Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2021, Vol. 30(7) 1667–1690
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/09622802211013065
journals.sagepub.com/home/smm
SAGE

Exploring consequences of simulation design for apparent performance of methods of meta-analysis

Elena Kulinskaya[1], David C. Hoaglin[2] and Ilyas Bakbergenuly[1]

Abstract

Contemporary statistical publications rely on simulation to evaluate performance of new methods and compare them with established methods. In the context of random-effects meta-analysis of log-odds-ratios, we investigate how choices in generating data affect such conclusions. The choices we study include the overall log-odds-ratio, the distribution of probabilities in the control arm, and the distribution of study-level sample sizes. We retain the customary normal distribution of study-level effects. To examine the impact of the components of simulations, we assess the performance of the best available inverse–variance–weighted two-stage method, a two-stage method with constant sample-size-based weights, and two generalized linear mixed models. The results show no important differences between fixed and random sample sizes. In contrast, we found differences among data-generation models in estimation of heterogeneity variance and overall log-odds-ratio. This sensitivity to design poses challenges for use of simulation in choosing methods of meta-analysis.

Keywords

Meta-analysis, odds-ratio, random-effects model, random probabilities, random sample sizes

1 Introduction

Many methodological publications in applied statistics develop a new method, illustrate it in examples, and evaluate its performance by simulation. Our interest lies in methods for meta-analysis (MA). For meta-analysis of odds ratios, we demonstrate how researchers' choices of simulation design can affect conclusions on the comparative merits of various methods.

Presentations of meta-analysis methods usually include assumptions about the behavior of the estimates from the individual studies. For example, a generic 2-stage random-effects model relates the observed effect sizes y_i ($i = 1, \ldots, K$) to the overall effect μ in the model

$$y_i = \mu + \delta_i + \varepsilon_i \tag{1}$$

where $\delta_i \sim N(0, \tau^2)$ represent random variation in the underlying study-level effects, the $\varepsilon_i \sim N(0, \sigma_i^2)$ represent random variation within the studies, and the δ_i and the ε_i are independent. From the y_i and their estimated variances, $s_i^2 = \widehat{\sigma}_i^2$, the two-stage method estimates μ and also τ^2. Such a model can serve as a basis for analysis and also as the basis for generating data as part of a simulation study. The analysis model and the data-generation

[1]School of Computing Sciences, University of East Anglia, Norwich, UK
[2]Department of Population and Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, MA, USA

Corresponding author:
Elena Kulinskaya, School of Computing Sciences, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, UK.
Email: e.kulinskaya@uea.ac.uk

model may differ, however. For example, when the measure of effect is the log-odds-ratio, the data-generation model produces more-basic study-level data (such as numbers of events in the two arms, as shown in Section 2), from which y_i and s_i^2 are calculated, and the popular inverse–variance–weighted methods build on equation (1). On the other hand, other methods, such as generalized linear models, build on the likelihood for the distributions in the data-generation model. In order to study the impact of choices among data-generation models – our primary interest – our simulations use several analysis models and methods based on them.

For a particular method, one can regard a measure of performance, such as the bias of a point estimator or the coverage of an interval estimator, as a function of variables that describe the meta-analysis and its setting. By a combination of analysis and, mainly, simulation, one aims to evaluate that function and describe its behavior. The variables include the number of studies, the study-level sample sizes, the extent of imbalance of the arm-level sample sizes, the overall effect, the between-study variance of the effect (for a random-effects method), and the arm-level variances within the studies (if the effect is continuous); and the relation of the performance measure to the variables usually involves nonlinearities and interactions. Thus, the design of a simulation has important implications for accuracy in evaluating the function, for estimating those relations, and, especially, for relevance of the results to practice.

The conventional meta-analysis of odds ratios from K studies involves $2K$ binomial variables, $X_{ij} \sim \text{Bin}(n_{ij}, p_{ij})$ for $i = 1, \ldots, K$ and $j = C$ or T (for the Control or Treatment arm). The random-effects model assumes that $\text{logit}(p_{ij}) = \alpha_i + \theta_i z_{ij}$ for $\theta_i \sim N(\theta, \tau^2)$ and an indicator z_{ij} taking on values 0 (for Control) and 1 (for Treatment). In this notation, $\alpha_i = \text{logit}(p_{iC})$ and $\alpha_i + \theta_i = \text{logit}(p_{iT})$.

A design specifies a systematic collection of situations involving the number of studies, K; the sample sizes, n_{ij}; the control-arm probabilities, p_{iC}, or, equivalently, their logits, α_i; the overall log-odds-ratio, θ; and the between-study variance, τ^2. For each situation the simulation uses M replications, where M is typically large, say 10,000.

For simplicity, we consider equal arm-level sample sizes, $n_{iC} = n_{iT} = n_i$; some studies use a random allocation ratio centered at a given percentage, q. Studies vary in how they specify the n_i. Choices include setting $n_1 = \cdots = n_K$ with the same value in all M replications, using a fixed set of n_i (not all equal), and using some distribution (typically normal or uniform) to generate a new set of n_i in each replication.

Similarly, the p_{iC} or their logits α_i can be fixed or generated from some distribution. Again, normal and uniform distributions are the typical choices.

For τ^2 most studies use a few selected values or an equally spaced set, such as $\tau^2 = 0, 0.1, \ldots, 1$, though some generate τ^2 randomly.[1] Some studies specify values of the heterogeneity measure I^2 and obtain values of τ^2 indirectly.

In Section 2, we review approaches for generating log-odds-ratios and control-arm probabilities, and consider their statistical consequences. For two-stage methods of meta-analysis, which use the studies' sample log-odds-ratios and their estimated variances, the relation between the estimates and their inverse-variance weights can produce bias. Section 3 examines this complication analytically, for fixed study-level sample sizes. Section 4 discusses approaches for generating sample sizes randomly and analyzes their impact. In Section 5 we study, by simulation, how various choices in generating data affect comparative merits of several established meta-analytic methods in estimating the between-study variance τ^2 and the overall log-odds-ratio θ. The methods we study include the best available two-stage methods for MA: the Mandel-Paule estimator of τ^2 and the corresponding inverse–variance–weighted estimator of θ with a confidence interval based on the normal distribution. We also consider the performance of two GLMM methods and a two-stage estimator of θ with constant sample-size-based weights whose confidence interval is based on the t distribution. Section 6 describes and summarizes the results. Discussion, in Section 7, offers concluding remarks. Appendices 1 and 2 provide technical details for Section 3. Additional figures are provided in online Supplemental material.

2 Generation of log-odds-ratios and control-arm probabilities

Consider K studies that used a particular individual-level binary outcome. Each study reports X_{iT} and X_{iC}, the numbers of events in the n_{iT} subjects in the Treatment arm and the n_{iC} subjects in the Control arm, for $i = 1, \ldots, K$. It is customary to treat X_{iT} and X_{iC} as independent binomial variables

$$X_{iT} \sim \text{Bin}(n_{iT}, p_{iT}) \quad \text{and} \quad X_{iC} \sim \text{Bin}(n_{iC}, p_{iC}) \tag{2}$$

The log-odds-ratio for Study i is

$$\theta_i = \log_e\left(\frac{p_{iT}(1-p_{iC})}{p_{iC}(1-p_{iT})}\right) \text{ estimated by } \widehat{\theta}_i = \log_e\left(\frac{\widehat{p}_{iT}(1-\widehat{p}_{iC})}{\widehat{p}_{iC}(1-\widehat{p}_{iT})}\right) \tag{3}$$

The (conditional, given p_{ij} and n_{ij}) variance of $\widehat{\theta}_i$, derived by the delta method, is

$$v_i^2 = \mathrm{Var}(\widehat{\theta}_i) = \frac{1}{n_{iT}p_{iT}(1-p_{iT})} + \frac{1}{n_{iC}p_{iC}(1-p_{iC})} \tag{4}$$

estimated by substituting $\widehat{p}_{ij}$ for p_{ij}. (In analyses, we follow the particular method's procedure for calculating $\widehat{p}_{ij}$.)

Under the binomial-normal random-effects model (REM), the true study-level effects, θ_i, follow a normal distribution

$$\theta_i \sim N(\theta, \tau^2) \tag{5}$$

For analysis, the resulting logistic mixed-effects model belongs to the class of generalized linear mixed models (GLMMs).[2,3] Kuss,[1] Jackson et al.,[4] and Bakbergenuly and Kulinskaya[5] review these GLMM methods.

In practice, p_{iC} and p_{iT} vary among studies in a variety of ways, not necessarily described by any particular distribution. Almost all analyses and simulations use the binomial-normal REM for the relation between p_{iT} and p_{iC}. Simulations can treat the p_{iC} as constant (e.g. at a sequence of values) or sample them from a distribution, either directly (usually from a uniform distribution or a more general beta distribution; Section 2.2 discusses beta and beta-binomial models) or indirectly, by generating $\mathrm{logit}(p_{iC})$ (usually from a Gaussian distribution).

For reference, Table 1 lists the various data-generation models considered in more detail later.

2.1 Models with fixed and random intercepts

We consider two fixed-intercepts random-effects models (FIM1 and FIM2, Section 2.1.1) and two random-intercept random-effects models (RIM1 and RIM2, Section 2.1.2) as in Bakbergenuly and Kulinskaya.[5] These models are equivalent to Models 2 and 4 (for FIM) and Models 3 and 5 (for RIM), respectively, of Jackson et al.[4] Briefly, the FIMs include fixed control-arm effects (log-odds of the control-arm probabilities), and the RIMs replace these fixed effects with random effects.

Under the fixed-effect (common-effect) model, $\tau^2 = 0$ and $\theta_i \equiv \theta$. Still, the control-arm effects can be either fixed or random, resulting in two fixed-effect models: the fixed-intercepts fixed-effect model FIM1F, and the random-intercept fixed-effect model RIM1F. Random-intercept fixed-effect models were considered by Kuss[1] and Piaget-Rossel and Taffé.[6] However, GLMMs with random θ_i are traditional in meta-analysis.

Table 1. Summary of data-generation models for log-odds-ratio.

Data-generation model	Intercept α_i or $\alpha + u_i$	Study-level random effects b_i	Fraction of random effect in Control arm (c)
FIM1	fixed α_i	$N(0, \tau^2)$	0
FIM2	fixed α_i	$N(0, \tau^2)$	1/2
RIM1	$u_i \sim N(0, \sigma^2)$	$N(0, \tau^2)$	0
RIM2	$u_i \sim N(0, \sigma^2)$	$N(0, \tau^2)$	1/2
URIM1	p_{iC} uniform	$N(0, \tau^2)$	0
FIM1F	fixed α_i	$\tau^2 = 0$	N/A
RIM1F	$u_i \sim N(0, \sigma^2)$	$\tau^2 = 0$	N/A
URIM1F	p_{iC} uniform	$\tau^2 = 0$	N/A

Note: In the fixed-intercept models, $\log(p_{iT}/(1-p_{iT})) = \alpha_i + \theta + (1-c)b_i$ and $\log(p_{iC}/(1-p_{iC})) = \alpha_i - cb_i$. In the random-intercept models, $\log(p_{iT}/(1-p_{iT})) = \alpha + u_i + \theta + (1-c)b_i$ and $\log(p_{iC}/(1-p_{iC})) = \alpha + u_i - cb_i$.

2.1.1 Fixed-intercepts models (FIM1 and FIM2)

The fixed-intercepts models assume fixed effects for the studies' control arms and allow heterogeneity in odds ratios among studies. (We follow Rice et al.[7] in using the plural form for fixed intercepts that differ among the studies.) Given the binomial distributions in the two arms (equation (2)), the model is ($i = 1, \ldots, K$)

$$\begin{aligned} \log\left(\frac{p_{iT}}{1-p_{iT}}\right) &= \alpha_i + \theta + (1-c)b_i \\ \log\left(\frac{p_{iC}}{1-p_{iC}}\right) &= \alpha_i - cb_i \end{aligned} \tag{6}$$

where the α_i are the fixed control-arm effects, θ is the overall log-odds-ratio, and the $b_i \sim N(0, \tau^2)$ are random effects. Under FIM1, $c = 0$, resulting in higher variance in the treatment arm. Under FIM2, $c = 1/2$, splitting the random effect b_i equally between the two equations and yielding equal variance in the two arms. When $\tau^2 \equiv 0$, these two models become a fixed-intercepts fixed-effect model, FIM1F.

An analysis has to estimate the fixed study-specific intercepts α_i (usually regarded as nuisance parameters), along with θ and τ^2. In a logistic mixed-effects regression, these $K+2$ parameters are estimated iteratively, using marginal quasi-likelihood, penalized quasi-likelihood, or a first- or second-order-expansion approximation. Jackson et al.[4] demonstrate that inference using FIM2 is preferable, even though they generate data from FIM1.

