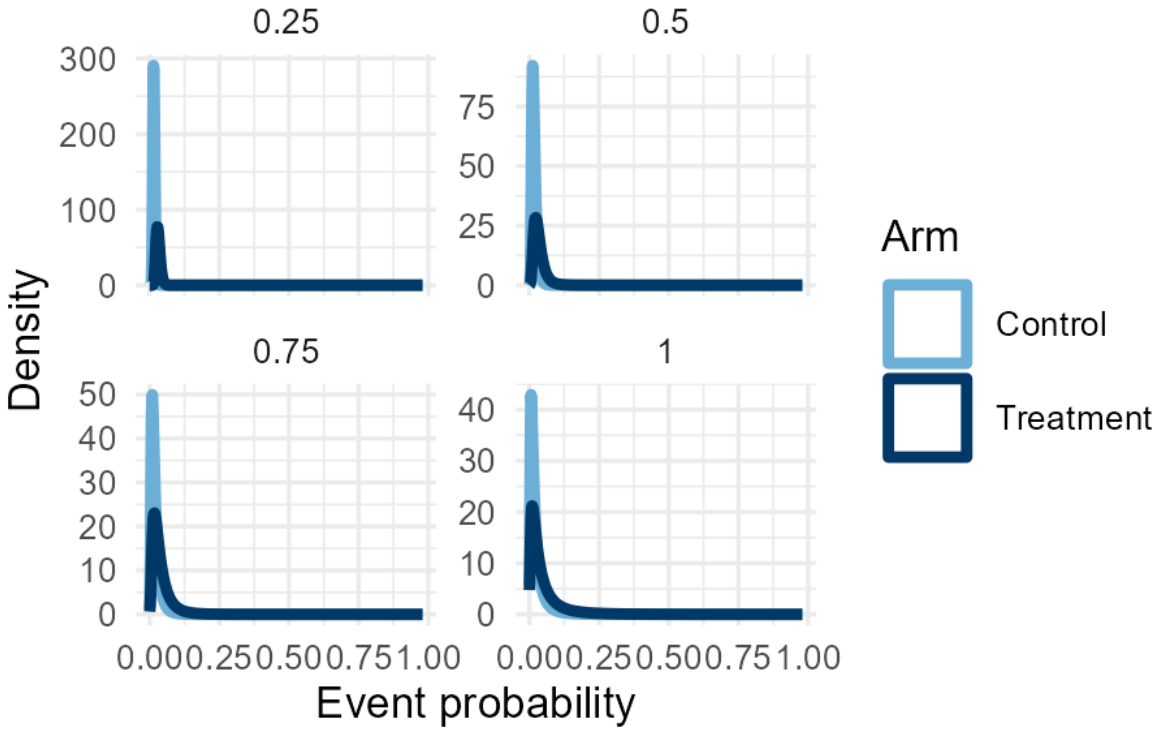


Figure 27: Risk Distribution (Positive Correlation Scenario)

