

RESEARCH ARTICLE  

# Meta-analysis with Jeffreys priors: Empirical frequentist properties

Maya B. Mathur

Quantitative Sciences Unit and Department of Pediatrics, Stanford University, Palo Alto, CA, USA  
Email: [mmathur@stanford.edu](mailto:mmathur@stanford.edu)

**Received:** 8 March 2024; **Revised:** 14 July 2024; **Accepted:** 6 September 2024

**Keywords:** meta-analysis; bayesian; simulation study; Firth correction; bayesian methods; small-sample estimation; simulations

## Abstract



In small meta-analyses (e.g., up to 20 studies), the best-performing frequentist methods can yield very wide confidence intervals for the meta-analytic mean, as well as biased and imprecise estimates of the heterogeneity. We investigate the frequentist performance of alternative Bayesian methods that use the invariant Jeffreys prior. This prior has the usual Bayesian motivation, but also has a purely frequentist motivation: the resulting posterior modes correspond to the established Firth bias correction of the maximum likelihood estimator. We consider two forms of the Jeffreys prior for random-effects meta-analysis: the previously established “Jeffreys1” prior treats the heterogeneity as a nuisance parameter, whereas the “Jeffreys2” prior treats both the mean and the heterogeneity as estimands of interest. In a large simulation study, we assess the performance of both Jeffreys priors, considering different types of Bayesian estimates and intervals. We assess point and interval estimation for both the mean and the heterogeneity parameters, comparing to the best-performing frequentist methods. For small meta-analyses of binary outcomes, the Jeffreys2 prior may offer advantages over standard frequentist methods for point and interval estimation of the mean parameter. In these cases, Jeffreys2 can substantially improve efficiency while more often showing nominal frequentist coverage. However, for small meta-analyses of continuous outcomes, standard frequentist methods seem to remain the best choices. The best-performing method for estimating the heterogeneity varied according to the heterogeneity itself. Röver & Friede’s R package *bayesmeta* implements both Jeffreys priors. We also generalize the Jeffreys2 prior to the case of meta-regression.

## Highlights

- What is already known: The best-performing frequentist methods for random-effects meta-analysis can be highly imprecise in small meta-analyses, and can provide biased estimates of the heterogeneity.
- What is new: We conduct a large simulation study evaluating two forms of the Jeffreys prior for meta-analysis, which correspond to the Firth bias correction to the maximum likelihood estimator.
- Potential impact for *RSM* readers: For small meta-analyses of binary outcomes, the Jeffreys2 prior may offer advantages over standard frequentist methods for point and interval estimation for the mean parameter.

## 1. Introduction

Standard random-effects meta-analysis involves estimating the heterogeneity of studies’ population effects (e.g., their standard deviation) and obtaining an inverse-variance-weighted estimate of the meta-analytic mean, in which studies’ weights depend on the estimated heterogeneity.<sup>1</sup> Commonly

  This article was awarded Open Data and Open Materials badges for transparent practices. See the Data availability statement for details.

© The Author(s), 2025. Published by Cambridge University Press on behalf of The Society for Research Synthesis Methodology. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

used methods to estimate the heterogeneity include semiparametric method-of-moments estimators<sup>1–5</sup> and parametric likelihood-based estimators.<sup>1,6</sup> The theoretical justification for these methods relies on asymptotics, yet in some scientific disciplines, the majority of meta-analyses include a relatively small number of studies. Meta-analyses of healthcare interventions in the Cochrane Database for Systematic Reviews include a median of only 3 studies (75th percentile: 6, 90th percentile: 10).<sup>7</sup> In psychology, meta-analyses published in *Psychological Bulletin* include a median of 12 studies, though some meta-analyses are much larger (75th percentile: 33, 90th percentile: 76).<sup>8,9</sup>

On the one hand, previous simulation studies indicate that even in very small meta-analyses (here defined as those with  $\leq 5$  studies), many existing methods provide nearly unbiased point estimates for the meta-analytic mean, termed  $\mu$ .<sup>10</sup> On the other hand, confidence intervals that are based on asymptotic normality (e.g., Wald intervals) can have less than nominal coverage in small meta-analyses ( $\leq 20$  studies), and coverage can decline further in very small meta-analyses.<sup>7,11,12</sup> Using Hartung–Knapp–Sidik–Jonkman’s (HKSJ) method to adjust standard errors<sup>13,14</sup> can provide better-calibrated intervals in many settings, though existing simulation studies have yielded somewhat mixed findings regarding whether these intervals consistently achieve nominal coverage.<sup>7,11,12,15–17</sup> Moreover, such intervals can be extremely wide for meta-analyses of typical sample sizes.<sup>15–18</sup> For example, even when the true heterogeneity is zero, moments estimators with HKSJ standard errors yielded 95% confidence intervals with average widths of approximately 4–5 in simulated meta-analyses of 5 studies.<sup>18</sup> This suggests that for a point estimate of 0.5 on the standardized mean difference scale, a typical confidence interval would be approximately  $[-1.5, 2.5]$ , which is so wide that it might be considered uninformative. Additionally, standard point estimates for the heterogeneity can be substantially biased and imprecise in small meta-analyses.<sup>7,11</sup> Many existing simulation studies on heterogeneity estimation do not seem to have evaluated the coverage or width of confidence intervals for the heterogeneity<sup>11</sup> (but see Viechtbauer (2007)<sup>19</sup>).

In this paper, we investigate the frequentist performance of alternative Bayesian methods that use the invariant Jeffreys prior.<sup>20</sup> In general, Bayesian estimation proceeds by specifying a prior on the unknown parameters and obtaining the posterior of those parameters, given the observed data. This essentially involves updating the prior based on the likelihood of the observed data.<sup>21</sup> Various types of point estimates and credible intervals can then be obtained from the posterior. For an arbitrary distribution with unknown parameters  $\Psi$  and expected Fisher information  $I(\Psi)$ , the Jeffreys prior is proportional to  $\sqrt{\det I(\Psi)}$ .<sup>20</sup> An original motivation for this prior was its invariance to transforming the parameters,<sup>20</sup> a property that does not hold for all priors.<sup>22,23,i</sup> For example, letting  $\tau$  denote the standard deviation of studies’ population effects, the Jeffreys prior on  $(\mu, \tau)$  is the same as the Jeffreys prior on  $(\mu, \tau^2)$ , so the resulting posterior estimates and intervals would not depend on the analyst’s arbitrary choice of parameterization. This desirable property has led some to describe the Jeffreys prior as “noninformative,” though we agree with others’ critiques of this term.<sup>24,25</sup>

An interesting, underappreciated property of the Jeffreys prior is that the resulting posterior can alternatively be motivated from a solely frequentist perspective.<sup>26</sup> In particular, it is well-known that the maximum likelihood (ML) estimate has an  $O(n^{-1})$  bias, essentially due to the curvature of the score function.<sup>26</sup> Firth (1993)<sup>26</sup> showed that for exponential family distributions, an appropriate penalty on the likelihood to correct this bias coincides with estimation under the Jeffreys prior. This is essentially because the Jeffreys prior introduces a bias in the score function that compensates for the bias due to its curvature.<sup>26</sup> In particular, the posterior mode under this prior can be viewed in frequentist terms as a bias-corrected ML estimate; consequently, the posterior mode under the Jeffreys prior has sometimes been termed the “Firth correction.” The Firth correction has demonstrated success in a number of frequentist estimation problems, and is used fairly often for logistic regression.<sup>26–29</sup>

Given the Jeffreys prior’s effectiveness as a bias-correction method in small samples, it seems plausible that using this prior in small meta-analyses might improve point and interval estimation. Bodnar et al. (2016, 2017)<sup>15,30</sup> derived the Jeffreys prior on the heterogeneity  $\tau$  alone (i.e., holding the mean  $\mu$  constant), an approach that may be optimal if  $\tau$  is strictly a nuisance parameter.<sup>25</sup> Their simulations suggested that, along with an independent flat prior on  $\mu$ , the resulting credible intervals

may have better frequentist coverage than existing frequentist methods.<sup>15</sup> We term this prior “Jeffreys1” because it is the prior with respect to a single parameter. Kosmidis et al. (2017)<sup>31</sup> independently derived a penalized likelihood correction that is equivalent to the single-parameter Jeffreys prior on  $\mu$  alone; that is, treating  $\mu$  rather than  $\tau$  as a nuisance parameter. This penalization is closely related to the restricted ML (REML) estimator of  $\tau$ .<sup>31</sup>

In this paper, we consider the Jeffreys1 prior along with the two-parameter Jeffreys prior on both  $\mu$  and  $\tau$ . To the best of our knowledge, the latter has not appeared in the published literature on meta-analysis. We consider this prior, termed “Jeffreys2”, for several reasons. First, while the mean parameter is often of primary interest in meta-analysis, the heterogeneity should generally also be estimated and reported, so it may not be optimal to treat  $\tau$  as a nuisance parameter.<sup>32</sup> Second, in other small-sample estimation problems, multiparameter Jeffreys priors that include scale parameters (e.g., the dispersion parameter in exponential-family models) have been proposed and have good empirical properties.<sup>26,28,33</sup> (We return to this issue in Section 3.3.) In the context of adjusting for  $p$ -hacking in meta-analyses by meta-analyzing only a truncated part of the random-effects distribution, we recently found that a Jeffreys prior on  $\mu$  and  $\tau$  performed considerably better than ML,<sup>34</sup> whose performance is remarkably poor for truncated distributions in general.<sup>28,35</sup> Third, as we will discuss, the shape of the Jeffreys2 prior suggests it might provide more precise intervals than the Jeffreys1 prior. Whether Jeffreys2 credible intervals show nominal frequentist coverage, and whether point estimation for  $\mu$  and  $\tau$  performs well, are open questions.