2.1.2 Random-intercept models (RIM1 and RIM2)

As K becomes large, it may be inconvenient, even problematic, for analysis to have a separate α_i for each study. One can replace those fixed intercepts with random intercepts $\alpha + u_i$, centered at α:

$$\begin{aligned} \log\left(\frac{p_{iT}}{1-p_{iT}}\right) &= \alpha + u_i + \theta + (1-c)b_i \\ \log\left(\frac{p_{iC}}{1-p_{iC}}\right) &= \alpha + u_i - cb_i \end{aligned} \tag{7}$$

As before, θ is the overall log-odds-ratio, and $b_i \sim N(0, \tau^2)$. RIM1 and RIM2 correspond to $c = 0$ and 1/2, respectively. Now the $u_i \sim N(0, \sigma^2)$, and u_i and b_i can be correlated: $\mathrm{Cov}(u_i, b_i) = \rho\sigma\tau$. (If this bivariate normal distribution is not correct, however, estimates of θ will be biased.[8]) Under RIM1, heterogeneity of log-odds is represented in the control arms by the variance σ^2 and in the treatment arms by $\sigma^2 + 2\rho\sigma\tau + \tau^2$. Typically, ρ is taken as zero in simulation. The standard two-stage random-effects analysis model, which works with the sample log-odds-ratios, involves only a single between-study variance, τ^2. Turner et al.[2] point out that ρ should be estimated. Estimation of α, θ, σ^2, τ^2 and ρ is similar to estimation of the parameters in the fixed-intercept model.[2] Again, RIM2 is preferable to RIM1 for inference.

When $\tau^2 \equiv 0$, these two models become a random-intercept fixed-effect model, denoted by RIM1F.

The vast majority of simulation studies use FIM1 or RIM1 for data generation, both for standard two-stage methods of MA and when studying performance of GLMMs, even when they use FIM2 or RIM2 for inference. Examples include Sidik and Jonkman,[9] Platt et al.,[10] Bakbergenuly and Kulinskaya,[5] and Cheng et al.[11] for FIM, and Abo-Zaid et al.[12] ($\sigma = 0.25$ and 1.5), Kosmidis et al.[13] ($\sigma^2 = 0.1$), and Jackson et al.[4] (Settings 1 to 12) ($\sigma = 0.3$) for RIM.

Langan et al.[14] use a somewhat more complicated simulation scheme, which either fixes the average within-study probabilities $\bar{p}_i$ (at .5, .05, and .001) or generates them from $U(.1, .5)$, and then derives the values of p_{iC} and p_{iT} from the values of $\bar{p}_i$ and θ_i, the latter normally distributed as in equation (5). Thus, p_{iC} satisfies the equation $\mathrm{logit}(p_{iC}) = \mathrm{logit}(2\bar{p}_i - p_{iC}) - \theta_i$. So $\mathrm{logit}(p_{iC})$ has a share of the variance, making this a version of FIM2 or RIM2.

2.1.3 Moments of the control-arm probability under RIM

The Gaussian random-intercept models generate the control-arm probabilities, p_{iC}, indirectly: $\mathrm{logit}(p_{iC})$ has a normal distribution centered at $\alpha = \mathrm{logit}(p_C^0)$. On the probability scale, where $p_C^0 = \mathrm{expit}(\alpha) = \exp(\alpha)/(1 + \exp(\alpha))$, the distribution is unimodal and skewed to the right when $p_C^0 < 0.5$. Thus, simulations

from RIM involve, on average, higher control-arm probabilities than corresponding simulations from FIM, though the median control-arm probability is the same. (In FIM1, the distribution has a single value: $p_{iC} = \text{expit}(\alpha_i)$.) To aid in comparing FIM and RIM, we evaluate the mean and variance of this distribution; we use the standard delta method.

For a transformed random variable $Y = h(X)$

$$\mathrm{E}(Y) = h(\mathrm{E}(X)) + h''(\mathrm{E}(X))\mathrm{Var}(X)/2 \text{ and } \mathrm{Var}(Y) = [h'(\mathrm{E}(X))]^2\mathrm{Var}(X) \tag{8}$$

For $\alpha_i = \mathrm{E}(\text{logit}(p_{iC}))$ and $p_C^0 = \text{expit}(\alpha)$, we have

$$p_C^0 = h(\alpha) = \frac{\exp(\alpha)}{1+\exp(\alpha)} = 1 - \frac{1}{1+\exp(\alpha)}$$

The derivatives of $h(\cdot)$ at α are

$$h'(\alpha) = \frac{\exp(\alpha)}{(1+\exp(\alpha))^2} = p_C^0(1-p_C^0)$$

and

$$h''(\alpha) = \frac{\exp(\alpha)(1-\exp(\alpha))}{(1+\exp(\alpha))^3} = p_C^0(1-p_C^0)(1-2p_C^0)$$

Hence

$$\mathrm{E}(p_{iC}) = p_C^0 + p_C^0(1-p_C^0)(1-2p_C^0)\sigma^2/2 \quad \text{and} \quad \mathrm{Var}(p_{iC}) = [p_C^0(1-p_C^0)]^2\sigma^2$$

The mean probability increases with the variance, σ^2, of the normal distribution of u_i. For $p_C^0 = .1$, say, the mean is .100 when $\sigma^2 = 0.01$, but it increases to .109 for $\sigma^2 = 0.25$ and to .136 for $\sigma^2 = 1$. Therefore, simulations from FIM and RIM are not quite equivalent.

2.2 Non-Gaussian random-intercept models

Other distributions besides the Gaussian can serve as a mixing distribution for control-arm probabilities.

In Bayesian analysis, the beta distributions are conjugate priors for a binomial, so they are a natural choice for a mixing distribution. The result is a marginal beta-binomial distribution in the control arm. In meta-analysis a beta-binomial model[1,15] usually assumes beta-binomial distributions in both arms. However, Bakbergenuly and Kulinskaya[15] showed that the standard RE method does not perform well when the data come from a beta-binomial model. Therefore, we would not use a RIM with beta-generated probabilities.

We are not aware of any simulation studies that intentionally used a beta distribution for control-arm probabilities. However, the Beta(1,1) distribution is the same as $U(0, 1)$, and a popular choice is a uniform distribution on an interval, $(p_l, p_u) \subset [0, 1]$. Viechtbauer,[16] Sidik and Jonkman,[17] and Nagashima et al.[18] (Set iii) generated the p_{iC} from $U(.05, .65)$ in combination with the Gaussian REM. Similarly, Jackson et al.[4] (Setting 13) generated the p_{iC} from $U(.1, .3)$. All these studies add a uniform distribution of control-arm probabilities to the FIM1 setting, producing a random-intercept model that we denote by URIM1. This model retains the normal distribution of the θ_i.

Piaget-Rossel and Taffé[6] used a fixed-effect model with $p_{iC} \sim U(p - p/5, p + p/5)$, URIM1F in our nomenclature, with $p = .1, .007, .0035, .0015$. Piaget-Rossel[19] used the same distribution for the p_{iC} and uniformly distributed log-odds-ratios, $\theta_i \sim U(\theta \pm \sqrt{3\tau^2})$.

If $X \sim \mathrm{Bin}(n, p)$ and $p \sim U(0, 1)$, then X has the discrete uniform distribution $U(0, 1, \ldots, n)$. More generally, when $p \sim U(p_l, p_u)$, the probabilities for the numbers of successes are

$$\begin{aligned} P(X = k) &= \frac{1}{p_u - p_l} \int_{p_l}^{p_u} \binom{n}{k} u^k (1-u)^{n-k} \mathrm{d}u \\ &= \frac{1}{p_u - p_l} [B(p_u; k+1, n-k+1) \\ &\quad - B(p_l; k+1, n-k+1)] \end{aligned} \tag{9}$$

where $B(\cdot; \cdot, \cdot)$ denotes the incomplete beta function. To examine the effects on the performance of the MA methods, our simulations include uniform distributions of control-arm probabilities.

3 Variances and covariances of estimated log-odds-ratios and their weights

Traditional one-size-fits-all meta-analysis proceeds in two stages: obtain the study-level estimates and their estimated variances (or standard errors) and then estimate the overall effect as a weighted mean with inverse-variance weights. One of its main faults is that it ignores the variability of the estimated variances. As a result, the variance of the overall effect is underestimated.[20] Additionally, a relation between the estimated study-level effects and their estimated variances may lead to bias in the estimate of the overall effect. In this section, we explore these relations for the log-odds-ratio and its variance and inverse-variance weight. We also demonstrate that the relation varies with the data-generation mechanism. In particular, the sample log-odds-ratio and its estimated variance can be almost independent under FIM2 and RIM2 when $\theta = 0$. Because the calculations are somewhat easier, we first examine the relation to the estimated variance and then turn to the relation to the inverse-variance weight.

3.1 Relation of sample log-odds-ratio and its estimated variance

The data-generation mechanisms for the random-effects model generate the p_{iC} and p_{iT} and then generate X_{iC} and X_{iT}, according to equation (2). Thus, to obtain the covariance between $\widehat{\theta}_i$ and $\widehat{\mathrm{Var}}(\widehat{\theta}_i)$, we apply the law of total covariance

$$\mathrm{Cov}(\widehat{\theta}_i, \widehat{\mathrm{Var}}(\widehat{\theta}_i)) = \mathrm{E}[\mathrm{Cov}(\widehat{\theta}_i, \widehat{\mathrm{Var}}(\widehat{\theta}_i)|\alpha_i, \theta_i)] + \mathrm{Cov}(\mathrm{E}(\widehat{\theta}_i|\alpha_i, \theta_i), \mathrm{E}(\widehat{\mathrm{Var}}(\widehat{\theta}_i)|\alpha_i, \theta_i)) \tag{10}$$

In the process, to show the full effect of the data-generating mechanism, we also obtain $\mathrm{Var}(\widehat{\theta}_i)$, using the more-familiar law of total variance

$$\mathrm{Var}(\widehat{\theta}_i) = \mathrm{E}[\mathrm{Var}(\widehat{\theta}_i|\alpha_i, \theta_i)] + \mathrm{Var}(\mathrm{E}(\widehat{\theta}_i|\alpha_i, \theta_i)) = \mathrm{E}(v_i^2) + \tau^2 \tag{11}$$

In equation (10), the covariance of the conditional expectations is just $\mathrm{Cov}(\theta_i, v_i^2)$ because $\mathrm{E}(\widehat{\theta}_i|\alpha_i, \theta_i) = \theta_i = \theta + b_i$ and (to first order) $\mathrm{E}(\widehat{\mathrm{Var}}(\widehat{\theta}_i)|\alpha_i, \theta_i) = v_i^2$. Thus, we need to calculate $\mathrm{Cov}(\widehat{\theta}_i, \widehat{\mathrm{Var}}(\widehat{\theta}_i)|\alpha_i, \theta_i)$ and take its expectation. Conditioning on α_i and θ_i, in equation (6) and equation (7), is equivalent to conditioning on p_{iC} and p_{iT}. Therefore, we can rewrite equation (10) as (shortening $\widehat{\mathrm{Var}}(\widehat{\theta}_i)$ to $\widehat{v}_i^2$)

$$\mathrm{Cov}(\widehat{\theta}_i, \widehat{v}_i^2) = \mathrm{E}[\mathrm{Cov}(\widehat{\theta}_i, \widehat{v}_i^2|p_{iC}, p_{iT})] + \mathrm{Cov}(\theta_i, v_i^2)$$

The first term in the above equation accounts for the binomial variation (of order $1/n_i$) in $\widehat{\theta}_i$ and in $\widehat{v}_i^2$, given p_{iC} and p_{iT}, whereas the second term accounts for the variation of its expected value and variance from random effects, of order 1 in model (7). Therefore, the first, binomial term is of smaller order ($O(n_i^{-2})$) than the second term (the covariance of the expected moments) and can be neglected in a calculation to order $1/n_i$.

To calculate the covariance of θ_i and v_i^2, we assume, for simplicity, that u_i and b_i are independent in RIM1 and RIM2. Then, defining $p_C = \text{expit}(\alpha)$ and $p_T = \text{expit}(\alpha + \theta)$, to order $1/n_i$

$$\text{Cov}(\theta_i, v_i^2) = \frac{c\tau^2}{n_{iC}}\left[\frac{1-2p_C}{p_C(1-p_C)}\right] - \frac{(1-c)\tau^2}{n_{iT}}\left[\frac{1-2p_T}{p_T(1-p_T)}\right] \tag{12}$$

In particular, when $c = 1/2$, $\theta = 0$ and $n_{iT} = n_{iC}$, $\text{Cov}(\theta_i, v_i^2) = 0$.

After some algebra, we also obtain, to order $1/n_i$,

$$\begin{aligned}\text{Var}(\widehat{\theta}_i) = {} & [n_{iT}p_T(1-p_T)]^{-1} + [n_{iC}p_C(1-p_C)]^{-1} + \tau^2 \\ & + \left(\sigma^2 + (1-c)^2\tau^2 + 2(1-c)\rho\sigma\tau\right)[2n_{iT}]^{-1}\left([p_T(1-p_T)]^{-1} - 2\right) \\ & + \left(\sigma^2 + c^2\tau^2 - 2c\rho\sigma\tau\right)[2n_{iC}]^{-1}\left([p_C(1-p_C)]^{-1} - 2\right)\end{aligned} \tag{13}$$

The binomial variance component v_i^2 is inflated by allowing random effects/random intercepts. The extent of the inflation involves τ^2, σ^2, and c.

Appendix 1 shows derivations for equations (12) and (13), and Appendix 2 applies those results to an arbitrary distribution of p_C, including URIM1 as a special case.