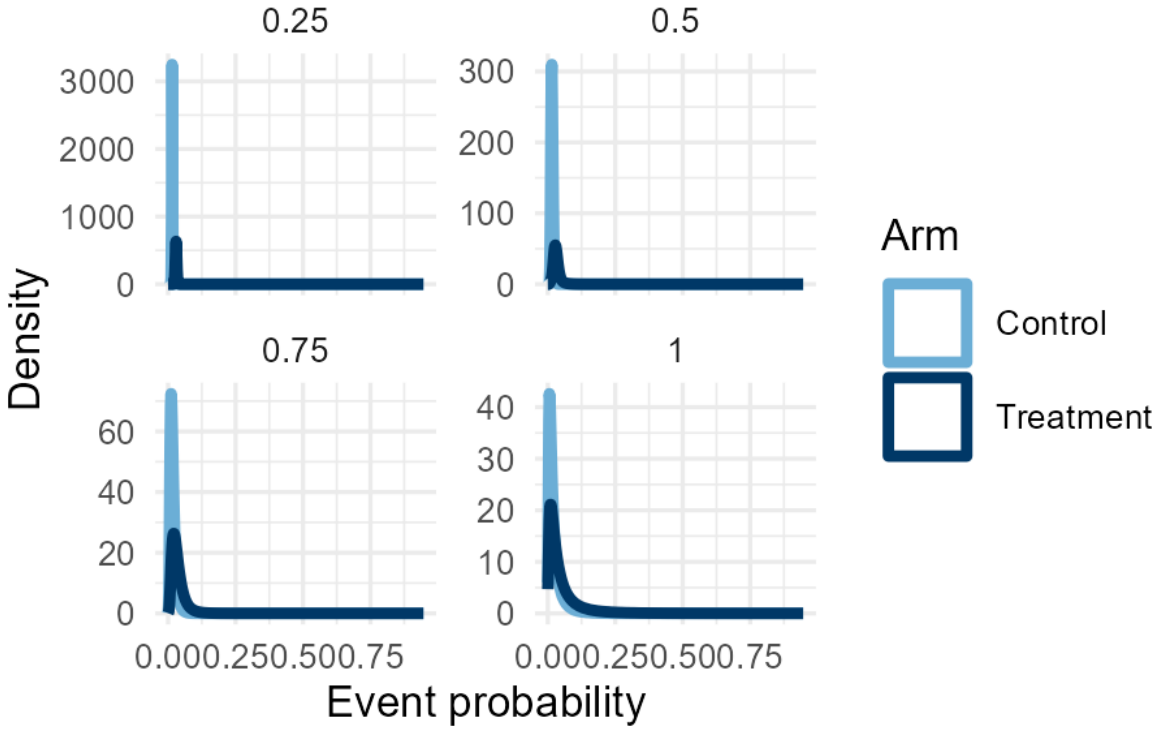


Figure 28: Risk Distribution (Randomisation 1:2 Scenario)