Previous simulation studies of Jeffreys priors in meta-analysis have provided promising preliminary results, but do have limitations. Those simulations investigated only the Jeffreys1 prior, but not Jeffreys2, and have considered point and interval estimation for  $\mu$ , but not  $\tau$ .<sup>15</sup> In this paper, we present a simulation study comparing the frequentist properties of point and interval estimation for both  $\mu$  and  $\tau$  under the Jeffreys1 and Jeffreys2 priors, as well as several of the best-performing frequentist methods. Using a simulation design that closely paralleled a recent, extensive simulation study by Langan et al. (2019)<sup>7</sup>, we substantially expanded on the range of comparison methods and simulation scenarios used in previous simulation studies of the Jeffreys1 prior. Previous simulations regarding the Jeffreys1 prior considered only posterior means for point estimation,<sup>15</sup> whereas the aforementioned bias-correction properties specifically apply to posterior modes. This may be especially relevant for point estimation of  $\tau$ , whose posterior is highly asymmetric. We therefore consider three types of Bayesian point estimates (the posterior mode, mean, and median) as well as two types of credible intervals (central and shortest). Our simulations include the best-performing methods in Langan et al.’s (2019)<sup>7</sup> simulation study, along with several other methods whose theoretical properties suggest they might also perform well, such as exact intervals<sup>18</sup> and intervals based on the profile likelihood.<sup>6</sup>

This paper is organized as follows. We briefly review existing moments and likelihood-based estimators for random-effects meta-analysis (Section 2), all of which have been covered in more detail elsewhere.<sup>6,18,36</sup> We also briefly review existing simulation results regarding these methods (Section 2.4). We review the established form of the Jeffreys1 prior<sup>15</sup> and derive the form of the Jeffreys2 prior; we then discuss posterior estimation under both priors (Section 3). We present the simulation study (Section 4) and a brief applied example (Section 5), and conclude with a general discussion.

## 2. Existing frequentist methods

### 2.1. Method-of-moments estimators

Moments estimators for meta-analysis are semiparametric; they involve specifying only the first two moments of the distribution of population effects, namely  $\mu$  and  $\tau^2$ . Because these methods do not require specifying the higher moments, they do not require assuming that population effects are normal. Specifically, consider  $k$  studies whose population effects,  $\mu_i$ , have expectation  $\mu$  and variance  $\tau^2$ . These two moments are the usual meta-analytic estimands of interest. Let  $\hat{\theta}_i$  and  $\sigma_i$  respectively

denote the point estimate and standard error of the  $i$ th study, such that  $\widehat{\theta}_i \sim N(\mu_i, \sigma_i^2)$  holds approximately. The within-study standard errors  $\sigma_i$  are generally treated as fixed and known.

For a given estimate of the heterogeneity variance,  $\widehat{\tau}^2$ , the estimated marginal variance of  $\widehat{\theta}_i$  is  $\widehat{\tau}^2 + \sigma_i^2$ . The uniformly minimum variance unbiased estimator (UMVUE) of  $\mu$  arises from weighting studies by the inverse of their estimated marginal variances,<sup>6</sup> denoted  $w_i = 1/(\widehat{\tau}^2 + \sigma_i^2)$ :

$$\widehat{\mu} = \frac{\sum_{i=1}^k w_i \widehat{\theta}_i}{\sum_{i=1}^k w_i}.$$

The various moments estimators are distinguished by their estimators for  $\tau^2$ , and hence the form of the weights  $w_i$ . Detailed reviews<sup>7,36,37</sup> and original papers on these approaches are available, so here we summarize briefly. Moments estimators for  $\tau^2$  are based on the generalized Q-statistic:

$$Q = \sum_{i=1}^k a_i (\widehat{\theta}_i - \widehat{\mu})^2, \quad (1)$$

where the form of the coefficients,  $a_i$ , differs across moments estimators. For example, the traditional Dersimonian–Laird estimator (DL)<sup>1</sup> sets  $a_i = 1/\sigma_i^2$ . The two-step DL estimator (DL2)<sup>2</sup> instead sets  $a_i = 1/(\widehat{\tau}_{DL}^2 + \sigma_i^2)$ , where  $\widehat{\tau}_{DL}^2$  is an initial estimate obtained using the DL estimator. The Paule–Mandel (PM)<sup>3,4</sup> estimator can be viewed as a limiting case of DL2, involving iteration over the estimates  $\widehat{\mu}$  and  $\widehat{\tau}^2$  until convergence. This estimator is also equivalent to the empirical Bayes estimator.<sup>5</sup> In general terms, empirical Bayes estimation uses the observed data to estimate the parameters of the Bayesian prior, rather than specifying the prior independently of the data.<sup>21</sup> In the context of meta-analysis, the empirical Bayes estimator essentially estimates the distribution of population effects by their posterior means, with the prior determined empirically.<sup>5</sup>

## 2.2. Likelihood-based estimators

In contrast to moments estimators, commonly used likelihood-based estimators assume that the population effects,  $\mu_i$ , arise independently from the distribution  $\mu_i \sim N(\mu, \tau^2)$ . Thus, the marginal distribution of studies' point estimates,  $\widehat{\theta}_i$ , is  $\widehat{\theta}_i \sim N(\mu, \tau^2 + \sigma_i^2)$ . We denote the  $k$ -vector of point estimates as  $\widehat{\boldsymbol{\theta}}$ . Letting  $S_i(\tau) = \sqrt{\tau^2 + \sigma_i^2}$  be the true marginal standard deviation of the  $i$ th study, the joint likelihood is:

$$p(\widehat{\boldsymbol{\theta}} \mid \mu, \tau) = \prod_{i=1}^k \frac{1}{S_i(\tau)\sqrt{2\pi}} \cdot \exp\left\{-\frac{1}{2}\left(\frac{\widehat{\theta}_i - \mu}{S_i(\tau)}\right)^2\right\}. \quad (2)$$

The standard ML estimator for  $\tau$  is obtained as usual by solving  $\frac{\partial}{\partial \tau} \log p(\widehat{\boldsymbol{\theta}} \mid \mu, \tau) = 0$ , whose solution depends on  $\mu$ .<sup>6</sup> Since this estimator does not take into account the loss in degrees of freedom due to the additional estimation of  $\mu$  itself, the resulting estimate is often negatively biased.<sup>6</sup> This issue motivates REML estimation, which can improve upon ML estimation by transforming the log-likelihood to remove the parameter  $\mu$ .<sup>6</sup>

## 2.3. Interval estimation

A simple Wald confidence interval can be obtained by assuming  $\widehat{\mu}$  is normally distributed, which holds asymptotically in  $k$  by standard ML properties. If the weights  $w_i$  are treated as known rather than

estimated, we have  $\widehat{\text{Var}}(\widehat{\mu}) = 1/\sum_{i=1}^k w_i$ . A Wald 95% confidence interval is:

$$\widehat{\mu} \pm c\sqrt{\widehat{\text{Var}}(\widehat{\mu})},$$

where  $c = \Phi^{-1}(0.975) \approx 1.96$  is the critical value of the standard normal distribution. However, Wald intervals exhibit substantial under-coverage for small meta-analyses, both because the normal approximation holds only asymptotically and because the approximation  $\widehat{\text{Var}}(\widehat{\mu}) = 1/\sum_{i=1}^k w_i$  does not account for the estimation of  $\tau^2$ .<sup>7,11,12</sup> Wald intervals can also be constructed for  $\widehat{\tau}$ , but exhibit similarly poor performance.<sup>19</sup> We therefore do not further discuss Wald intervals, focusing instead on the better-performing alternatives discussed below.

Regarding interval estimation for  $\mu$ , the alternative HKSJ, sometimes called “Knapp–Hartung,” interval addresses the limitations of the Wald interval.<sup>13,14</sup> This method more flexibly assumes that  $\widehat{\mu}$  follows a  $t$  distribution and additionally rescales  $\widehat{\text{Var}}(\widehat{\mu})$  to account for the estimation of  $\tau^2$  in the weights  $w_i$ :

$$\widehat{\text{Var}}(\widehat{\mu}) = \frac{\sum_{i=1}^k w_i (\widehat{\theta}_i - \widehat{\mu})^2}{(k-1) \sum_{i=1}^k w_i}.$$

For  $\tau$ , improved intervals can be constructed using the chi-square distribution of the Q statistic, per Eq. (1).<sup>19</sup> These “Q-profile” intervals substantially outperform Wald intervals.<sup>19</sup> For both  $\mu$  and  $\tau$ , ML profile intervals can also be constructed in the usual way.<sup>6</sup>

An interesting, relatively new approach provides exact rather than asymptotic intervals and is theoretically guaranteed to provide more than nominal coverage, under the assumption of normal population effects.<sup>18</sup> This method essentially involves inverting exact tests. Other parametric methods provide finite-sample corrections to the likelihood ratio test statistic; these include Skovgaard’s second-order correction and Bartlett’s correction.<sup>38–40</sup> These methods can improve upon basic likelihood methods for hypothesis testing,<sup>40</sup> but Skovgaard’s second-order correction was not designed for interval estimation and can be numerically unstable in this context.<sup>31</sup> Interval estimation with Bartlett’s correction is possible,<sup>41</sup> but is not implemented in existing software (I. Visser, personal communication, 8 July 2024).<sup>42,43</sup> Because our focus is on interval estimation rather than testing, our simulations do not include Skovgaard’s or Bartlett’s corrections. Finally, various parametric or nonparametric resampling methods can be used to obtain bootstrapped confidence intervals.<sup>19,43,44</sup> Nonparametric resampling can be conducted by resampling rows with replacement, after which one can obtain simple percentile bootstrap intervals or bias-corrected and accelerated (BCa) intervals, among many other types of bootstrap intervals.<sup>45,46</sup> The BCa confidence corrects for bias and skewness in the bootstrapped sampling distribution, which we speculate could be helpful when estimating the sampling distribution of  $\tau$ . The BCa bootstrap has performed relatively well for certain meta-analytic estimators that are functions of  $\widehat{\tau}$ .<sup>47</sup> However, bootstrapping is an asymptotic procedure whose finite-sample performance typically must be assessed through simulations.