3.2 Relation of sample log-odds-ratio and its weight

We can write the IV weights as $\widehat{w}_i = \widehat{v}_i^{-2}/(\widehat{W}_{(i)} + \widehat{v}_i^{-2}) = [\widehat{W}_{(i)}\widehat{v}_i^2 + 1]^{-1}$, where $\widehat{W}_{(i)} = \sum_{j\neq i}\widehat{v}_j^{-2}$ is independent of $\widehat{v}_i^2$ and of $\widehat{\theta}_i$. Similarly, let $W_{(i)} = \sum_{j\neq i} v_j^{-2}$. We are interested in $\text{Cov}(\widehat{\theta}_i, \widehat{w}_i)$. Again using the law of total covariance

$$\text{Cov}(\widehat{\theta}_i, \widehat{w}_i) = \text{E}[\text{Cov}(\widehat{\theta}_i, \widehat{w}_i|\alpha_i, \theta_i)] + \text{Cov}(\theta_i, w_i^0)$$

where $w_i^0 = \text{E}(\widehat{w}_i|\alpha_i, \theta_i) = [W_{(i)}v_i^2 + 1]^{-1} + O(1/n_i)$. The first term of the covariance is of a smaller order than the second, so to order $1/n$, $\text{Cov}(\widehat{\theta}_i, \widehat{w}_i) = \text{Cov}(\theta_i, [W_{(i)}v_i^2 + 1]^{-1})$.

Expanding $[W_{(i)}v_i^2 + 1]^{-1}$ and taking into account the independence of $W_{(i)}$ from θ_i and v_i^2, we have

$$\text{Cov}(\theta_i, [W_{(i)}v_i^2 + 1]^{-1}) = -\frac{\text{E}(W_{(i)})}{(\text{E}(W_{(i)})\text{E}(v_i^2) + 1)^2}\text{Cov}(\theta_i, v_i^2) \tag{14}$$

where $\text{E}(W_{(i)}) = \sum_{j\neq i}\text{E}(v_j^{-2})$.

Equations (12) to (14) show that the choice of θ, the choice of p_C (through α), the choice of FIM versus RIM (through σ^2), the choice of fixed-effect versus random-effects model (through τ^2), and the choice of FIM1/RIM1 versus FIM2/RIM2 (through c) all affect the covariances between $\widehat{\theta}_i$ and their estimated weights, and result in varying biases in the estimated overall effect. In particular, when $n_T = n_C$, $\theta = 0$, and $c = 1/2$, the covariance is zero, so the $\widehat{\theta}_i$ and their estimated weights are almost independent, making the standard IV estimate of the overall effect unbiased when generated from FIM2/RIM2. On the other hand, the sign of the bias of the $\widehat{\theta}_i$ depends on the sign of $1 - 2p_T$, and the bias increases with an increase in τ^2 when generated from FIM1/RIM1.

4 Generation of sample sizes

Several authors[5,11] use constant study-level sample sizes, either equal or unequal, in all replications. More often, however, authors generate sample sizes from a uniform or normal distribution. Jackson et al.[4] use (mostly with $n_{iC} = n_{iT}$) sample sizes from discrete $U(50, 500)$. Langan et al.[14] use either constant and equal sample sizes within and across studies, or sample sizes from $U(40, 400)$ and $U(2000, 4000)$; Sidik and Jonkman[17] use $U(20, 200)$, and Abo-Zaid et al.[12] use $U(30, 100)$ and $U(30, 1000)$. Viechtbauer[16] generates study-level sample sizes ($n_i = n_{iC} = n_{iT}$) from $N(n, n/4)$ ($n/4$ is the variance) with $n = 10, 20, 40, 80, 160$. In an extensive simulation study for sparse data, Kuss[1] uses FIM1F and FIM1 along with a large number of fitting methods. He generates

both the number of studies K and their sample sizes n from log-normal distributions: with mean 0.65 and standard deviation 1.2 for rather small K, with log-normal mean 3.05 and log-normal standard deviation 0.97 for larger K, and with log-normal mean 4.615 and log-normal standard deviation 1.1 for sample sizes. He applies the ceiling function to the generated number and adds 1, and he limits the number of studies to a maximum of 100.

In general, if mutually independent random variables Y_i have a common distribution $F(\cdot)$, and $N \sim G_n(\cdot)$ is independent of the Y_i, the sum $Y_1 + \cdots + Y_N$ has a *c*ompound distribution.[21] A binomial distribution with probability p and a random number of trials is a compound Bernoulli distribution. The first two moments of such a distribution are $\mathrm{E}(X) = p\mathrm{E}(N)$ and $\mathrm{Var}(X) = p(1-p)\mathrm{E}(N) + p^2\mathrm{Var}(N)$. This variance is larger than the variance of the $\mathrm{Bin}(\mathrm{E}(N), p)$ distribution. Therefore, random generation of sample sizes produces an overdispersed binomial (compound Bernoulli) distribution for the control arm, and may also inflate, though in a more complicated way, the variance in the treatment arm.

In particular, when $N \sim N(\mathrm{E}(N), \sigma_n^2)$, the variance $\mathrm{Var}(X) = p(1-p)\mathrm{E}(N) + p^2\sigma_n^2$. And when $N \sim U(n_l, n_u)$, $\mathrm{Var}(X) = p(1-p)\mathrm{E}(N) + p^2(n_u - n_l)^2/12$.

4.1 Variances and covariances of estimated log-odds-ratios and their weights for random sample sizes

To calculate the variance of $\widehat{\theta}$ when sample sizes are random, we again use the law of total variance

$$\mathrm{Var}(\widehat{\theta}_i) = \mathrm{E}(\mathrm{Var}(\widehat{\theta}_i|n_i)) + \mathrm{Var}(\mathrm{E}(\widehat{\theta}_i|n_i))$$

The second term is $\mathrm{Var}(\theta) = 0$, and the first term is obtained by substituting $\mathrm{E}(n_{iC}^{-1})$ and $\mathrm{E}(n_{iT}^{-1})$ in equation (13).

Using the delta method

$$\mathrm{E}(N^{-1}) = (\mathrm{E}(N))^{-1}(1 + [\mathrm{CV}(N)]^2) \tag{15}$$

where the coefficient of variation, CV, is the ratio of the standard deviation of N to its mean. Therefore, to order $1/\mathrm{E}(N)$, random generation of sample sizes inflates the variance of $\widehat{\theta}$ if and only if the coefficient of variation of the distribution of sample sizes is of order 1. In the simulations of Viechtbauer,[16] where $\mathrm{Var}(N) = n/4$, $\mathrm{CV}(N) = O(1/\sqrt{n})$, so the variance is not inflated. In contrast, generating sample sizes from $N(n, n^2/4)$ would result in $\mathrm{CV} = 1/2$ and would inflate variance. (Use of such a combination of mean and variance, however, is unlikely to produce realistic sets of sample sizes, and the probability of generating a negative sample size exceeds 2%.)

The variance of a uniform distribution on an interval of width Δ centered at n_0 is $\Delta^2/12$, and its CV is $\Delta/(\sqrt{12}n_0)$. Therefore, $\mathrm{CV}(N)$ is of order 1 whenever the width of the interval is of the same order as its center. Hence, variance is inflated in the simulations by Jackson,[4] Langan et al.,[14] Sidik and Jonkman,[17] and Abo-Zaid et al.,[12] who all use wide intervals for n.

Similarly, we use the law of total covariance to calculate the covariance between $\widehat{\theta}_i$ and $\widehat{v}_i^2$

$$\mathrm{Cov}(\widehat{\theta}_i, \widehat{v}_i^2) = \mathrm{E}[\mathrm{Cov}(\widehat{\theta}_i, \widehat{v}_i^2|n_i)] + \mathrm{Cov}(\mathrm{E}(\widehat{\theta}_i), v_i^2|n_i)$$

The second term is zero, as $\mathrm{E}(\widehat{\theta}_i|n_i) = \theta$, which does not depend on n_i. So $\mathrm{Cov}(\widehat{\theta}_i, \widehat{v}_i^2)$ is obtained by substituting $\mathrm{E}(n_{iC}^{-1})$ and $\mathrm{E}(n_{iT}^{-1})$ in equation (12), and the covariances are affected only when $\mathrm{CV}(N)$ is of order 1.

5 Design of simulations

Our simulations keep the arm-level sample sizes equal in the K (= 5, 10, 30) studies. The control-arm probability $p_{iC} = .1, .4$. For the log-odds-ratios θ_i, we use equation (5) with $\theta = 0, 0.5, 1, 1.5$, and 2 and $\tau^2 = 0, 0.1, \ldots, 1$. We vary two components of the data-generating mechanism: the model (at five levels: FIM1, FIM2, RIM1, RIM2, and URIM1) and the arm-level sample sizes, n, centered at 40, 100, 250, and 1000 (constant, normally distributed, or uniformly distributed). We also vary the variance $\sigma^2 = 0.1, 0.4$ for RIM.

We keep the control-arm probabilities p_{iC} and the log-odds-ratios θ_i independent (i.e. $\rho = 0$ in the RIMs).

To make the normal and uniform distributions of sample sizes comparable, we center them at the same value n and equate their variances. If a normal distribution has variance σ_n^2, a uniform distribution with the same variance

Table 2. Summary of methods for meta-analysis of log-odds-ratios in our simulations.

Method	Features
Two-stage methods	
Inverse–variance–weighted average	
DerSimonian-Laird (DL) estimate of τ^2	All three assume $\widehat{v}_i^2 = v_i^2$
REML	and use
Mandel-Paule (MP) estimate of τ^2	Normal-based CIs
Kulinskaya-Dollinger (KD) estimate of τ^2	Normal-based CI
Sample-size-weighted estimator	
SSW	Constant weights
	t-based CI
General linear mixed models	
Binomial-normal random-effects model	
FIM2	Fixed intercept
RIM2	Random intercept

has interval width $\Delta_n = \sqrt{12\sigma_n^2}$. We set $\Delta_n = 1.1n$, resulting in $\mathrm{CV} = \Delta_n/(\sqrt{12}n) = 0.318$ and a squared CV of 0.101. Therefore, by equation (15), our simulations with random n inflate variances and covariances by 10% in comparison with simulations with fixed n. Wider intervals of n would inflate variances more, but in generating sample sizes from a corresponding normal distribution, we want negative sample sizes to have reasonably small probability. For our choice of Δ_n this probability is .0008. Unfortunately, we were still getting a small number of values below zero out of thousands of simulated values, so we additionally truncate the n values generated from a normal distribution at 10. Truncation happens with probability .009.

Similarly, for control-arm probabilities, even though using a normal distribution on the logit shifts the mean value of the control-arm probability, as discussed in Section 2.1.3, we can have equal variances on the probability scale by taking $\Delta_p = \sqrt{12[p_C^0(1-p_C^0)]^2\sigma_p^2}$ in comparator simulations.

For each generated dataset, we use a number of the two-stage methods for log-odds-ratio, including the best available method:[22,23] MP estimation of τ^2 with corresponding inverse–variance–weighted estimation of θ and a confidence interval based on the normal distribution. We also consider the performance of the GLMM methods based on FIM2 and RIM2 as implemented in **metafor**.[4,5,24] Finally, we include a weighted-average estimator of θ whose weights depend only on the studies' arm-level sample sizes: $w_i = n_{iT}n_{iC}/(n_{iT}+n_{ic})$.[22] We refer to this sample-size-weighted estimator as SSW. The accompanying confidence interval is based on the t distribution with $K-1$ degrees of freedom. Table 2 lists the analysis methods.

For each combination of the parameters and a data-generating mechanism, we generated data for 1000 simulated meta-analyses.

Table 3 shows the components of the simulations. For completeness we included the DerSimonian-Laird (DL), restricted maximum-likelihood (REML), MP, and Kulinskaya-Dollinger (KD) estimators of τ^2 with the corresponding inverse–variance–weighted estimators of θ and confidence intervals with critical values from the normal distribution. Bakbergenuly et al.[22] studied those inverse–variance–weighted estimators in detail. The results for the other IV-weighted estimators under the five data-generation mechanisms are similar to those for the Mandel-Paule estimator, so we do not include them in Section 6. Our preprints[25,26] give the full details. Among the estimators, FIM2 and RIM2 denote the estimators in the corresponding GLMMs.

6 Results of the simulations

In the figures that accompany the summaries of results, each plot shows a trace of a measure of the performance of an estimator (bias or coverage) for each of the five data-generation mechanisms. The horizontal variable is $\tau^2 \in [0,1]$. A row corresponds to a value of n (usually 40 or 100) and a combination of values of other parameters (e.g. p_C and σ^2 or θ). The figures illustrate the important patterns in the full sets of figures.[25,26] These preprints contain full simulation results for constant, normally distributed, and uniformly distributed sample sizes n.

Table 3. Components of the simulations.

Parameter	Values
K	5, 10, 30
n	40, 100, 250, 1000
θ	0, 0.5, 1, 1.5, 2
τ^2	0, 0.1, . . ., 1
p_C	.1, .4
σ^2	0.1, 0.4
Generation of n	
Constant	
Normal(n, $1.21n^2/12$)	
Uniform($n \pm 0.55n$)	
Generation of logit(p_{iC}) and logit(p_{iT})	
FIM1	Section 2.1.1
FIM2	Section 2.1.1
RIM1	Section 2.1.2
RIM2	Section 2.1.2
URIM1	Section 2.2
Estimation targets	Estimators
Bias in estimating τ^2	DL, REML, MP, KD, FIM2, RIM2
Bias in estimating θ	DL, REML, MP, KD, FIM2, RIM2, SSW
Coverage of θ	DL, REML, MP, KD, FIM2, RIM2, SSW (with $\widehat{\tau}^2_{KD}$ and t_{K-1} critical values)

As it turned out, the three methods of generating sample sizes produced essentially the same results. For two illustrative examples, compare the third and fourth rows of Figures 1 and 2. Thus, with those exceptions, the plots in the figures come from the results for constant n.