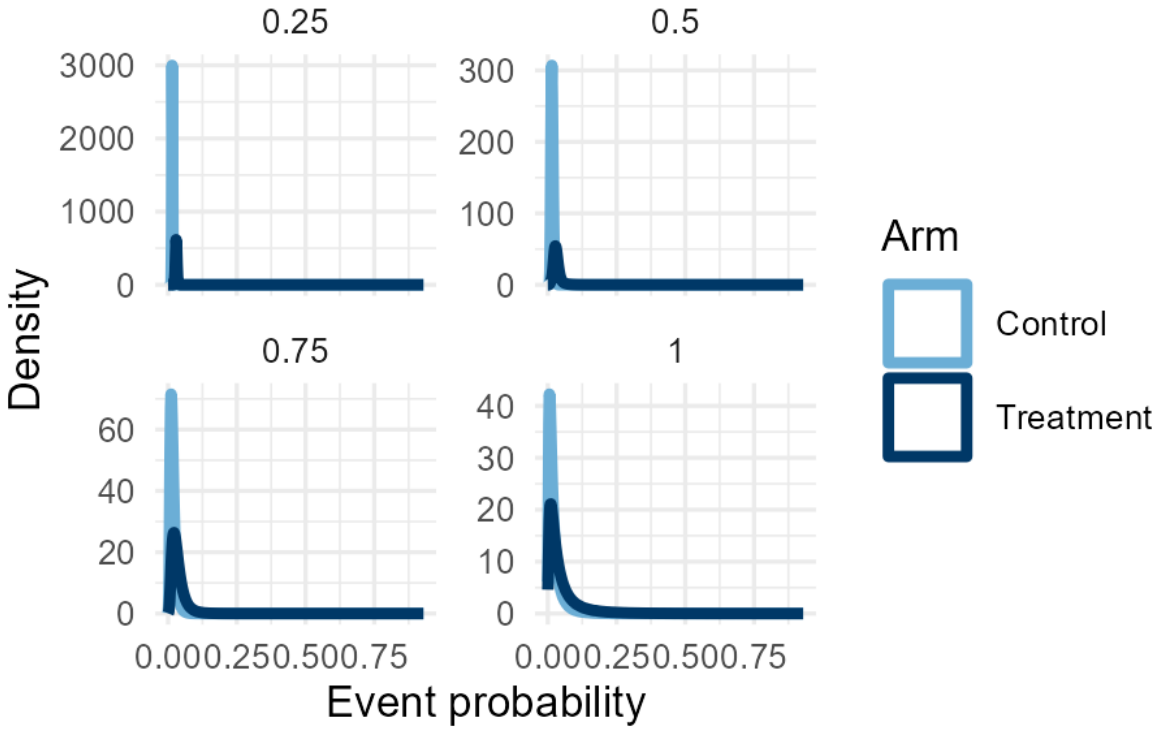
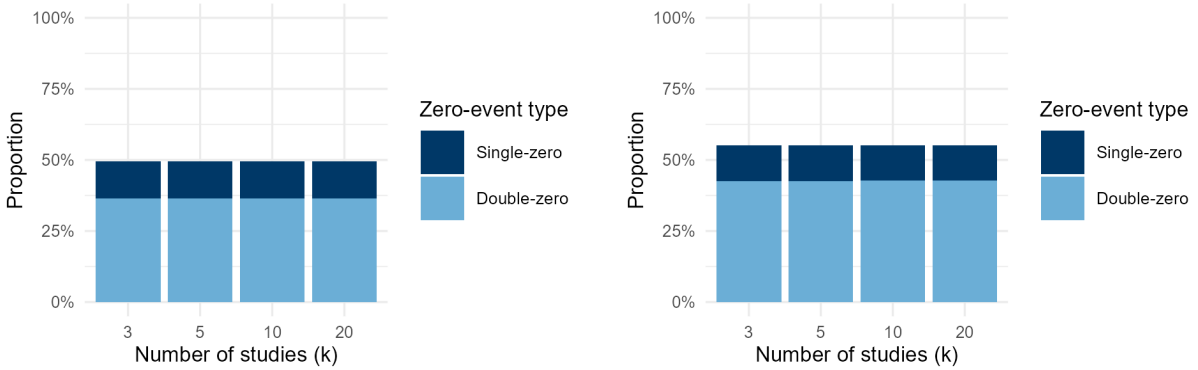


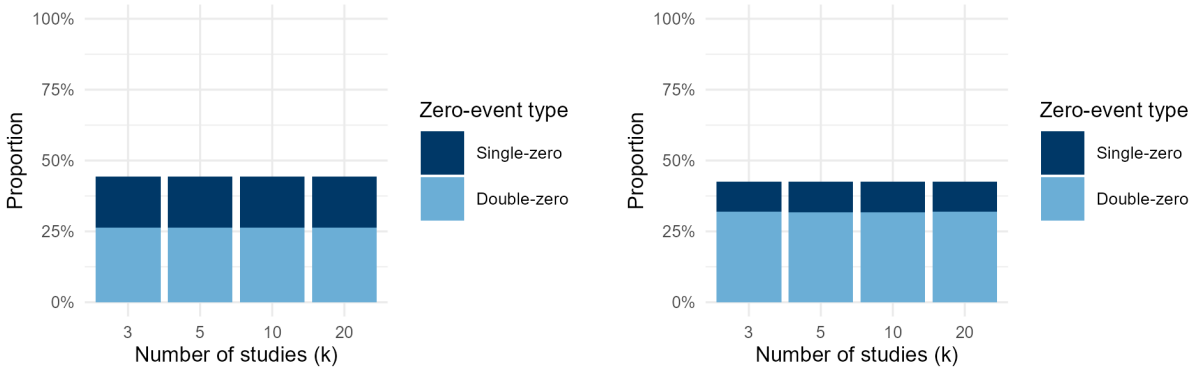
Figure 29: Risk Distribution (Randomisation 1:3 Scenario)