#### 2.4. Existing simulations comparing these methods

Langan et al. (2017)<sup>11</sup> provide an excellent systematic review of simulation studies for different heterogeneity estimators.<sup>7</sup> Briefly, the DL estimator was negatively biased for  $\tau$  when heterogeneity was moderate to high, and the PM estimator was typically less biased.<sup>11</sup> The reviewed studies do not appear to have assessed interval estimation for  $\tau$ . Based on their own, more extensive simulation study, Langan et al. (2019)<sup>7</sup> generally recommend REML, PM, or DL2 for heterogeneity estimation, along with HKSJ confidence intervals for  $\mu$ ; however, they recommend caution in interpreting heterogeneity estimates in small meta-analyses.

Langan et al.'s (2019)<sup>7</sup> simulation study did not assess intervals based on the profile likelihood, bootstrapping, or the exact method; the latter was developed only recently. Regarding profile intervals, recommendations in the literature are inconsistent. A prominent paper stated that “the profile likelihood is a good method for computing confidence intervals”.<sup>48</sup> One simulation study seemed to support this recommendation, finding that when the heterogeneity is greater than zero, profile likelihood intervals showed the closest to nominal coverage.<sup>10</sup> On the other hand, another simulation study suggested that profile intervals often exhibited under-coverage for meta-analyses of only 5 studies.<sup>39</sup> The originators of the exact method provide simulations suggesting that the resulting intervals are not substantially wider than those of existing methods, despite the method’s theoretical guarantee of at least nominal coverage.<sup>18</sup> While our simulation study is primarily motivated by investigating the Jeffreys methods, a secondary contribution is to more extensively evaluate profile, bootstrap, and exact intervals. We now turn to establishing the theory for the Jeffreys1 and Jeffreys2 priors.

### 3. Bayesian methods using Jeffreys priors

#### 3.1. The Jeffreys priors

Under the assumption of normal population effects, Bodnar et al. (2017)<sup>15</sup> showed that the improper Jeffreys1 prior is:

$$p(\mu, \tau) \propto \tau \sqrt{\sum_{i=1}^k (S_i(\tau))^{-4}}$$

where, again,  $S_i(\tau) = \sqrt{\tau^2 + \sigma_i^2}$ . Since this prior is independent of  $\mu$ , it can be expressed as two independent priors on  $\mu$  and  $\tau$ , where the prior on  $\mu$  is uniform:

$$p(\mu, \tau) \propto p(\mu) p(\tau), \text{ where } p(\mu) \propto 1 \text{ and } p(\tau) \propto \tau \sqrt{\sum_{i=1}^k (S_i(\tau))^{-4}}. \quad (3)$$

If  $\mu$  is treated as the only parameter of interest and  $\tau$  is considered a nuisance parameter, then the Jeffreys1 prior also coincides with the Berger–Bernardo reference prior.<sup>30</sup> In general, the Berger–Bernardo prior for a given distribution is designed to be maximally “noninformative” in the sense of minimizing the amount of information provided by the prior and maximizing the amount of information provided by the data.<sup>30,49</sup> Specifically, this prior maximizes the Kullback–Liebler divergence between the prior and the posterior.<sup>49</sup>

Regarding the Jeffreys2 prior, the joint likelihood in Eq. (2) implies that the entries of the expected Fisher information are:

$$k_{\mu\mu} := E \left[ \frac{\partial^2 \ell}{\partial \mu^2} \right] = - \sum_{i=1}^k (S_i(\tau))^{-2}, \quad k_{\tau\tau} := E \left[ \frac{\partial^2 \ell}{\partial \tau^2} \right] = -2\tau^2 \sum_{i=1}^k (S_i(\tau))^{-4}, \quad k_{\mu\tau} := E \left[ \frac{\partial^2 \ell}{\partial \mu \partial \tau} \right] = 0$$

where  $\ell$  is the likelihood function. Therefore, the Jeffreys2 prior is  $p(\mu, \tau) \propto \sqrt{k_{\mu\mu} k_{\tau\tau}}$ . (This result is straightforward to show directly, or alternatively can be viewed as a simple special case of the prior given in Mathur (2024).<sup>34,ii</sup> This yields the improper two-parameter prior:

$$p(\mu, \tau) \propto \tau \sqrt{\left( \sum_{i=1}^k (S_i(\tau))^{-2} \right) \left( \sum_{i=1}^k (S_i(\tau))^{-4} \right)}.$$



Like the Jeffreys1 prior, the Jeffreys2 prior can be expressed as:

$$p(\mu, \tau) \propto p(\mu) p(\tau), \text{ where } p(\mu) \propto 1 \text{ and } p(\tau) \propto \tau \sqrt{\left(\sum_{i=1}^k (S_i(\tau))^{-2}\right) \left(\sum_{i=1}^k (S_i(\tau))^{-4}\right)}. \quad (4)$$

To illustrate, Figure 1 shows both priors on  $\tau$  for four meta-analyses of standardized mean differences. The meta-analyses were simulated with studies' sample sizes,  $N$ , arising from four different distributions. Although the magnitude of the priors will of course be affected by the number of studies  $k$ , their shape is minimally affected by  $k$ , so Figure 1 depicts the prior for meta-analyses with  $k = 10$ . Note that for each meta-analysis, the Jeffreys2 prior is somewhat narrower than the Jeffreys1 prior, suggesting that the former may provide narrower intervals; this hypothesis will be explored in more depth in the simulation study (Section 4). Both priors lead to proper posteriors if  $k > 1$  (see Bodnar (2017)<sup>15</sup> regarding Jeffreys1 and the present Section 1 of the Supplementary Material, regarding Jeffreys2). Additionally, both priors generalize easily to the case of meta-regression: the Jeffreys1 prior would coincide with that of Bodnar et al. (2024) for generalized marginal random effects models,<sup>50</sup> and we derive the Jeffreys2 prior for meta-regression in Section 1 of the Supplementary Material. We do not further consider meta-regression in the main text.

### 3.2. The posterior under each prior

For either prior, since  $p(\mu, \tau) \propto p(\tau)$ , the marginal posterior on  $\tau$  is:<sup>15</sup>

$$p(\tau | \hat{\theta}) \propto p(\tau) \int_{-\infty}^{+\infty} p(\hat{\theta} | \mu, \tau) d\mu \\ \propto p(\tau) \cdot \prod_{i=1}^k \frac{1}{S_i(\tau) \sqrt{\sum_{i=1}^k (S_i(\tau))^{-2}}} \cdot \exp \left\{ -\frac{1}{2} \left( \sum_{i=1}^k (S_i(\tau))^{-2} \hat{\theta}_i^2 - \frac{(\sum_{i=1}^k (S_i(\tau))^{-2} \hat{\theta}_i)^2}{\sum_{i=1}^k (S_i(\tau))^{-2}} \right) \right\}.$$

In turn, the conditional posterior of  $\mu$ , given  $\tau$ , is normal:<sup>9,15,21</sup>

$$p(\mu | \tau, \hat{\theta}) \propto \frac{1}{\sqrt{2\pi \text{Var}(\mu | \tau, \hat{\theta})}} \cdot \exp \left\{ -\frac{1}{2 \text{Var}(\mu | \tau, \hat{\theta})} (\mu - E[\mu | \tau, \hat{\theta}])^2 \right\},$$

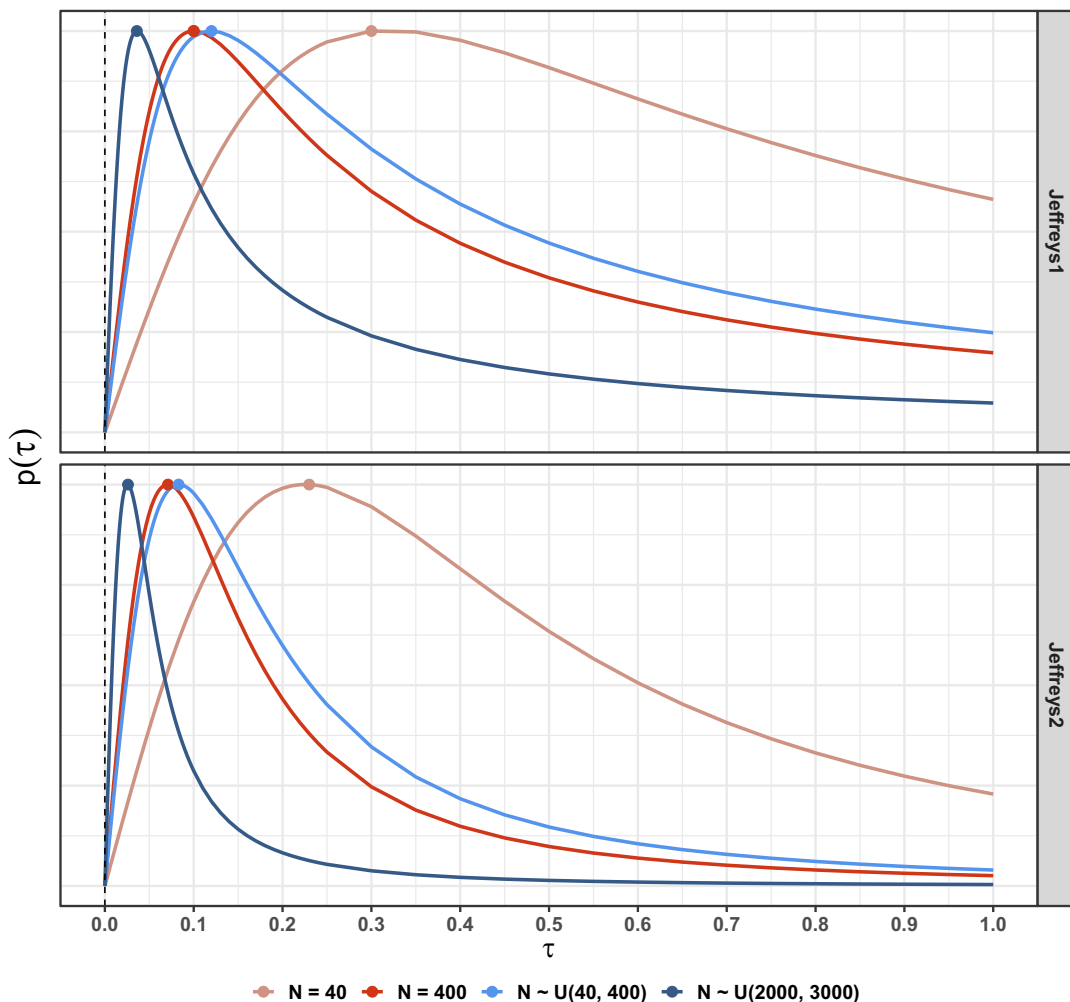
where:

$$E[\mu | \tau, \hat{\theta}] = \frac{\sum_{i=1}^k \hat{\theta}_i (S_i(\tau))^{-2}}{\sum_{i=1}^k (S_i(\tau))^{-2}}, \quad \text{Var}(\mu | \tau, \hat{\theta}) = \frac{1}{\sum_{i=1}^k (S_i(\tau))^{-2}}.$$