If the five data-generation mechanisms produce the same results, their traces in a plot coincide (except for minor variation). We focus on systematic departures from this null pattern (e.g. the traces separate into two groups). The specific performance measure may be important (e.g. an estimator has substantially greater bias when the data are generated by a certain mechanism). We generally give performance less emphasis, however, because our primary goal is to examine the consequences for inference of the choice of a data-generating method. Bakbergenuly et al.[22] have studied in detail the performance under FIM1 of the estimators other than the GLMM estimators based on FIM2 and RIM2.

6.1 Bias of $\widehat{\tau}^2_{MP}$

The estimated bias of $\widehat{\tau}^2_{MP}$ often varies among the data-generation mechanisms. In the most common single pattern, the traces versus τ^2 form two clusters: one for FIM2 and RIM2 and another for FIM1, RIM1, and URIM1, as shown in the first row of Figure 1. When $\sigma^2 = 0.1$ and $p_C = .1$, separations tend to become clearer as K increases, and they are most evident when $K = 30$. As n increases, the traces flatten and coalesce around 0 bias, becoming essentially flat when $n \geq 250$. As θ increases, the traces for FIM2 and RIM2 merge with the others and then emerge below them, and the whole set of traces flattens toward 0.

Changing only p_C, from .1 to .4 (Figure 2), produces traces that stay near 0. Groupings are not consistently visible. As θ increases, the reversal observed when $p_C = .1$ (particularly when $n = 40$ and $K = 30$) does not occur. Instead, the separation between the traces for FIM2 and RIM2 and those for FIM1, RIM1, and URIM1 increases because the latter mechanisms produce larger negative bias as τ^2 increases.

When the simulations use $\sigma^2 = 0.4$ instead of $\sigma^2 = 0.1$, the most noticeable differences (when $p_C = .1$ and $n \leq 100$ and, especially, $K = 30$) are substantially larger negative bias under URIM1 (compare the first two rows of Figure 1) and greater separation among the traces for the other mechanisms. The trace for URIM1 approaches

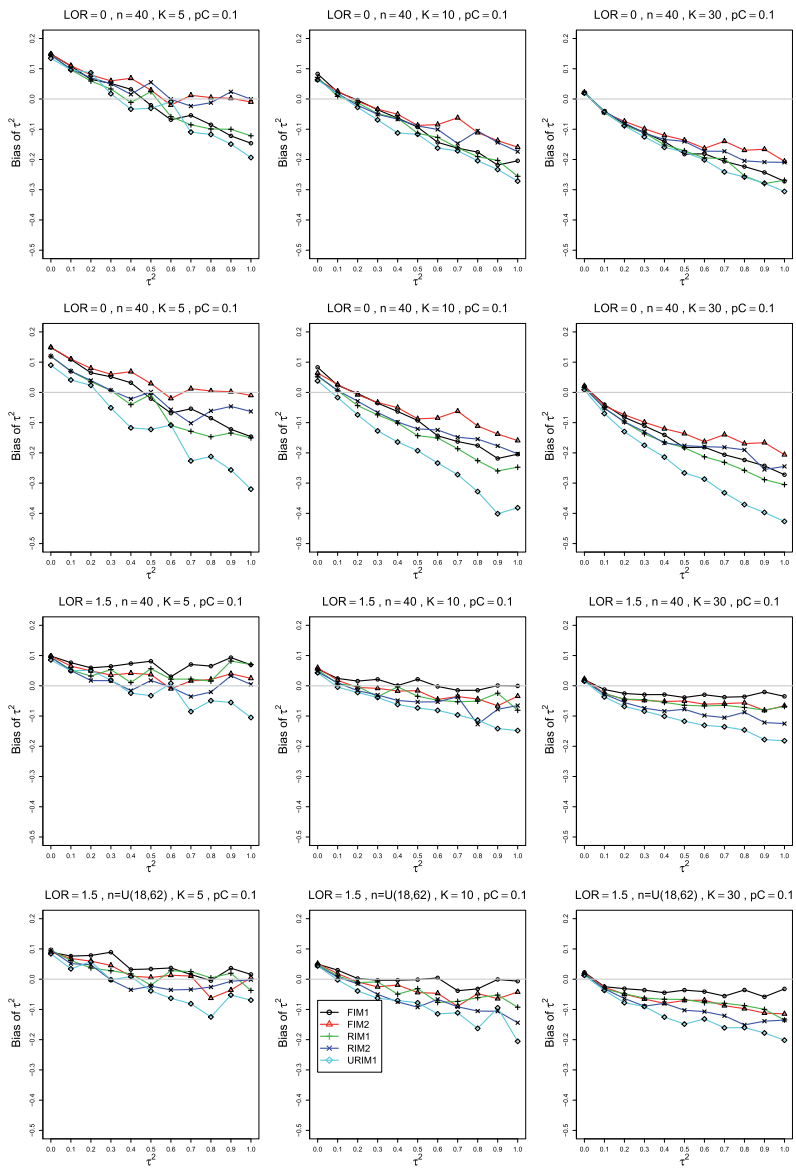

Figure 1. Bias in estimating the between-studies variance, τ^2, by $\widehat{\tau}^2_{MP}$ for $p_C = .1$, $\theta = 0$, $\sigma^2 = 0.1$ (top row); $\theta = 0$, $\sigma^2 = .4$ (second row); $\theta = 1.5$, $\sigma^2 = 0.4$ (bottom two rows). Sample sizes are constant $n = 40$ in the top three rows and uniformly distributed in the bottom row. The data-generation mechanisms are FIM1 (circle), FIM2 (triangle), RIM1 (plus), RIM2 (cross), and URIM1 (diamond). Light gray line at 0.

the others as θ increases (compare the second and third rows of Figure 1). This change in σ^2 produces little change in the patterns for $p_C = .4$.

Turning from the data-generation mechanisms to the bias, when $p_C = .1$ and $\theta = 0$, $\widehat{\tau}^2_{MP}$ has positive bias for small to moderate values of τ^2 and substantial negative bias when $K \geq 10$, increasing in τ^2. FIM1, RIM1 and URIM1 produce larger negative bias than FIM2 and RIM2 when $n = 40$. When sample sizes increase to $n = 100$, FIM2 and RIM2 have positive bias for $K \leq 10$, whereas for $K = 30$, FIM2 and RIM2 have almost no bias. Differences between data-generation mechanisms disappear by $n = 250$.

Negative bias at large τ^2 decreases with increasing θ. When $\theta \geq 1$, $K = 5$, and $n = 40$, $\widehat{\tau}^2_{MP}$ has a small positive bias, especially under RIM1, decreasing in K. For $K = 30$, FIM1 produces almost no bias, and other mechanisms result in small negative bias. Bias is almost absent when $n \geq 100$.

When $p_C = .4$ and $\theta = 0$, $\widehat{\tau}^2_{MP}$ has a small positive bias, somewhat increasing for larger τ^2. RIM2 and FIM2 produce somewhat more bias than the other mechanisms. When $p_C = .4$ and $\theta = 1.5$, FIM2 and RIM2 produce

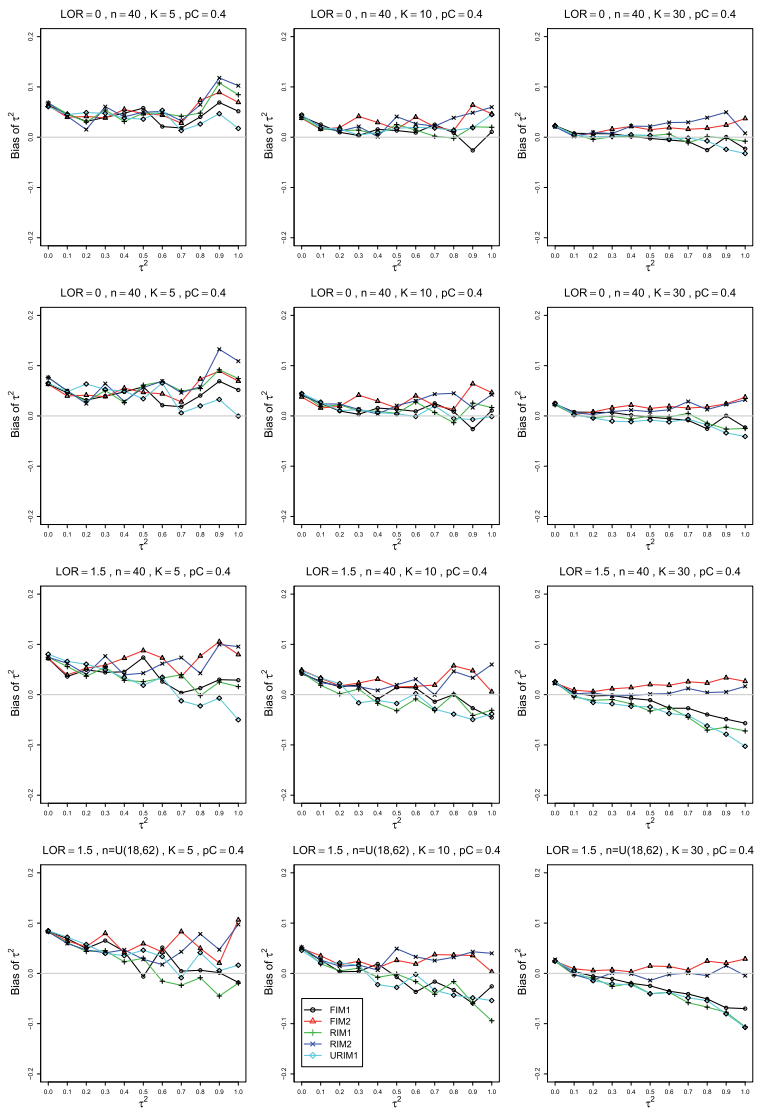

Figure 2. Bias in estimating the between-studies variance, τ^2, by $\widehat{\tau}^2_{MP}$ for $p_C = .4$, $\theta = 0$, $\sigma^2 = 0.1$ (top row); $\theta = 0$, $\sigma^2 = 0.4$ (second row); $\theta = 1.5$, $\sigma^2 = 0.4$ (bottom two rows). Sample sizes are constant $n = 40$ in the top three rows and uniformly distributed in the bottom row. The data-generation mechanisms are FIM1 (circle), FIM2 (triangle), RIM1 (plus), RIM2 (cross), and URIM1 (diamond). Light gray line at 0.

almost no bias for $K = 30$, and the rest produce negative bias for large τ^2. For $K \leq 10$, FIM2 and RIM2 produce positive bias, and FIM1, RIM1, and URIM1 produce positive bias for small to moderate values of τ^2 and negative bias for large values.

6.2 Bias of the estimators of τ^2 in the FIM2 and RIM2 GLMMs

Having used the FIM2 and RIM2 data-generation mechanisms, we examine the performance of the estimators in those GLMMs (in this section and in Sections 6.4 and 6.7).