(a) Randomisation 1:2

(b) Randomisation 1:3

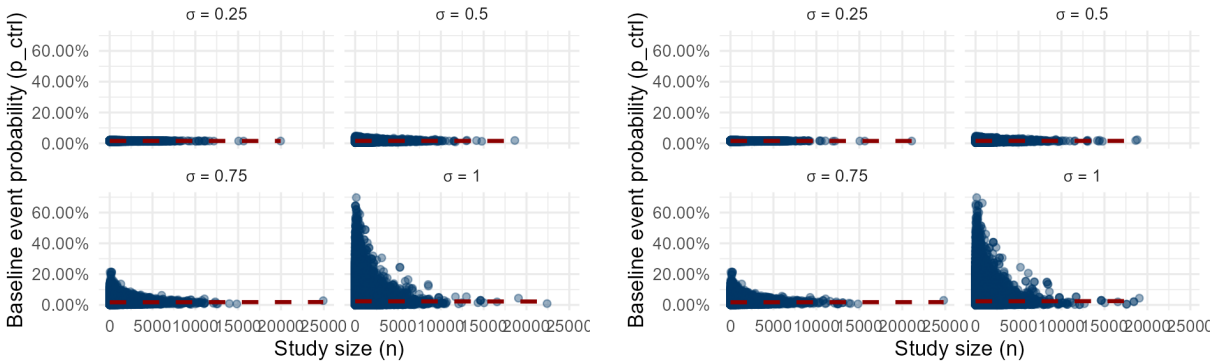
Figure 30: Zero percentages for the randomisation 1:2 and 1:3 Scenario



(a) Positive correlation

(b) Negative correlation

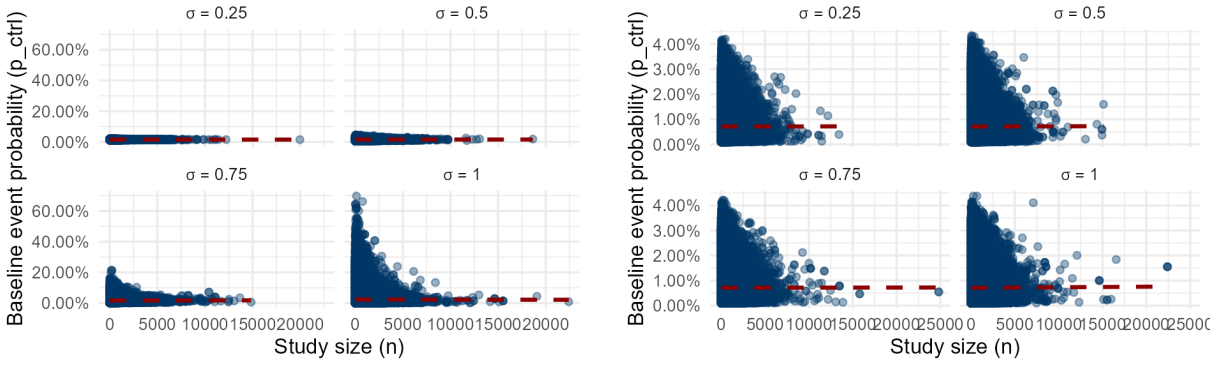
Figure 31: Zero percentages for the positive and negative correlation Scenario



(a) Randomisation 1:2

(b) Randomisation 1:3

Figure 32: Study size vs. baseline risk for randomisation 1:2 and 1:3 scenario



(a) Standard

(b) Many zeros

Figure 33: Study size vs. baseline risk for standard and many zeros scenario

Declaration of Authorship

I hereby declare that I have written this Master's thesis independently and without unauthorized assistance. All sources and references that have been used or quoted directly or indirectly are acknowledged as such.
I further declare that I have not previously submitted this thesis, in whole or in part, for the purpose of obtaining an academic degree at any other institution.
Artificial intelligence (AI) tools, such as OpenAI's ChatGPT, were used solely to assist in refining text passages, supporting code implementation in `R`, and identifying relevant literature. The content, structure, analysis, and conclusions of the thesis represent my own independent work and intellectual contribution.

Place, Date: _____

Signature: _____

Name: Phil Tobeck