Thus, the joint posterior  $p(\mu, \tau | \hat{\theta})$  can be decomposed into the two tractable components  $p(\tau | \hat{\theta})$  and  $p(\mu | \tau, \hat{\theta})$ .<sup>9</sup> Given this observation, Röver and others<sup>9,51</sup> developed theory and software for a discrete approximation to the joint posterior  $p(\mu, \tau | \hat{\theta})$  and the marginal posterior on  $\mu$ , given by the mixture distribution:

$$p(\mu | \hat{\theta}) \propto \int_0^{\infty} p(\mu | \tau, \hat{\theta}) p(\tau | \hat{\theta}) d\tau.$$

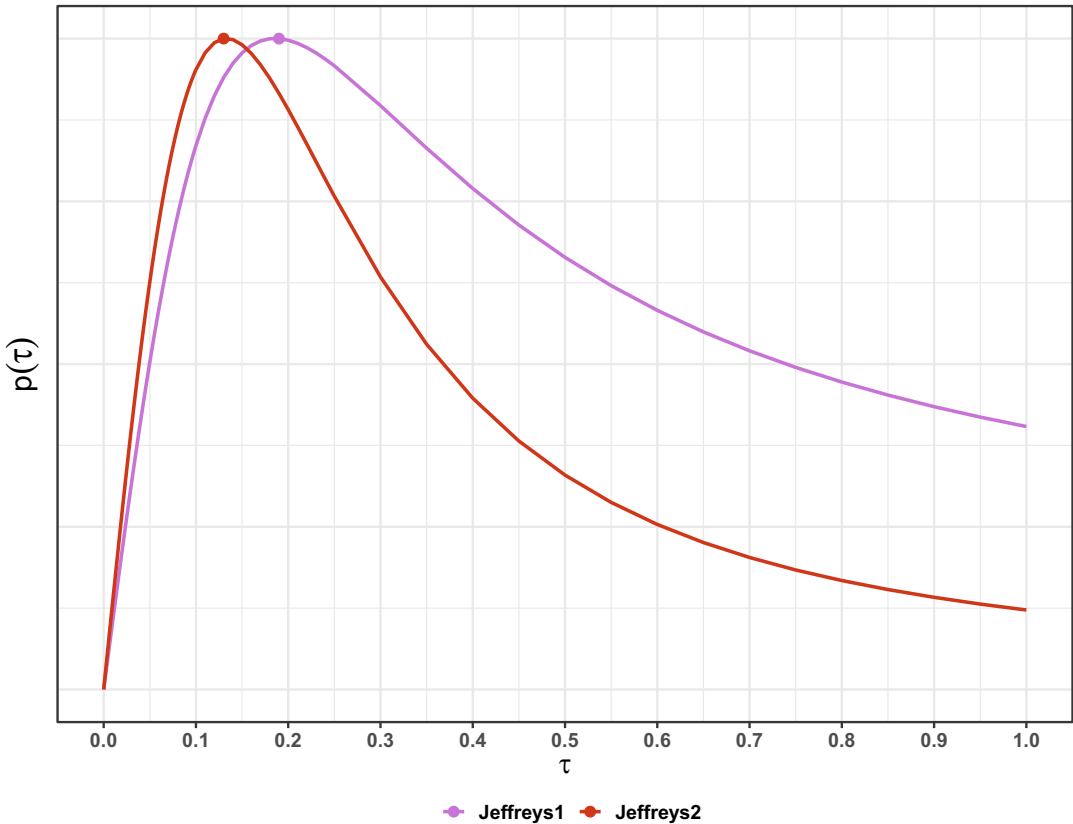
The discrete approximation approach does not require sampling via mixed-chain Monte Carlo (MCMC) and is implemented in the R package `bayesmeta`.<sup>9,51</sup> We use this package in our simulations and applied example.



**Figure 1.** Priors for four simulated meta-analyses of standardized mean differences ( $k = 10$ ), in which the within-study sample sizes ( $N$ ) were generated from four possible distributions. Studies' standard errors were estimated using Eq. (5) and, given the data-generation parameters, were approximately equal to  $2/\sqrt{N}$ . Points are the maxima. The priors have been scaled to have the same maximum height.

With approximations to the joint and marginal posteriors in hand, point estimates can be defined in terms of various measures of central tendency, such as the posterior mode, median, or mean. For either prior,  $p(\mu | \hat{\theta})$  appears to be nearly symmetric in many cases (e.g., Figure 4), so the three measures of central tendency will often agree closely. However, this is not the case for  $p(\tau | \hat{\theta})$ , which is asymmetric under either prior. Existing work on the Jeffreys1 prior focused primarily on posterior means and medians,<sup>15</sup> but we focus on posterior modes given their aforementioned theoretical advantages.<sup>26</sup> Indeed, as discussed in Section 4.4, our simulations indicated that posterior modes for  $\tau$  provided substantially lower bias, root mean square error (RMSE), and mean absolute error (MAE) than did posterior means and medians. As in ML estimation, point estimates can be defined either in terms of the marginal or the joint mode. In the Bayesian context, the marginal mode represents the value of a given parameter (e.g.,  $\mu$ ) that maximizes the posterior for that parameter alone, marginalizing over the other parameter (e.g.,  $\tau$ ). In contrast, the joint mode represents the values of both parameters that





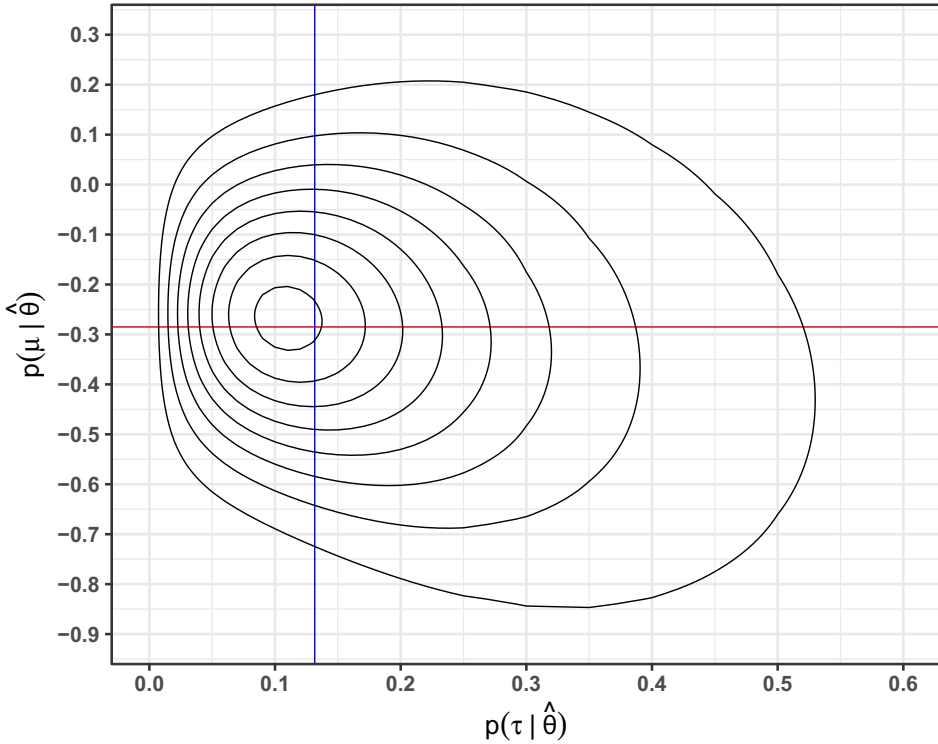
**Figure 2.** Priors on  $\tau$  for the meta-analysis on all-cause death ( $k = 3, \{\sigma_i\} = \{1.15, 1.63, 0.19\}$ ). Points are maxima. The priors have been scaled to have the same maximum height.

jointly maximize the joint posterior. We consider marginal modes in this paper to provide a more direct comparison to marginal ML estimation, which is the usual implementation for meta-analysis.