6.2.1 Bias of $\widehat{\tau}^2_{FIM2}$

For the bias of $\widehat{\tau}^2_{FIM2}$, departures of the traces from the null pattern generally occur when $n = 40$ and occasionally when $n = 100$. In the most common departure, at larger τ^2, the traces for FIM1, RIM1, and URIM1 form one

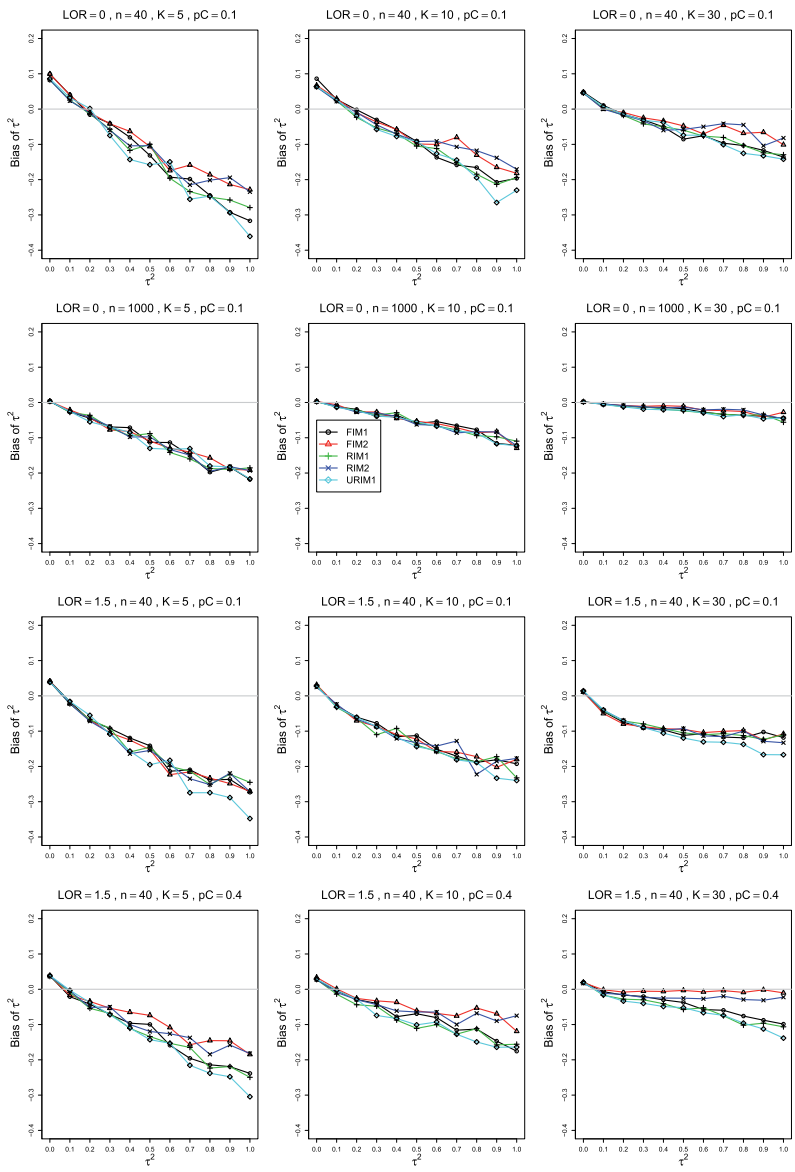

Figure 3. Bias of the estimator of the between-studies variance τ^2 in the FIM2 GLMM for $\sigma^2 = 0.4$, $p_C = .1$, $\theta = 0$ (top two rows); $p_C = .1$, $\theta = 1.5$ (third row); $p_C = .4$, $\theta = 1.5$ (bottom row); constant sample sizes $n = 40$ in rows 1, 3 and 4, and $n = 1000$ in row 2. The data-generation mechanisms are FIM1 (circle), FIM2 (triangle), RIM1 (plus), RIM2 (cross), and URIM1 (diamond). Light gray line at 0.

group, and those for FIM2 and RIM2 form another, closer to 0, as in the first row of Figure 3. This pattern tends to become clearer as K increases; it occurs more often when $K = 30$ than when $K = 10$ or $K = 5$.

The separation between FIM2 and RIM2 and FIM1, RIM1, and URIM1 tends to be clearer when $p_C = .4$ than when $p_C = .1$ (compare the third and fourth rows of Figure 3), and when $\sigma^2 = 0.4$ than when $\sigma^2 = 0.1$. When $p_C = .1$, the traces tend to be closer together as θ increases, but increasing θ has the opposite effect when $p_C = .4$.

In some situations, particularly when $p_C = .1$, $n = 40$, $K = 5$, and τ^2 is larger, the trace for URIM1 is visibly lower than the other traces (as in the third row of Figure 3).

The bias of $\hat{\tau}^2_{FIM2}$ under FIM2 and RIM2 relative to the other mechanisms (e.g. in the plot for $K = 30$ in the fourth row of Figure 3) is consistent with fitting the same GLMM that generated the data and with the fact that FIM2 is a submodel of RIM2.

Except at small τ^2, $\hat{\tau}^2_{FIM2}$ has negative bias, increasing with τ^2 (as in the first row of Figure 3, where the bias exceeds –20% when $\tau^2 = 1$) but decreasing as K increases. The bias remains large when θ is larger. It is worst when

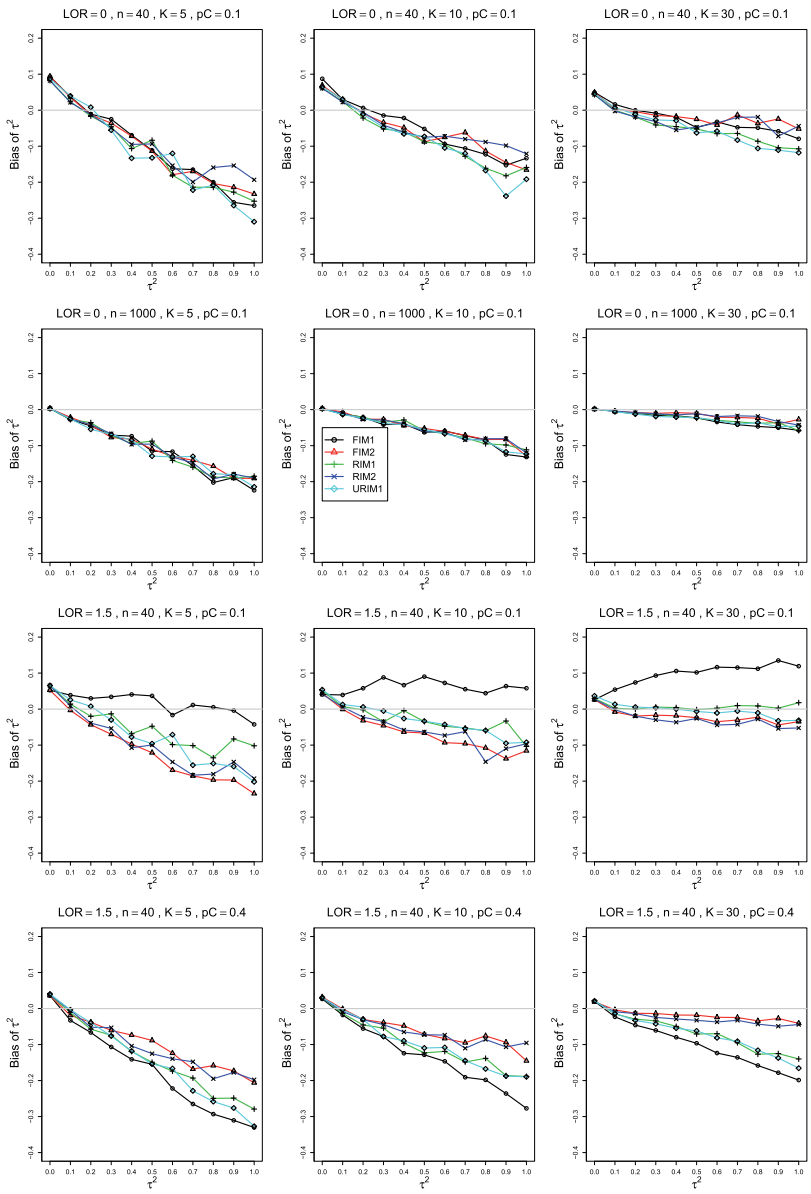

Figure 4. Bias of the estimator of the between-studies variance τ^2 in the RIM2 GLMM for $\sigma^2 = 0.4$, $p_C = .1$, $\theta = 0$ (top two rows); $p_C = .1$, $\theta = 1.5$ (third row); $p_C = .4$, $\theta = 1.5$ (bottom row); constant sample sizes $n = 40$ in rows 1, 3 and 4, and $n = 1000$ in row 2. The data-generation mechanisms are FIM1 (circle), FIM2 (triangle), RIM1 (plus), RIM2 (cross), and URIM1 (diamond). Light gray line at 0.

$K = 5$, even for $n = 1000$ (second row of Figure 3). When $n = 40$ and $K = 30$ and $p_C = .4$ (but not when $p_C = .1$), $\widehat{\tau}^2_{FIM2}$ is almost unbiased under FIM2 and RIM2 (compare the third and fourth rows of Figure 3).

6.2.2 Bias of $\widehat{\tau}^2_{RIM2}$

In summarizing the traces of the bias of $\widehat{\tau}^2_{RIM2}$, p_C and n play a larger role than for $\widehat{\tau}^2_{FIM2}$. The pattern in which FIM2 and RIM2 form a group, above the rest (FIM1, RIM1, and URIM1), is readily evident whenever $p_C = .4$, and it extends to smaller τ^2 (as in the fourth row of Figure 4). In addition to $n = 40$ the pattern is generally present when $n = 100$.

If $p_C = .1$, the same pattern is visible when $\sigma^2 = 0.1$, $\theta = 0$, $K = 30$, and n is 100 and 250. When θ is larger and $n = 40$, however, the traces follow a different, three-group pattern: FIM1 > RIM1/URIM1 > FIM2/RIM2 (as in the third row of Figure 4).

Contrary to what one might expect, the trace for RIM2 is not always closest to 0; indeed, it is sometimes fairly far from 0, particularly when $K < 30$ (as in the third and fourth rows of Figure 4).

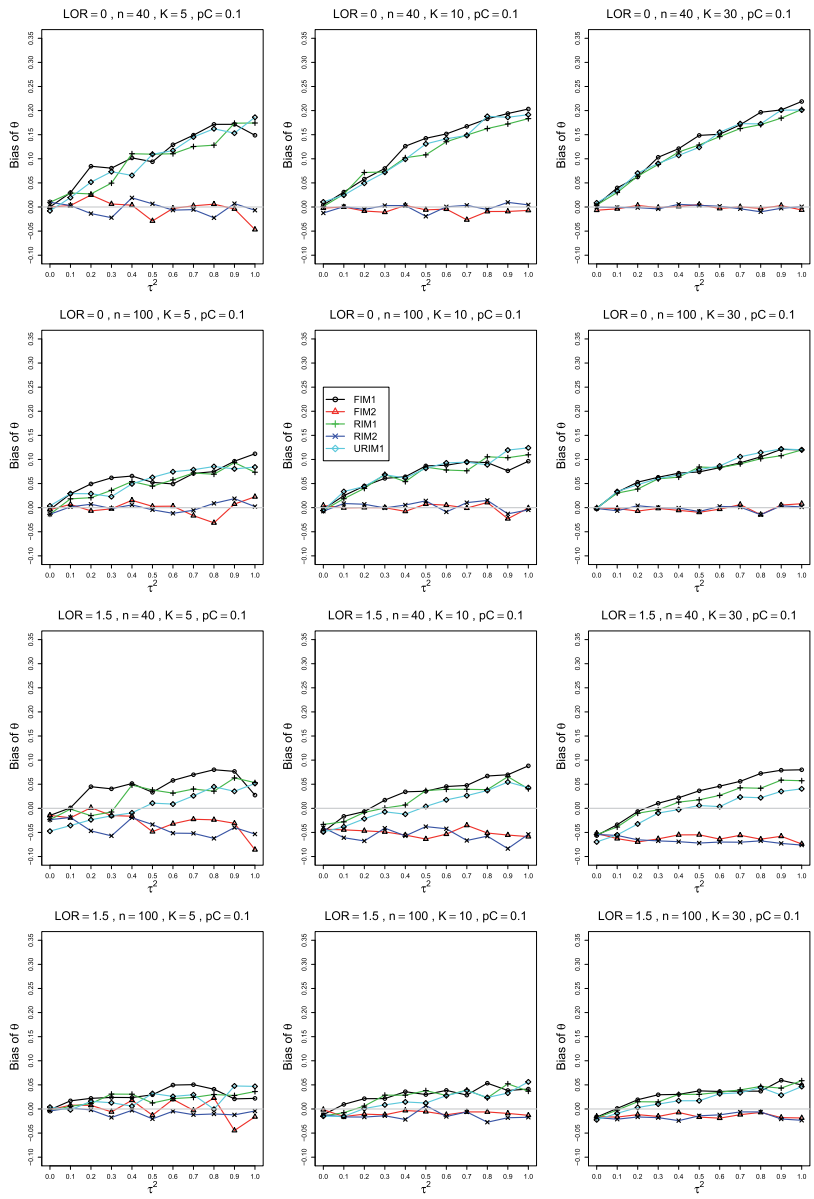

Figure 5 Bias in estimating the overall log-odds-ratio, θ, by $\widehat{\theta}_{MP}$ for $p_C = .1$, $\sigma^2 = 0.4$ (top three rows); $p_C = .4$, $\sigma^2 = 0.4$ (bottom row), and constant sample sizes $n = 40, 100$. Top two rows: $\theta = 0$; bottom two rows: $\theta = 1.5$. The data-generation mechanisms are FIM1 (circle), FIM2 (triangle), RIM1 (plus), RIM2 (cross), and URIM1 (diamond). Light gray line at 0.

Similar to $\widehat{\tau}^2_{FIM2}$, $\widehat{\tau}^2_{RIM2}$ has substantial negative bias when $K = 5$ or $K = 10$. When $K = 30$, $\widehat{\tau}^2_{RIM2}$ is nearly unbiased under RIM2 and FIM2, particularly when $p_C = .4$.

6.3 *Bias of* $\widehat{\theta}_{MP}$

The other IV-weighted estimators of θ have bias patterns similar to those of $\widehat{\theta}_{MP}$.

In the traces for the bias of $\widehat{\theta}_{MP}$, the patterns divide most clearly on p_C. When $p_C = .1$, no plot for $n \leq 250$ shows the null pattern, whereas when $p_C = .4$, departures from the null pattern are rare, occurring mainly when $n = 40$ and $K = 30$.