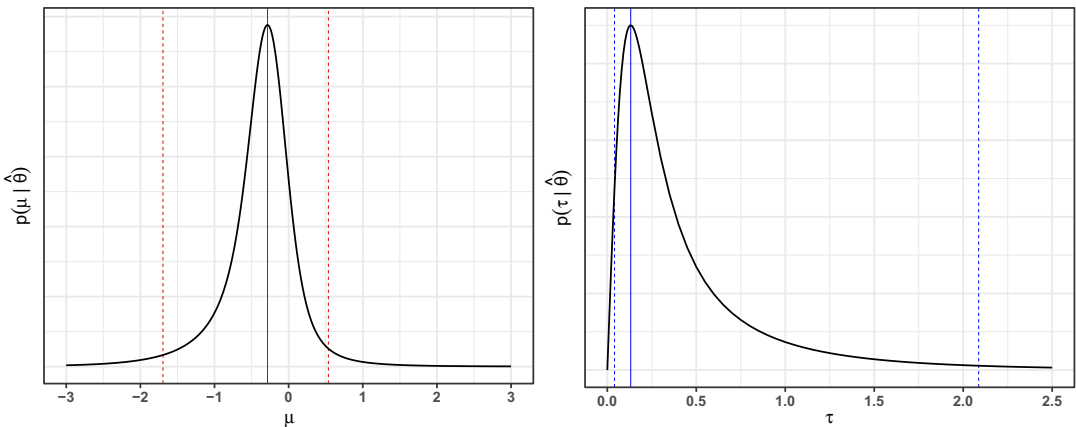
Also analogously to ML estimation, symmetric Wald credible intervals are sometimes constructed for Bayesian estimates by approximating the posterior as asymptotically normal around the posterior mode, with a variance–covariance matrix equal to the inverse of the Hessian of the negative log-posterior evaluated at the posterior mode.<sup>21</sup> However, just as Wald intervals around the ML estimate can perform poorly if the likelihood is asymmetric, Wald intervals around the posterior mode can likewise perform poorly if the posterior is asymmetric.<sup>52</sup> To obtain appropriately asymmetric posterior intervals, we consider two approaches. First, a central (also called “equal-tailed”) 95% posterior quantile interval can be obtained by taking the 2.5th and 97.5th quantiles of the estimated posterior distribution. Second, the shortest possible 95% posterior quantile interval can be obtained numerically; this interval is equivalent to a highest posterior density interval for unimodal distributions.<sup>21</sup> In our simulations and applied example, we obtain both types of intervals from the R package *bayesmeta*.<sup>9</sup>

### 3.3. Theoretical and substantive distinctions between the priors

The distinction between the Jeffreys1 and Jeffreys2 priors invokes theoretical and substantive considerations that pertain in general to multiparameter Jeffreys priors. Jeffreys and others have argued that multiparameter Jeffreys priors are appropriate if one wishes to estimate all of the parameters (i.e., both  $\mu$  and  $\tau$  in meta-analysis), but not if one wishes to estimate only a subset of the parameters (i.e., only  $\mu$ ),

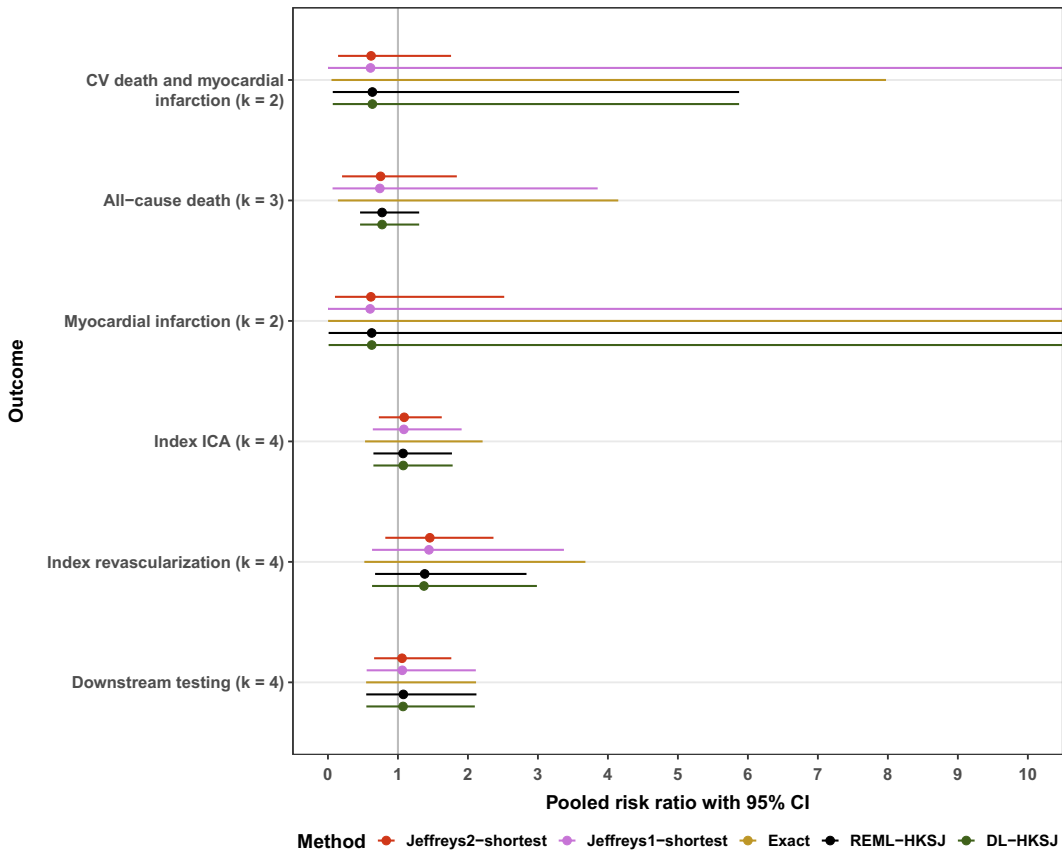


**Figure 3.** Joint posterior under the Jeffreys2 prior for the meta-analysis on all-cause death ( $k = 3$ ,  $\{\sigma_i\} = \{1.15, 1.63, 0.19\}$ ). Horizontal red line: marginal posterior mode of  $\mu$ . Vertical blue line: marginal posterior mode of  $\tau$ .



**Figure 4.** Marginal posteriors under the Jeffreys2 prior for the meta-analysis on all-cause death. Solid vertical lines: marginal posterior modes. Dashed vertical lines: limits of 95% intervals.

with the others treated as nuisance parameters.<sup>24,25,53</sup> As noted in the Introduction, a random-effects meta-analysis should generally involve estimation and reporting of  $\tau$  (or related metrics<sup>32,54,55</sup>) in addition to  $\mu$ , which suggests consideration of the Jeffreys2 prior. On the other hand, in general location-scale problems, Jeffreys recommended obtaining the prior with respect to only the scale parameters,



**Figure 5.** Interval limits greater than  $RR = 10$  are truncated. The exact method does not yield point estimates. CI: credible interval.

holding constant the location parameters.<sup>24,53</sup> This would correspond to the Jeffreys1 prior. Jeffreys’ recommendation was motivated by problems that can arise when the number of location parameters increases with the sample size, similarly to the well-known Neyman–Scott problem in which the ML estimator fails to be consistent.<sup>24,53</sup> Interestingly, Firth later showed that in a specific, severe version of the Neyman–Scott problem, the multiparameter Jeffreys prior (i.e., the Firth correction) in fact leads to a consistent and exactly unbiased estimator.<sup>26</sup> This was unexpected given that the asymptotic arguments justifying the Firth correction are violated with an increasing number of parameters.<sup>26</sup> Of course, in the present setting of random-effects meta-analysis, the number of parameters is fixed, so this potential issue does not arise in the first place. Our view is that existing substantive and theoretical considerations do not clearly rule out either prior as inappropriate for random-effects meta-analysis, so our simulation study evaluates both.

#### 4. Simulation study

We designed the simulation study to closely parallel that of Langan et al. (2019),<sup>7</sup> which in turn was designed to address many of the limitations of previous simulation studies.<sup>11</sup> As detailed below, we considered meta-analyses with binary outcomes (with effect sizes on the log-odds ratio scale) and with continuous outcomes (with effect sizes on the Hedges’  $g$  scale<sup>56</sup>), with as few as 2 studies, with varying amounts of heterogeneity, with varying means and outcome probabilities (for binary outcomes), and with varying distributions of within-study sample sizes. Because we assessed a variety of parametric,

**Table 1.** *Methods assessed in simulation study.*

Abbreviation	Method
<b>Point estimation</b>	
ML	Maximum likelihood
REML	Restricted maximum likelihood
DL	Dersimonian–Laird
DL2	Two-step Dersimonian–Laird
PM	Paule–Mandel
Jeffreys1	Marginal posterior mode under Jeffreys1 prior
Jeffreys2	Marginal posterior mode under Jeffreys2 prior
<b>Interval estimation</b>	
HKSJ	Hartung–Knapp–Sidik–Jonkman interval for $\mu$ (suffix for moments estimators and likelihood methods)
Qprofile	Q-profile interval for $\tau$ (suffix for moments estimators and likelihood methods)
ML-profile	Maximum likelihood profile interval
Exact	Exact interval for $\mu$
Jeffreys1-shortest, Jeffreys2-shortest	Shortest interval
Jeffreys1-central, Jeffreys2-central	Central (equal-tailed) interval. Results only shown for interval estimation for $\tau$ .
Boot-BCa <sup>a</sup>	Nonparametric bias-corrected and accelerated bootstrap
Boot-perc <sup>a</sup>	Nonparametric percentile bootstrap

<sup>a</sup>In pilot tests for the scenarios with  $k = 10$ , the bootstrap methods were not competitive with other methods, so these computationally intensive methods were not run for other sample sizes.

semiparametric, and nonparametric methods, we preliminarily investigated robustness to parametric misspecification by considering exponentially distributed population effects in addition to normally distributed effects.

#### 4.1. Point and interval estimation methods

Table 1 lists the methods assessed in our simulation study. We assessed both Jeffreys priors. For point estimation under each prior, we primarily considered marginal posterior modes but secondarily investigated posterior means and medians (Section 2.2 of the Supplementary Material). Regarding interval estimation for  $\mu$ , central and shortest intervals were generally quite similar, so we only show results for shortest intervals. Regarding interval estimation for  $\tau$ , we consider both types of intervals for each prior, termed “Jeffreys1-shortest,” “Jeffreys1-central,” “Jeffreys2-shortest,” and “Jeffreys2-central.”

We compared the performance of both Jeffreys priors to that of several existing frequentist methods that were described in Section 2. We selected methods that have performed well in existing, large simulation studies or that have desirable theoretical properties, such as providing appropriately asymmetric intervals for  $\tau$ .<sup>6,7,18,39,48,57</sup> For point estimation, the comparison methods were ML estimation, REML, DL, DL2, and PM. Regarding interval estimation for  $\mu$ , we considered HKSJ intervals for each frequentist estimation method, ML profile intervals (ML-profile), exact intervals,<sup>18</sup> nonparametric BCa bootstrap intervals, and nonparametric percentile bootstrap intervals.<sup>45,46</sup> Regarding interval estimation for  $\tau$ , we considered Q-profile intervals for each frequentist estimation method, as well as ML-profile and both bootstrap intervals. We implemented all frequentist methods and intervals using the R package

**Table 2.** Possible values of simulation parameters.

Simulation parameter	Possible values
Outcome type	Continuous (Hedges' $g$ ), binary (log-odds ratio)
$k$	2, 3, 5, 10, 20, 100 <sup>a</sup>
$\tau$	0.01, 0.05, 0.10, 0.20, 0.50
Distribution of population effects	Normal, exponential
Distribution of within-study $N$	All $N = 40$ , all $N = 400$ , $N \sim U(40, 400)$ , $N \sim U(2000, 4000)$
For continuous outcomes only:	
$\mu$	0.5
For binary outcomes only:	
$\mu$	0, 0.5, 1.1, 2.3
$P(Y = 1   X = 0)$	0.05, 0.1, 0.50

<sup>a</sup> Results for scenarios with  $k = 100$  appear in the Supplementary Material; these scenarios are excluded from aggregated results in the main text.

metafor<sup>58</sup> with the following exceptions: we implemented ML-profile using custom R code, the exact method using the R package `rma.exact`,<sup>18</sup> and the bootstrap methods using the R package `boot`.<sup>59</sup>

#### 4.2. Data generation

Table 2 summarizes the simulation parameters we manipulated, which were similar to those of Langan et al.'s (2019) simulation study.<sup>7</sup> We considered continuous outcomes with point estimates on the Hedges'  $g$  scale<sup>56</sup> as well as binary outcomes with point estimates on the log-odds ratio scale. We considered both normally distributed and exponentially distributed population effects; in the latter case, the assumptions for all point estimators except the moments estimators were violated. Statistical theory suggests that all methods would perform comparably for very large meta-analyses with normal effects, and accordingly our focus is on point and interval estimation for smaller meta-analyses ( $k \leq 20$ ). Our primary simulations reported in the main text are those with  $k \in \{2, 3, 5, 10, 20\}$ . We additionally ran simulations with  $k = 100$  to confirm asymptotic behavior (Section 3 of the Supplementary Material). Because the bootstrap intervals required much more computational time than the other methods, we first pilot-tested them in all scenarios with a single sample size ( $k = 10$ ) to assess whether these methods were competitive with other methods.

Data generation proceeded as follows. For each simulation iterate, we generated a meta-analysis whose underlying population effects ( $\mu_i$ ) were either normal or exponential. Normal population effects were generated as  $\mu_i \sim N(\mu, \tau^2)$ , where we varied  $\mu$  and  $\tau$  as indicated in Table 2. Exponential population effects were generated from an appropriately scaled and shifted distribution to achieve the desired population moments,  $\mu$  and  $\tau^2$ . For each study in the meta-analysis, we generated a total sample size  $N$  from one of the four distributions listed in Table 2. We then simulated individual participant data, such that  $N/2$  participants were allocated to a treatment group, and the other  $N/2$  to a control group. In scenarios with a continuous outcome, we simulated outcomes with a mean of 0 in the control group and  $\mu_i$  in the treatment group, and with a standard deviation of 1 within each group. We then estimated the standardized mean difference using the Hedges'  $g$  correction.<sup>56,58</sup> We used the standard large-sample approximation for studies' standard errors (Eq. (8) in Hedges (1982)<sup>60</sup>):

$$\widehat{\sigma}_i = \sqrt{\frac{N^c + N^t}{N^c N^t} + \frac{\widehat{\theta}_i^2}{2(N^c + N^t)}} = \sqrt{\frac{8 + \widehat{\theta}_i^2}{2N}} \quad (5)$$

where  $N^c$  and  $N^t$  are the within-group sample sizes, which were both equal to  $N/2$  in our simulations. The expectation of this estimator is approximately  $\sqrt{(8 + \mu^2)/2N}$ .

In scenarios with a binary outcome, we simulated outcomes from a logistic model such that:

$$P(Y = 1 | X) = \text{expit}\{\text{logit}\{P(Y = 1 | X = 0)\} + \mu_i X\}$$

where  $P(Y = 1 | X = 0)$  was a scenario parameter that we manipulated among the values listed in Table 2. We then estimated the odds ratio; to handle potential zero cell counts when present, we added 0.5 to each table cell when any cells had a count of zero.<sup>58</sup>

We expected that for binary outcomes and small within-study sample sizes, certain extreme combinations of scenario parameters (e.g.,  $N = 40$  and  $\mu = 2.3$ , corresponding to an extreme odds ratio of 10) would result in biased within-study odds ratios.<sup>26,61</sup> In pilot simulations, we identified combinations of scenario parameters that resulted in within-study absolute bias of greater than 0.05. We excluded these combinations of scenario parameters since our focus is on bias arising from meta-analytic estimation methods rather than from within-study bias. After excluding these combinations of simulation parameters, we ultimately simulated 240 unique scenarios for continuous outcomes and 2267 for binary outcomes.

### 4.3. Performance metrics

For each scenario, we assessed the point estimators' performance and variability in terms of their bias, MAE, and RMSE, defined in the usual frequentist sense. That is, for a generic parameter  $\omega_r$  that varies across 500 simulation iterates,  $r$ :

$$\begin{aligned} \text{Bias} &= \frac{1}{500} \sum_{r=1}^{500} (\widehat{\omega}_r - \omega_r) \\ \text{MAE} &= \frac{1}{500} \sum_{r=1}^{500} |\widehat{\omega}_r - \omega_r| \\ \text{RMSE} &= \left( \frac{1}{500} \sum_{r=1}^{500} (\widehat{\omega}_r - \omega_r)^2 \right)^{1/2}. \end{aligned}$$

For each scenario, we assessed interval estimation in terms of the frequentist coverage and width of 95% confidence or credible intervals. Some methods' intervals exhibited over-coverage in some scenarios but exhibited under-coverage in others. Therefore, when aggregating results across scenarios, we also consider the percentage of scenarios in which each method achieved approximately nominal coverage, defined stringently as having coverage >94%. In the Discussion, we expand upon our reasons for assessing frequentist properties of Bayesian methods, and the implications of this approach. We did not assess statistical power. Although  $p$ -values can certainly be useful when interpreted as continuous measures of evidence, we concur with others' longstanding concerns about bright-line significance testing,<sup>62,63</sup> a practice that has contributed to striking misinterpretations of published meta-analyses<sup>55,64</sup> and likely also to publication bias.

### 4.4. Results

Given the large number of scenarios, some aggregation is necessary to display the results compactly. In the main text, we provide line plots that stratify by  $k$ ,  $\tau$ , the distribution of population effects, and the outcome type, and that aggregate over distributions of  $N$  and, for binary outcomes, over  $\mu$  and  $P(Y = 1 | X = 0)$ . Because the direction of a given estimator's bias could differ across scenarios, we depict

**Table 3.** Scenarios with continuous outcomes;  $\hat{\mu}$  point and interval estimation.

$\hat{\mu}$ MAE	$\hat{\mu}$ RMSE	$\hat{\mu}$ coverage	$\hat{\mu}$ coverage > 0.94	$\hat{\mu}$ CI width					
All	0.09	All	0.12	Jeffreys1-shortest	0.98	Jeffreys1-shortest	0.89	ML-profile	0.51
				Exact	0.97	Exact	0.84	Jeffreys2-shortest	0.67
				Jeffreys2-shortest	0.96	Jeffreys2-shortest	0.72	DL-HKSJ	1.19
				DL-HKSJ	0.94	DL-HKSJ	0.64	DL2-HKSJ	1.19
				DL2-HKSJ	0.94	DL2-HKSJ	0.64	ML-HKSJ	1.19
				ML-HKSJ	0.94	ML-HKSJ	0.64	PM-HKSJ	1.19
				PM-HKSJ	0.94	PM-HKSJ	0.64	REML-HKSJ	1.19
				REML-HKSJ	0.94	REML-HKSJ	0.64	Exact	1.36
				ML-profile	0.93	ML-profile	0.50	Jeffreys1-shortest	1.57

Note: Methods are sorted from best to worst performance within each column (or alphabetically for ties); coverage is sorted from highest to lowest. MAE: Mean absolute error. RMSE: Root mean square error. CI: 95% confidence or credible interval. Coverage >0.94: Percent of scenarios for which coverage probability was at least 0.94. "All": All methods performed equally to two decimal places.

each estimator’s bias across scenarios using boxplots instead of line plots to avoid any aggregation across scenarios. For the other performance metrics, we additionally provide a series of tables that consider average performances within subsets of scenarios defined by the outcome type and  $k$  (Tables 3–10). Comprehensive simulation results for each individual scenario are publicly available as a dataset (<https://osf.io/9qfah>).

As described above, our focus is on small meta-analyses. Thus, except where otherwise noted, all subsequent results pertain to scenarios with  $k \leq 20$ , and we refer to these as “all scenarios.” Although tables and figures show results for both normal and exponential population effects, our prose descriptions focus primarily on scenarios with normal effects; in these scenarios, all methods were correctly specified. We do secondarily discuss how results changed for exponentially distributed effects. Note that figures stratify on effect distribution, while tables aggregate over normal and exponential effects due to space constraints.

#### 4.4.1. Convergence metrics

Other than the exact method and the BCa bootstrap, all methods’ algorithms converged (in the sense of yielding point estimates and/or intervals for  $\hat{\mu}$  and  $\hat{\tau}$ ) in >99% of simulated datasets. The exact method is designed only to provide an interval for  $\hat{\mu}$ , and its algorithm did so in >98% of simulated datasets. In the subset of scenarios we ran with the bootstrap methods (i.e., scenarios with  $k = 10$ ), the BCa bootstrap only provided an interval for  $\hat{\mu}$  and  $\hat{\tau}$  in 67% of datasets. When no interval was provided, this was because the estimated bias correction was infinite, which can happen when empirical influence values are close to zero due to outliers or small sample sizes.