The first three rows of Figure 5 illustrate the behavior when $p_C = .1$. The traces for FIM2 and RIM2 form one group, in which the bias does not vary with τ^2; and those for FIM1, RIM1, and URIM1 form a second group, in which the bias increases with τ^2. Under FIM2 and RIM2 $\widehat{\theta}_{MP}$ is essentially unbiased when $\theta = 0$ (as in the first row

Table 4. Lower and upper bounds for p_{iC} (p_{CL} and p_{CU}) and for p_{iT} (p_{TL} and p_{TU}) values under URIM1.

p_C	σ^2	p_{CL}	p_{CU}	p_{TL}	p_{TU}
.1	0.1	.0507	.1493	.2830	.5646
.1	0.4	.0014	.1986	.0103	.6468
.4	0.1	.2685	.5315	.7306	.8934
.4	0.4	.1371	.6629	.5400	.9356

Note: Intervals of p_T are given for $b = 0$ and $\theta = 2$.

of Figure 5); but when $\theta > 0$, its bias is roughly -0.05 when $n = 40$ (as in the third row of Figure 5), decreasing to nearly 0 when $n = 100$. As n increases, the traces for FIM1, RIM1, and URIM1 flatten and also approach 0.

The fourth row of Figure 5 illustrates a situation, when $p_C = .4$, in which, for $K = 30$, $\widehat{\theta}_{MP}$ is nearly unbiased under FIM2 and RIM2 and has some negative bias under FIM1, RIM1, and URIM1, particularly when τ^2 is larger. Other such situations involve $\theta = 0$ or, mainly, $\theta = 2$. Ordinarily, however, $\widehat{\theta}_{MP}$ is essentially unbiased under all five data-generation mechanisms.

6.4 Bias of the estimators of θ in the FIM2 and RIM2 GLMMs

For the bias of $\widehat{\theta}_{FIM2}$ and $\widehat{\theta}_{RIM2}$, the patterns of the traces in Figures 6 and 7 strongly resemble those for $\widehat{\theta}_{MP}$. When $p_C = .1$, both estimators are essentially unbiased under FIM2 and RIM2, except for bias of $+0.01$ to $+0.03$ in $\widehat{\theta}_{FIM2}$ when $\theta \geq 1$ and $n = 40$. The behavior of the other group differs more clearly between $\widehat{\theta}_{FIM2}$ and $\widehat{\theta}_{RIM2}$: when $n = 40$ and $n = 100$, $\widehat{\theta}_{RIM2}$ usually has greater bias under FIM1 than under RIM1 or URIM1. (The plots for $\widehat{\theta}_{MP}$ show a suggestion of this behavior.)

When $p_C = .4$, both $\widehat{\theta}_{FIM2}$ and $\widehat{\theta}_{RIM2}$ are usually unbiased under all five data-generation mechanisms. The exceptions arise mainly when $n = 40$ (especially when $K = 30$). When $\theta = 0$, the traces for FIM1, RIM1, and URIM1 rise to around 0.02; when $\theta = 1.5$ or $\theta = 2$, those traces drop to around -0.02 or lower.

6.5 *Bias of the SSW estimator of* θ

Only a few situations show bias in $\widehat{\theta}_{SSW}$. Those involve $p_C = .1$. When $\theta = 0$, $n = 40$, and $K = 10$ and 30, the traces for FIM1, RIM1, and URIM1 are positive, rising to around 0.05 as τ^2 increases to 1 (first row of Figure 8).

A different pattern arises when $\theta = 2$ and $\sigma^2 = 0.4$; the trace for URIM1 is low, around -0.05 when $n = 40$ (and $K = 5$, 10, and 30) and around -0.02 when $n = 100$ and $K = 30$, shown in the third and fourth rows of Figure 8.

An explanation for this bias is that URIM1 may quite often produce extremely low or extremely high probabilities, as shown in Table 4. These probabilities may be even more extreme when τ^2 is large. Then the relevant binomial distributions produce more zero or n values. Adding 0.5 to these data introduces the observed biases. This does not happen when $\sigma^2 = 0.1$ because then the probabilities are far enough from 0 and 1.

6.6 *Coverage of the confidence interval for* θ *centered at* $\widehat{\theta}_{MP}$

The 95% confidence interval for θ centered at $\widehat{\theta}_{MP}$ uses normal critical values. The coverage of the confidence intervals based on the other IV-weighted estimators of θ has similar patterns.

With few exceptions, the patterns of the traces for coverage of the confidence interval based on $\widehat{\theta}_{MP}$ are similar for $p_C = .1$ and $p_C = .4$. When $K = 5$, all five start together above .95 at $\tau^2 = 0$. For $\tau^2 \geq 0.1$ they decrease and then level off below .95 (as illustrated in Figure S1 in the Supplementary Material). As K increases, that level approaches .95, but increasing n has the opposite effect, producing coverage like that shown in the second row of Figure S1. Exceptions occur when $\theta = 0$ and 0.5, $p_C = .1$, $n = 40$, and $K = 10$ and 30. Beyond a certain τ^2 the traces separate into two groups; FIM2 and RIM2 level off around .95, and FIM1, RIM1, and URIM1 continue to decrease. Other, similar exceptions occur when $\theta = 0$, $p_C = .1$, $n = 100$, and $K = 30$ and perhaps when $\theta = 2$, $p_C = .4$, $\sigma^2 = 0.1$, $n = 40$, and $K = 30$.

6.7 *Coverage of the confidence intervals for* θ *from the FIM2 and RIM2 GLMMs*

The coverage of the 95% confidence interval accompanying $\widehat{\theta}_{FIM2}$ generally resembles that of the confidence interval based on $\widehat{\theta}_{MP}$ (compare Figure S2 and Figure S1). The main difference is that for all values of θ, the traces in the plot for $n = 40$ and $K = 30$ separate into the two groups (as illustrated in the first row of Figure S2).

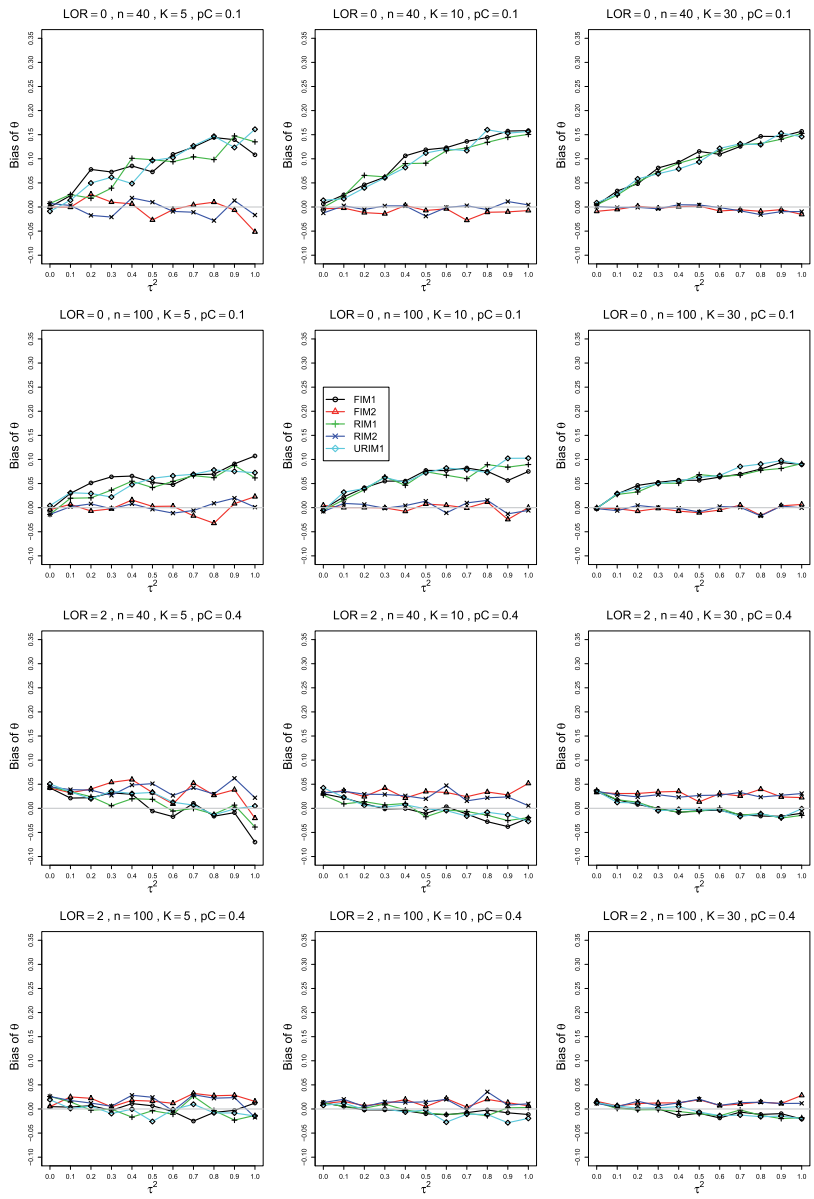

Figure 6. Bias of the estimator of the overall log-odds-ratio, θ, in the FIM2 GLMM when $\sigma^2 = 0.4$, constant sample sizes $n = 40, 100$, and $p_C = .1$ and $\theta = 0$ (top two rows) or $p_C = .4$ and $\theta = 2$ (bottom two rows). The data-generation mechanisms are FIM1 (circle), FIM2 (triangle), RIM1 (plus), RIM2 (cross), and URIM1 (diamond). Light gray line at 0.

The coverage of the confidence interval accompanying $\widehat{\theta}_{RIM2}$ has a surprising feature: When $p_C = .1$ and $n = 40$, the traces for the five data-generation mechanisms often differ substantially (as in the first and third rows of Figure S3). Coverage may be close to .95 when $\tau^2 = 0$, but it can decline to .60 and below when $\tau^2 = 1$. Coverage under FIM2 generally exceeds .90, and it improves as θ increases. When $p_C = .4$ or $n \geq 100$, coverage of θ is similar to that from the FIM2 GLMM.

6.8 Coverage of the confidence interval centered at the SSW estimator of θ

In all situations in our simulations, the traces for the coverage of the confidence interval centered at $\widehat{\theta}_{SSW}$ follow the null pattern. This favorable result makes it easy to summarize the coverage itself.

Coverage of the SSW interval exceeds .95 for small values of τ^2. When $p_C = .1$, $n = 40$, and $K = 5$, coverage is still too high at $\tau^2 = 1$ (first row of Figure S4); this excess decreases somewhat when $p_C = .4$ (third row of Figure

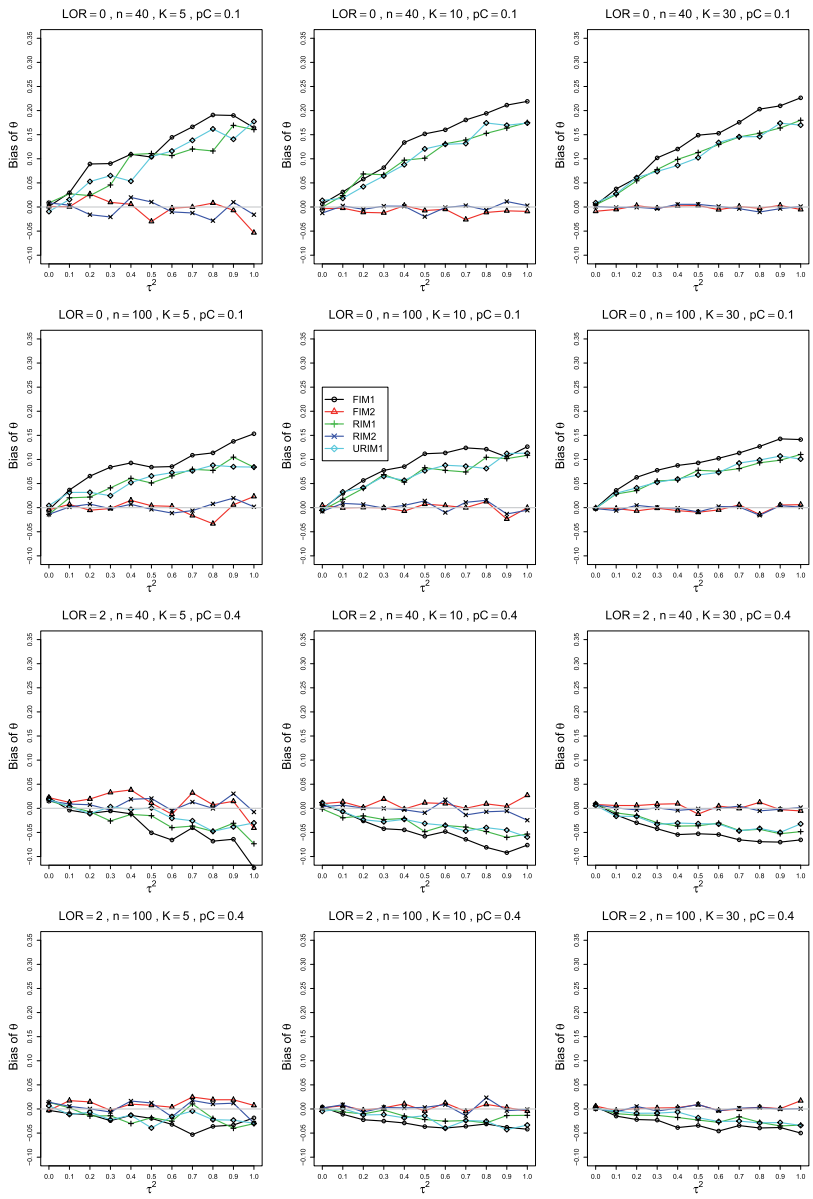

Figure 7. Bias of the estimator of the overall log-odds-ratio, θ, in the RIM2 GLMM when $\sigma^2 = 0.4$, constant sample sizes $n = 40, 100$, and $p_C = .1$ and $\theta = 0$ (top two rows) or $p_C = .4$ and $\theta = 2$ (bottom two rows). The data-generation mechanisms are FIM1 (circle), FIM2 (triangle), RIM1 (plus), RIM2 (cross), and URIM1 (diamond). Light gray line at 0.