#### 4.4.2. Point and interval estimation for $\mu$

Consistent with previously published simulations,<sup>10</sup> all methods performed very similarly for point estimation of  $\mu$  and were approximately unbiased (Figure 6 and Section 2.1 of the Supplementary Material). Across all scenarios, the maximum within-scenario absolute differences between any two methods in bias, RMSE, and MAE respectively were only 0.056, 0.064, and 0.036. Given these relatively minor differences in point estimation for  $\mu$ , we primarily discuss interval estimation for this estimand. In pilot tests for  $k = 10$  scenarios, the bootstrap methods were not competitive with other methods (Sections 3.7 and 3.8 of the Supplementary Material). Therefore, we did not run these computationally intensive methods = for other sample sizes, and the bootstrap methods are omitted from results in the main text.



**Table 4.** Scenarios with continuous outcomes;  $\hat{\tau}$  point and interval estimation.

$\hat{\tau}$ MAE		$\hat{\tau}$ RMSE		$\hat{\tau}$ coverage		$\hat{\tau}$ coverage > 0.94		$\hat{\tau}$ CI width	
Jeffreys1	0.09	Jeffreys2	0.10	DL-Qprofile	0.93	Jeffreys1-shortest	0.74	ML-profile	0.42
Jeffreys2	0.09	Jeffreys1	0.11	DL2-Qprofile	0.93	Jeffreys2-shortest	0.62	Jeffreys2-shortest	0.50
ML	0.09	ML	0.11	Jeffreys1-shortest	0.93	ML-profile	0.62	Jeffreys2-central	0.62
DL	0.10	DL	0.12	ML-Qprofile	0.93	Jeffreys2-central	0.58	Jeffreys1-shortest	1.37
DL2	0.10	DL2	0.12	PM-Qprofile	0.93	DL-Qprofile	0.56	DL-Qprofile	1.47
PM	0.10	PM	0.12	REML-Qprofile	0.93	DL2-Qprofile	0.56	DL2-Qprofile	1.47
REML	0.10	REML	0.12	ML-profile	0.92	Jeffreys1-central	0.56	ML-Qprofile	1.47
				Jeffreys2-shortest	0.91	ML-Qprofile	0.56	PM-Qprofile	1.47
				Jeffreys2-central	0.79	PM-Qprofile	0.56	REML-Qprofile	1.47
				Jeffreys1-central	0.77	REML-Qprofile	0.56	Jeffreys1-central	2.40

*Note:* Methods are sorted from best to worst performance within each column (or alphabetically for ties); coverage is sorted from highest to lowest. MAE: Mean absolute error. RMSE: Root mean square error. CI: 95% confidence or credible interval. Coverage >0.94: Percent of scenarios for which coverage probability was at least 0.94.

**Table 5.** Scenarios with binary outcomes;  $\hat{\mu}$  point and interval estimation.

$\hat{\mu}$ MAE		$\hat{\mu}$ RMSE		$\hat{\mu}$ coverage		$\hat{\mu}$ coverage > 0.94		$\hat{\mu}$ CI width	
All	0.16	All	0.20	Jeffreys1-shortest	0.99	Jeffreys1-shortest	0.95	ML-profile	0.96
				Exact	0.98	Exact	0.93	Jeffreys2-shortest	1.38
				Jeffreys2-shortest	0.98	Jeffreys2-shortest	0.89	ML-HKSJ	2.05
				DL-HKSJ	0.95	DL-HKSJ	0.73	REML-HKSJ	2.07
				DL2-HKSJ	0.95	DL2-HKSJ	0.73	DL-HKSJ	2.08
				ML-HKSJ	0.95	ML-HKSJ	0.73	DL2-HKSJ	2.08
				ML-profile	0.95	ML-profile	0.73	PM-HKSJ	2.08
				PM-HKSJ	0.95	PM-HKSJ	0.73	Exact	2.64
				REML-HKSJ	0.95	REML-HKSJ	0.73	Jeffreys1-shortest	3.26

*Note:* Methods are sorted from best to worst performance within each column (or alphabetically for ties); coverage is sorted from highest to lowest. MAE: Mean absolute error. RMSE: Root mean square error. CI: 95% confidence or credible interval. Coverage >0.94: Percent of scenarios for which coverage probability was at least 0.94. "All": All methods performed equally to two decimal places.

**Table 6.** Scenarios with binary outcomes;  $\hat{\tau}$  point and interval estimation.

$\hat{\tau}$ MAE		$\hat{\tau}$ RMSE		$\hat{\tau}$ coverage		$\hat{\tau}$ coverage > 0.94		$\hat{\tau}$ CI width	
ML	0.13	ML	0.16	ML-profile	0.96	ML-profile	0.82	ML-profile	0.80
DL	0.15	Jeffreys2	0.17	DL-Qprofile	0.94	Jeffreys1-shortest	0.78	Jeffreys2-shortest	1.02
DL2	0.15	DL	0.20	DL2-Qprofile	0.94	Jeffreys2-shortest	0.72	Jeffreys2-central	1.28
PM	0.15	DL2	0.20	ML-Qprofile	0.94	DL-Qprofile	0.71	DL2-Qprofile	2.17
REML	0.15	Jeffreys1	0.20	PM-Qprofile	0.94	ML-Qprofile	0.71	DL-Qprofile	2.19
Jeffreys2	0.16	PM	0.20	REML-Qprofile	0.94	PM-Qprofile	0.71	ML-Qprofile	2.19
Jeffreys1	0.18	REML	0.20	Jeffreys1-shortest	0.90	REML-Qprofile	0.71	PM-Qprofile	2.19
				Jeffreys2-shortest	0.89	DL2-Qprofile	0.70	REML-Qprofile	2.19
				Jeffreys2-central	0.68	Jeffreys2-central	0.51	Jeffreys1-shortest	2.86
				Jeffreys1-central	0.64	Jeffreys1-central	0.45	Jeffreys1-central	5.06

*Note:* Methods are sorted from best to worst performance within each column (or alphabetically for ties); coverage is sorted from highest to lowest. MAE: Mean absolute error. RMSE: Root mean square error. CI: 95% confidence or credible interval. Coverage >0.94: Percent of scenarios for which coverage probability was at least 0.94.

**Table 7.** Scenarios with continuous outcomes,  $k \leq 5$ ;  $\hat{\mu}$  point and interval estimation.

$\hat{\mu}$ MAE		$\hat{\mu}$ RMSE		$\hat{\mu}$ coverage		$\hat{\mu}$ coverage > 0.94		$\hat{\mu}$ CI width	
All	0.12	All	0.15	Jeffreys1-shortest	0.99	Jeffreys1-shortest	0.93	ML-profile	0.66
				Exact	0.98	Exact	0.91	Jeffreys2-shortest	0.92
				Jeffreys2-shortest	0.96	Jeffreys2-shortest	0.76	ML-HKSJ	1.78
				DL-HKSJ	0.94	DL-HKSJ	0.68	DL-HKSJ	1.79
				DL2-HKSJ	0.94	DL2-HKSJ	0.68	DL2-HKSJ	1.79
				ML-HKSJ	0.94	ML-HKSJ	0.68	PM-HKSJ	1.79
				PM-HKSJ	0.94	PM-HKSJ	0.68	REML-HKSJ	1.79
				REML-HKSJ	0.94	REML-HKSJ	0.68	Exact	2.06
				ML-profile	0.92	ML-profile	0.52	Jeffreys1-shortest	2.40

*Note:* Methods are sorted from best to worst performance within each column (or alphabetically for ties); coverage is sorted from highest to lowest. MAE: Mean absolute error. RMSE: Root mean square error. CI: 95% confidence or credible interval. Coverage >0.94: Percent of scenarios for which coverage probability was at least 0.94. "All": All methods performed equally to two decimal places.

**Table 8.** Scenarios with continuous outcomes,  $k \leq 5$ ;  $\hat{\tau}$  point and interval estimation.

$\hat{\tau}$ MAE		$\hat{\tau}$ RMSE		$\hat{\tau}$ coverage		$\hat{\tau}$ coverage > 0.94		$\hat{\tau}$ CI width	
Jeffreys2	0.10	Jeffreys2	0.12	Jeffreys1-shortest	0.97	Jeffreys1-shortest	0.87	ML-profile	0.54
Jeffreys1	0.11	Jeffreys1	0.13	DL-Qprofile	0.94	Jeffreys2-shortest	0.72	Jeffreys2-shortest	0.66
ML	0.11	ML	0.13	DL2-Qprofile	0.94	ML-profile	0.68	Jeffreys2-central	0.85
DL	0.12	DL	0.15	Jeffreys2-shortest	0.94	Jeffreys2-central	0.65	Jeffreys1-shortest	2.10
DL2	0.12	DL2	0.15	ML-Qprofile	0.94	Jeffreys1-central	0.60	DL2-Qprofile	2.26
PM	0.12	PM	0.15	PM-Qprofile	0.94	DL-Qprofile	0.56	DL-Qprofile	2.27
REML	0.12	REML	0.15	REML-Qprofile	0.94	DL2-Qprofile	0.56	ML-Qprofile	2.27
				ML-profile	0.92	ML-Qprofile	0.56	PM-Qprofile	2.27
				Jeffreys2-central	0.78	PM-Qprofile	0.56	REML-Qprofile	2.27
				Jeffreys1-central	0.76	REML-Qprofile	0.56	Jeffreys1-central	3.80

*Note:* Methods are sorted from best to worst performance within each column (or alphabetically for ties); coverage is sorted from highest to lowest. MAE: Mean absolute error. RMSE: Root mean square error. CI: 95% confidence or credible interval. Coverage >0.94: Percent of scenarios for which coverage probability was at least 0.94.