S4). It decreases when $K = 10$, and coverage is close to nominal when $K = 30$. Coverage approaches nominal for lower values of τ^2 as the sample size increases. For $n = 1000$, coverage is above nominal only at $\tau^2 = 0$ (second and fourth rows of Figure S4). Coverage does not depend on σ^2 or θ.

6.9 Summary

Our simulations explored two main components of design: the data-generation mechanism and the distribution of study-level sample sizes. As we mentioned earlier, the second of these had essentially no impact on bias of estimators of τ^2, bias of estimators of θ, or coverage of confidence intervals for θ.

The five data-generation mechanisms often produced different results for at least one of those measures of performance. In the most frequent pattern, FIM2 and RIM2 yield similar results, and FIM1, RIM1, and URIM1 also yield results that are similar but different from those of FIM2 and RIM2. In some situations, URIM1 stands

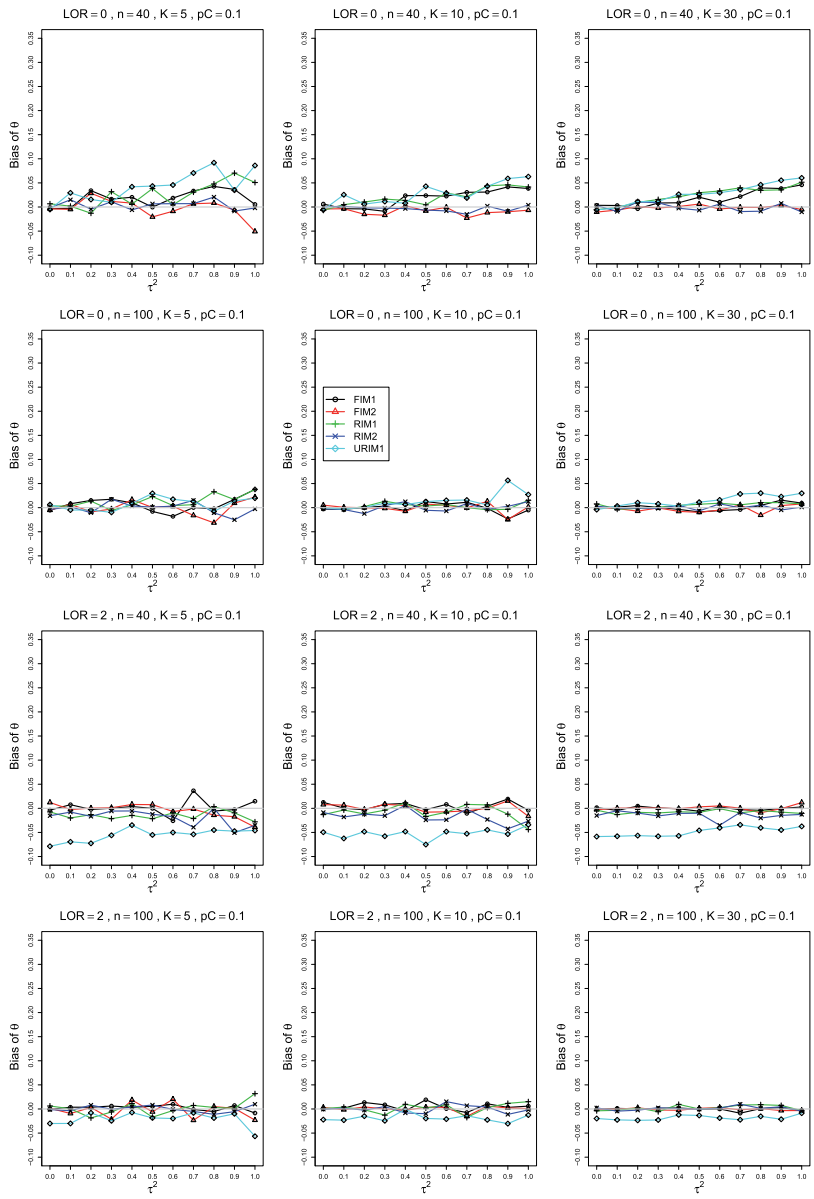

Figure 8. Bias of the SSW estimator of the overall log-odds-ratio, θ, for $p_C = .1$, $\sigma^2 = 0.4$, and constant sample sizes $n = 40$; 100. Top two rows, $\theta = 0$; bottom two rows, $\theta = 2$. The data-generation mechanisms are FIM1 (circle), FIM2 (triangle), RIM1 (plus), RIM2 (cross), and URIM1 (diamond). Light gray line at 0.

apart (e.g., for the bias of $\widehat{\tau}^2_{MP}$ and the bias of $\widehat{\theta}_{SSW}$), and so does FIM1 (for the bias of $\widehat{\tau}^2_{RIM2}$ and the bias of $\widehat{\theta}_{RIM2}$). For $K = 30$ Figure S3 shows a particularly unusual pattern, in which the traces for the five data-generation mechanisms are mostly separate.

In summary, except for the coverage of the SSW confidence interval and, in most situations, the bias of $\widehat{\theta}_{SSW}$, the choice of data-generation mechanism affects the results. These differences can complicate the process of integrating results from separate simulation studies.

7 Discussion

With the advent of powerful computers, the typical methodology paper in applied statistics has a standard structure. It proposes a new method, sometimes but not necessarily provides a mathematical derivation of its

properties, and then uses simulation to demonstrate, usually successfully, that the new method is superior to previous methods.

Using methods for meta-analysis of odds ratios as an example, we aimed to compare various ways of generating data in simulations. In the literature, we identified five methods of generating odds ratios. We combined them with three methods of generating sample sizes, and we derived the statistical properties of inverse–variance–weighted estimators of the overall log-odds-ratio, θ, under these methods of data generation. In particular, we derived, to order $1/n$, the biases and the variances of the inverse–variance–weighted estimators of θ.

We simulated data from the combinations of data-generation mechanism and sample-sizes method, and we compared the resulting estimates of the performance in estimating τ^2 and θ of four methods of meta-analysis: inverse–variance weighting (represented by the Mandel-Paule method), the FIM2 and RIM2 GLMMs, and SSW (for θ only). Our results show that the properties of various methods and the recommendations on their use greatly depend on the data-generation mechanism.

Our theoretical derivations showed that, under FIM1/RIM1/URIM1, the IV-weighted estimators of θ should have positive bias for small values of $p_T < 1/2$ and negative bias for $p_T > 1/2$. On the other hand, under FIM2/RIM2 these estimators should be approximately unbiased when $\theta = 0$. Our simulations (Figure 5) confirmed these findings.

Importantly, results of our simulations also show very similar behavior for the FIM2 and RIM2 GLMM estimators of θ (Figures 6 and 7). This finding is not very astonishing. Regardless of the hype that concerns use of GLMMs in meta-analysis, GLMs (and GLMMs) are asymptotic methods. The maximum-likelihood equations used in GLMs for binary data (Section 4.4 in McCullagh and Nelder[27]) are weighted-least-squares equations with inverse–variance weights. For this reason, the GLMMs result in quite considerable biases in meta-analysis of odds-ratios, as demonstrated by our simulations and by Bakbergenuly and Kulinskaya.[5]

The SSW estimator of θ had considerably less bias, but even for this estimator the data-generation mechanism mattered, as URIM1 produced more-biased results (Figure 8).

Differences in the behavior of moment-based estimators of τ^2 such as $\widehat{\tau}^2_{MP}$ under various data-generation mechanisms (Figures 1 and 2) have the same explanation as those for estimators of θ. These estimators are derived from the Q statistic, which is affected by the correlation between the effects and the weights.

Even though wider, t-based confidence intervals[17,28,29] would somewhat improve coverage of θ, differences in coverage are due perhaps more to the centering of the intervals at very biased estimators. These biases are so large that they obscured the results of inflated variance in RIM methods. We also did not observe differences associated with random generation of sample sizes, perhaps because we used relatively tight intervals for them.

Finally, an interesting question is whether particular estimation methods work better when the data are generated exactly from the assumed model. Counterintuitively, the answer is no. In the majority of our simulations, generation under FIM2/RIM2 resulted in better estimation by all methods. But the RIM2 GLMM produced confidence intervals for θ that had much better coverage when the data were generated under FIM2, and really bad coverage otherwise.

What method(s) of meta-analysis should be used in practice, where we can never be certain of the true data-generating mechanism? In estimating θ, SSW provides the lowest biases and coverage that is correct but rather conservative and appears to be robust to the data-generation mechanism. This advantage will be shared by other methods that use constant weights.

As a more robust alternative in the two-stage random-effects model, Henmi and Copas[30] and, independently, Stanley and Doucouliagos[31] use an inverse–variance–weighted fixed-effect (FE) estimator as the center of the CI for θ. Our results show that the FE estimator of θ is also biased and will be affected by the simulation method.

Our findings are not surprising when put in a wider context. In pursuit of the effect of interest, we often neglect nuisance parameters that are sometimes only implicitly present in our models. However, when the sufficient statistics include these nuisance parameters, their distribution matters. Different distribution assumptions for the nuisance parameters should and do result in different properties of the estimators of interest. This influence directly parallels the effects of choice of prior distribution on the properties of the increasingly common Bayesian variants of the two-stage and GLM meta-analytic methods.[8,32,33] One solution may be to try to develop minimax procedures that would minimize possible biases. Another solution is the use of procedures that are robust to a wide class of distributions for nuisance parameters.

We demonstrated substantial effects of data-generating mechanisms on the inference in meta-analysis of odds-ratios. These complications are not restricted to binary data, and they make it difficult to rely on any single simulation in choosing methods. Careful, resourceful effort may lead to a battery of designs that, collectively, approximates the mechanisms underlying the data in actual meta-analyses. In any event, simulations should be

designed with the awareness of the possible effects of design choices, and quite a few recommendations may need to be revised.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Economic and Social Research Council [grant number ES/L011859/1].

Data availability statement

Our full simulation results are available in Kulinskaya et al.[25,26]

ORCID iD

Elena Kulinskaya https://orcid.org/0000-0002-9843-1663

Supplemental material

Supplemental material for this article is available online.

References

1. Kuss O. Statistical methods for meta-analyses including information from studies without any events—add nothing to nothing and succeed nevertheless. *Stat Med* 2015; **34**: 1097–1116.
2. Turner RM, Omar RZ, Yang M, et al. A multilevel model framework for meta-analysis of clinical trials with binary outcomes. *Stat Med* 2000; **19**: 3417–3432.
3. Stijnen T, Hamza TH and Özdemir P. Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. *Stat Med* 2010; **29**: 3046–3067.
4. Jackson D, Law M, Stijnen T, et al. A comparison of seven random-effects models for meta-analyses that estimate the summary odds ratio. *Stat Med* 2018; **37**: 1059–1085.
5. Bakbergenuly I and Kulinskaya E. Meta-analysis of binary outcomes via generalized linear mixed models: a simulation study. *BMC Med Res Methodol* 2018; **18**: 70.
6. Piaget-Rossel R, Taffé P. A pseudo-likelihood approach for the meta-analysis of homogeneous treatment effects: exploiting the information contained in single-arm and double-zero studies. *J Stat: Adv Theory Appl* 2019; **21**: 91–117.
7. Rice K and Higgins JPT and Lumley T. A re-evaluation of fixed effect(s) meta-analysis. *J Royal Stat Soc: Series A (Stat Soc)* 2018; **181**: 205–227.
8. Dias S, Sutton AJ, Ades AE, et al. Evidence synthesis for decision making 2: A generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. *Med Decis Mak* 2013; **33**: 607–617.
9. Sidik K and Jonkman JN. A simple confidence interval for meta-analysis. *Stat Med* 2002; **21**: 3153–3159.
10. Platt RW, Leroux BG and Breslow N. Generalized linear mixed models for meta-analysis. *Stat Med* 1999; **18**: 643–654.
11. Cheng J, Pullenayegum E, Marshall JK, et al. 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* 2016; **6**: e010983.
12. Abo-Zaid G, Guo B, Deeks JJ, et al. Individual participant data meta-analyses should not ignore clustering. *J Clin Epidemiol* 2013; **66**: 865–873.
13. Kosmidis I, Guolo A and Varin C. Improving the accuracy of likelihood-based inference in meta-analysis and meta-regression. *Biometrika* 2017; **104**: 489–496.
14. Langan D, Higgins JPT, Jackson D, et al. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Res Synthesis Meth* 2019; **10**: 83–98.
15. Bakbergenuly I and Kulinskaya E. Beta-binomial model for meta-analysis of odds ratios. *Stat Med* 2017; **36**: 1715–1734.
16. Viechtbauer W. Confidence intervals for the amount of heterogeneity in meta-analysis. *Stat Med* 2007; **26**: 37–52.
17. Sidik K and Jonkman JN. A comparison of heterogeneity variance estimators in combining results of studies. *Stat Med* 2007; **26**: 1964–1981.
18. Nagashima K, Noma H and Furukawa TA. Prediction intervals for random-effects meta-analysis: a confidence distribution approach. *Stat Meth Med Res* 2019; **28**: 1689–1702.
19. Piaget-Rossel R. *Meta-analysis of rare events: the challenge of combining the lack of information*. PhD Thesis. The University of Lausanne Open Archive, 2020, https://serval.unil.ch/resource/serval:BIB_6FFCED413301.P002/REF.pdf.

20. Li Y, Shi L and Roth H. The bias of the commonly-used estimate of variance in meta-analysis. *Commun Stat Theory Meth* 1994; **23**: 1063–1085.
21. Grubbström RW and Tang O. The moments and central moments of a compound distribution. *Eur J Operation Res* 2006; **170**: 106–119.
22. Bakbergenuly I, Hoaglin DC and Kulinskaya E. Methods for estimating between-study variance and overall effect in meta-analysis of odds-ratios. *Res Synthesis Meth* 2020; **11**: 426–442.
23. Veroniki AA, Jackson D, Viechtbauer W, et al. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Res Synthesis Meth* 2016; **7**: 55–79.
24. Viechtbauer W. Conducting meta-analyses in R with the **metafor** package. *J Stat Software* 2010; **36**: 3.
25. Kulinskaya E, Hoaglin DC and Bakbergenuly I. Exploring consequences of simulation design for apparent performance of statistical methods. 1: Results from simulations with constant sample sizes, 2020. eprint *arXiv: 2006.16638* [stat.ME].
26. Kulinskaya E, Hoaglin DC and Bakbergenuly I. Exploring consequences of simulation design for apparent performance of statistical methods. 2: Results from simulations with normally and uniformly distributed sample sizes. eprint *arXiv: 2007.05354* [stat.ME].
27. McCullagh P and Nelder JA. *Generalized linear models*. 2nd ed. London, UK: Chapman & Hall, 1989.
28. Hartung J and Knapp G. A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Stat Med* 2001; **20**: 3875–3889.
29. Röver C, Knapp G and Friede T. Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies. *BMC Med Res Methodol* 2015; **15**: 99.
30. Henmi M and Copas JB. Confidence intervals for random effects meta-analysis and robustness to publication bias. *Stat Med* 2010; **29**: 2969–2983.
31. Stanley TD and Doucouliagos H. Neither fixed nor random: weighted least squares meta-analysis. *Stat Med* 2015; **34**: 2116–2127.
32. Friede T, Röver C, Wandel S, et al. Meta-analysis of few small studies in orphan diseases. *Res Synthesis Meth* 2017; **8**: 79–91.
33. Turner RM, Jackson D, Wei Y, et al. Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. *Stat Med* 2015; **34**: 984–998.

Appendix 1. Derivations for equations (12) and (13)

This appendix gives derivations for equations (12) and (13). From equations (3) and (4)

$$\widehat{\theta}_i = \text{logit}(\widehat{p}_{iT}) - \text{logit}(\widehat{p}_{iC})$$

and

$$\widehat{\text{Var}}(\widehat{\theta}_i) = \widehat{v}_i^2 = \frac{1}{n_{iT}}\left(\frac{1}{\widehat{p}_{iT}} + \frac{1}{1-\widehat{p}_{iT}}\right) + \frac{1}{n_{iC}}\left(\frac{1}{\widehat{p}_{iC}} + \frac{1}{1-\widehat{p}_{iC}}\right)$$

In Section 3.1, we rewrote equation (10) as

$$\text{Cov}(\widehat{\theta}_i, \widehat{v}_i^2) = \text{E}[\text{Cov}(\widehat{\theta}_i, \widehat{v}_i^2 | p_{iC}, p_{iT})] + \text{Cov}(\theta_i, v_i^2)$$

We first calculate $\text{Cov}(\widehat{\theta}_i, \widehat{v}_i^2 | p_{iC}, p_{iT})$. Since $\widehat{p}_{iT}$ and $\widehat{p}_{iC}$ are conditionally independent, we expand their terms of $\widehat{\theta}_i$ and $\widehat{v}_i^2$ separately

$$\text{logit}(\widehat{p}) = \text{logit}(p) + (\widehat{p} - p)\left(\frac{1}{p} + \frac{1}{1-p}\right) - \frac{(\widehat{p}-p)^2}{2}\left(\frac{1}{p^2} - \frac{1}{(1-p)^2}\right) + \cdots$$

and (omitting the $1/n$)

$$\frac{1}{\widehat{p}}+\frac{1}{1-\widehat{p}}=\frac{1}{p}+\frac{1}{1-p}+(\widehat{p}-p)\left(\frac{1}{(1-p)^2}-\frac{1}{p^2}\right)+(\widehat{p}-p)^2\left(\frac{1}{(1-p)^3}+\frac{1}{p^3}\right)$$

For each of T and C, the conditional covariance of these two terms has the form

$$\mathrm{Cov}(a(\widehat{p}-p)+b(\widehat{p}-p)^2, c(\widehat{p}-p)+d(\widehat{p}-p)^2|p)=ac\mathrm{Var}(\widehat{p}|p)+O(1/n^2)$$

because the variance of the binomial distribution is of order $1/n$ and the higher central moments are at most $O(1/n^2)$. For $a=[p(1-p)]^{-1}$ and $c=(p^2-(1-p)^2)/[p^2(1-p)^2]=(2p-1)/[p^2(1-p)^2]$

$$\mathrm{Cov}\left(\mathrm{logit}(\widehat{p}),\frac{1}{\widehat{p}}+\frac{1}{1-\widehat{p}}|p\right)=\frac{1}{n}\left[\frac{2p-1}{p^2(1-p)^2}\right]$$

Combining the results for T and C (and restoring the $1/n$) yields

$$\mathrm{Cov}\left(\widehat{\theta}_i,\widehat{v}_i^2|p_{iC},p_{iT}\right)=\frac{1}{n_{iT}^2}\left[\frac{2p_{iT}-1}{p_{iT}^2(1-p_{iT})^2}\right]-\frac{1}{n_{iC}^2}\left[\frac{2p_{iC}-1}{p_{iC}^2(1-p_{iC})^2}\right]$$

which is only $O(1/n^2)$, so we do not need to calculate its expectation.

Next we calculate the second term of equation (10), $\mathrm{Cov}(\theta_i, v_i^2)$. Under the RIM, equation (7)

$$\begin{aligned}
p_{iT}^{-1}&=1+\exp(-(\alpha+u_i+\theta+(1-c)b_i)),\\
(1-p_{iT})^{-1}&=1+\exp(\alpha+u_i+\theta+(1-c)b_i),\\
p_{iC}^{-1}&=1+\exp(-(\alpha+u_i-cb_i)),\\
(1-p_{iC})^{-1}&=1+\exp(\alpha+u_i-cb_i)
\end{aligned}\tag{16}$$

Substituting these expressions in equation (4) yields

$$v_i^2=n_{iC}^{-1}[e^{-(\alpha+u_i-cb_i)}+2+e^{\alpha+u_i-cb_i}]+n_{iT}^{-1}[e^{-(\alpha+u_i+\theta+(1-c)b_i)}+2+e^{\alpha+u_i+\theta+(1-c)b_i}]$$

Because $\theta_i=\theta+b_i$, to complete the calculation, we need covariances of the form $\mathrm{Cov}(b,e^{dx})$. As $e^{dx}=e^{dx_0}+d(x-x_0)e^{dx_0}+(x-x_0)^2d^2e^{dx_0}/2+\cdots$, where $x_0=\mathrm{E}(x)$, we have

$$\mathrm{Cov}(b,e^{dx})=d\mathrm{Cov}(b,x)e^{dx_0}+\mathrm{Cov}(b,(x-x_0)^2)d^2e^{dx_0}/2+\cdots\tag{17}$$

For simplicity, we assume that u_i and b_i are independent. Then, to order $1/n$

$$\mathrm{Cov}(\theta_i,v_i^2)=n_{iC}^{-1}c\tau^2[e^{-\alpha}-e^{\alpha}]-n_{iT}^{-1}(1-c)\tau^2[e^{-\alpha-\theta}-e^{\alpha+\theta}]\tag{18}$$

In particular, when $c=1/2$, $\theta=0$ and $n_{iT}=n_{iC}$, $\mathrm{Cov}(\theta_i,v_i^2)=0$. Defining $p_C=\mathrm{expit}(\alpha)$ and $p_T=\mathrm{expit}(\alpha+\theta)$ yields $e^{-\alpha}-e^{\alpha}=p_C^{-1}-(1-p_C)^{-1}$ and $e^{-\alpha-\theta}-e^{\alpha+\theta}=p_T^{-1}-(1-p_T)^{-1}$ and the equivalent expression, equation (12)

$$\mathrm{Cov}(\theta_i,v_i^2)=\frac{c\tau^2}{n_{iC}}\left[\frac{1-2p_C}{p_C(1-p_C)}\right]-\frac{(1-c)\tau^2}{n_{iT}}\left[\frac{1-2p_T}{p_T(1-p_T)}\right]$$

Similarly, from equation (11), $\mathrm{Var}(\widehat{\theta}_i) = \mathrm{E}(v_i^2) + \tau^2$. To calculate $\mathrm{E}(v_i^2)$, we need expansions in two variables, u_i and b_i. We omit the grubby details, observe that $e^{-\alpha} + e^{\alpha} = [p_C(1-p_C)]^{-1} - 2$ and $e^{-\alpha-\theta} + e^{\alpha+\theta} = [p_T(1-p_T)]^{-1} - 2$, and express the full variance of $\widehat{\theta}_i$ in terms of p_C and p_T

$$\begin{aligned}\mathrm{Var}(\widehat{\theta}_i) = {} & [n_{iT}p_T(1-p_T)]^{-1} + [n_{iC}p_C(1-p_C)]^{-1} + \tau^2 \\ & + \left(\sigma^2 + (1-c)^2\tau^2 + 2(1-c)\rho\sigma\tau\right)[2n_{iT}]^{-1}\left([p_T(1-p_T)]^{-1} - 2\right) \\ & + \left(\sigma^2 + c^2\tau^2 - 2c\rho\sigma\tau\right)[2n_{iC}]^{-1}\left([p_C(1-p_C)]^{-1} - 2\right)\end{aligned} \tag{19}$$

This is equation (13).

Appendix 2. Arbitrary distribution for p_c

To calculate $\mathrm{Var}(\widehat{\theta}_i)$ for an arbitrary distribution of p_{iC} but assuming that the p_{iC} and the normal random effects b_i are independent, we proceed as in equation (19) to obtain (for $c=0$ and substituting $p_C^0 = \mathrm{expit}(\mathrm{E}(\mathrm{logit}(p_{iC})))$, $p_T^0 = \mathrm{expit}(\mathrm{E}(\mathrm{logit}(p_{iC})) + \theta)$ and $\mathrm{Var}(\mathrm{logit}(p_{iC}))$ for p_C, p_T and σ^2, respectively)

$$\begin{aligned}\mathrm{Var}(\widehat{\theta}_i) = {} & [n_{iT}p_T^0(1-p_T^0)]^{-1} + [n_{iC}p_C^0(1-p_C^0)]^{-1} + \tau^2 \\ & + \left(\mathrm{Var}(\mathrm{logit}(p_{iC})) + \tau^2\right)[2n_{iT}]^{-1}\left([p_T^0(1-p_T^0)]^{-1} - 2\right) \\ & + (\mathrm{Var}(\mathrm{logit}(p_{iC})))[2n_{iC}]^{-1}\left([p_C^0(1-p_C^0)]^{-1} - 2\right)\end{aligned}$$

$\mathrm{Cov}(\theta_i, v_i^2)$ is the same as in equation (12) with $c=0$ and $p_T^0 = \mathrm{expit}(\mathrm{E}(\mathrm{logit}(p_{iC})) + \theta)$, i.e.

$$\mathrm{Cov}(\theta_i, v_i^2) = -\frac{\tau^2}{n_{iT}}\left[\frac{1-2p_T^0}{p_T^0(1-p_T^0)}\right] \tag{20}$$

This should produce substantial positive bias in $\widehat{w}_i\widehat{\theta}_i$ for small $p_T^0 < 1/2$, and negative bias for $p_T^0 > 1/2$.

To evaluate the first two moments of $\mathrm{logit}(p_{iC})$, we use the standard delta method, as in equation (8). Let the mean and the variance of the distribution of p_{iC} be p_C and σ_C^2. For $h(x) = \mathrm{logit}(x)$, the derivatives of $h(\cdot)$ at p_C are

$$h'(p_C) = \frac{1}{p_C(1-p_C)} \text{ and } h''(p_C) = -\left[\frac{1}{p_C^2} + \frac{1}{(1-p_C)^2}\right]$$

Hence

$$\mathrm{E}(\mathrm{logit}(p_{iC})) = \mathrm{logit}(p_C) - \left[\frac{1}{p_C^2} + \frac{1}{(1-p_C)^2}\right]\sigma_C^2/2$$

and

$$\mathrm{Var}(\mathrm{logit}(p_{iC})) = \frac{1}{p_C^2(1-p_C)^2}\sigma_C^2$$

The expected value decreases, and the variance increases, with σ_C^2, the variance of the distribution of p_{iC}.

For a uniform distribution on an interval of width Δ centered at p_C, $\sigma_C^2 = \Delta^2/12$.