**Table 9.** Scenarios with binary outcomes,  $k \leq 5$ ;  $\hat{\mu}$  point and interval estimation.

$\hat{\mu}$ MAE		$\hat{\mu}$ RMSE		$\hat{\mu}$ coverage		$\hat{\mu}$ coverage > 0.94		$\hat{\mu}$ CI width	
All	0.20	All	0.25	Jeffreys1-shortest	1.00	Jeffreys1-shortest	0.98	ML-profile	1.27
				Exact	0.99	Exact	0.97	Jeffreys2-shortest	1.93
				Jeffreys2-shortest	0.99	Jeffreys2-shortest	0.92	ML-HKSJ	3.08
				DL-HKSJ	0.95	ML-profile	0.75	DL-HKSJ	3.13
				DL2-HKSJ	0.95	DL-HKSJ	0.74	DL2-HKSJ	3.13
				ML-HKSJ	0.95	DL2-HKSJ	0.74	PM-HKSJ	3.13
				ML-profile	0.95	ML-HKSJ	0.74	REML-HKSJ	3.13
				PM-HKSJ	0.95	PM-HKSJ	0.74	Exact	4.04
				REML-HKSJ	0.95	REML-HKSJ	0.74	Jeffreys1-shortest	5.05

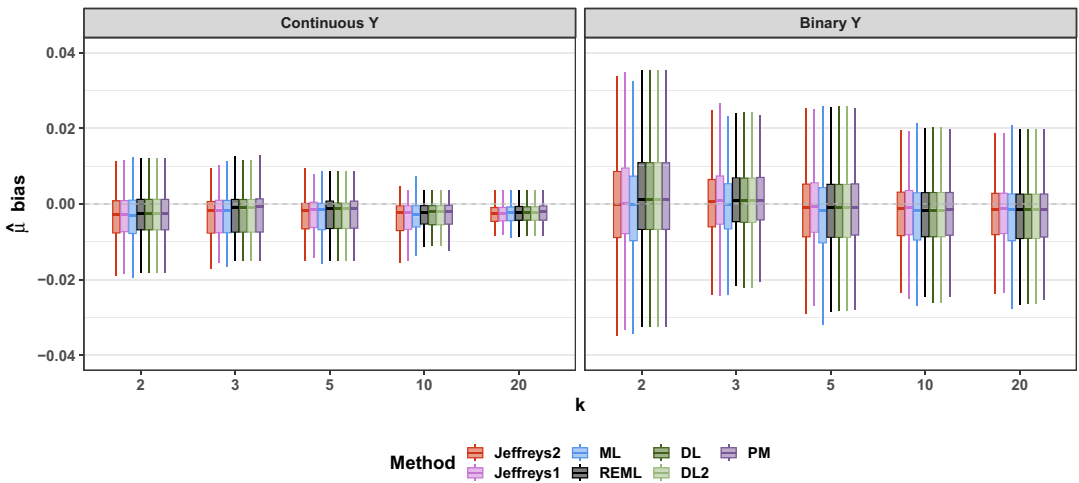
Note: Methods are sorted from best to worst performance within each column (or alphabetically for ties); coverage is sorted from highest to lowest. MAE: Mean absolute error. RMSE: Root mean square error. CI: 95% confidence or credible interval. Coverage >0.94: Percent of scenarios for which coverage probability was at least 0.94. "All": All methods performed equally to two decimal places.

**Table 10.** Scenarios with binary outcomes,  $k \leq 5$ ;  $\hat{\tau}$  point and interval estimation.

$\hat{\tau}$ MAE		$\hat{\tau}$ RMSE		$\hat{\tau}$ coverage		$\hat{\tau}$ coverage > 0.94		$\hat{\tau}$ CI width	
ML	0.14	ML	0.18	ML-profile	0.97	Jeffreys1-shortest	0.89	ML-profile	1.04
DL	0.18	Jeffreys2	0.20	DL-Qprofile	0.95	ML-profile	0.85	Jeffreys2-shortest	1.38
DL2	0.18	DL	0.24	DL2-Qprofile	0.95	Jeffreys2-shortest	0.81	Jeffreys2-central	1.80
Jeffreys2	0.18	DL2	0.24	Jeffreys1-shortest	0.95	DL-Qprofile	0.73	DL2-Qprofile	3.30
PM	0.18	REML	0.24	ML-Qprofile	0.95	ML-Qprofile	0.73	DL-Qprofile	3.32
REML	0.18	Jeffreys1	0.25	PM-Qprofile	0.95	PM-Qprofile	0.73	ML-Qprofile	3.32
Jeffreys1	0.22	PM	0.25	REML-Qprofile	0.95	REML-Qprofile	0.73	PM-Qprofile	3.32
				Jeffreys2-shortest	0.93	DL2-Qprofile	0.72	REML-Qprofile	3.32
				Jeffreys2-central	0.65	Jeffreys2-central	0.52	Jeffreys1-shortest	4.43
				Jeffreys1-central	0.59	Jeffreys1-central	0.43	Jeffreys1-central	8.08

*Note:* Methods are sorted from best to worst performance within each column (or alphabetically for ties); coverage is sorted from highest to lowest. MAE: Mean absolute error. RMSE: Root mean square error. CI: 95% confidence or credible interval. Coverage >0.94: Percent of scenarios for which coverage probability was at least 0.94.





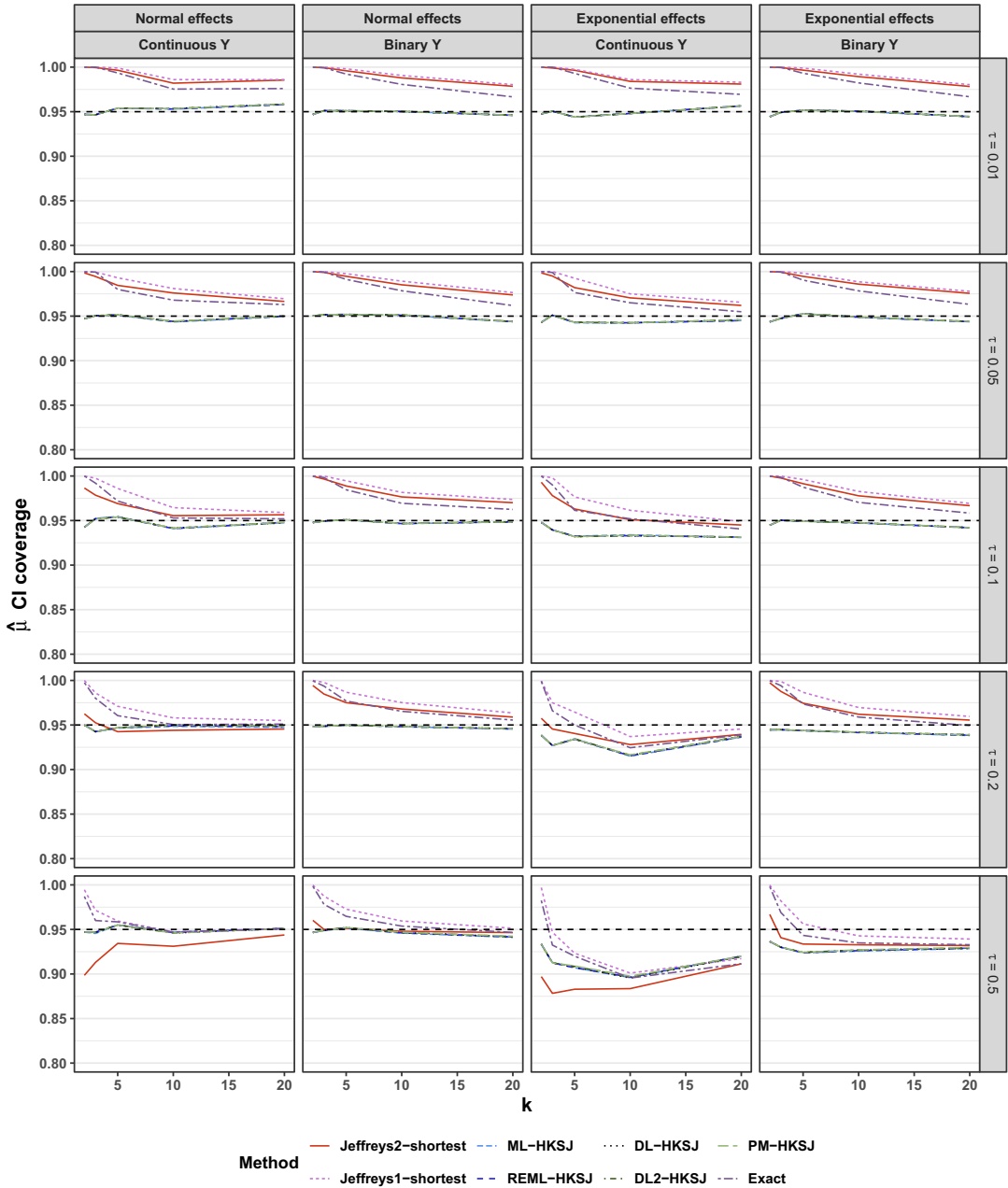
**Figure 6.** Bias of  $\hat{\mu}$ ; all scenarios. Hinges of each boxplot are the 25th, 50th, and 75th percentiles. The upper and lower whiskers extend from the hinge to the minimum or maximum value that is no more than  $1.5 \times$  (interquartile range) from the nearest hinge.

Figure 7 shows the coverage of 95% intervals. All frequentist methods with HKSJ intervals performed similarly to one another. In scenarios with normal population effects, these methods' performances were minimally affected by  $k$  and  $\tau$ , and coverage was  $>94\%$  in 80% of scenarios. This is a somewhat pessimistic portrayal because coverages for these methods were also rarely below  $\sim 93\%$ . ML-profile had coverage  $>94\%$  in 71% of scenarios with normal effects, but unlike the HKSJ methods, the coverage of ML-profile varied substantially across scenarios. In particular, this method had close to nominal coverage for intermediate levels of heterogeneity and for  $k = 20$ , but exhibited under-coverage for higher heterogeneity values (e.g.,  $\tau \geq 0.20$ ). The exact interval exhibited over-coverage for smaller values of  $k$  and close to nominal coverage for  $k = 20$ . All of these findings are consistent with previous simulation studies.<sup>10,18</sup>

Jeffreys1-shortest and Jeffreys2-shortest intervals had coverage  $>94\%$  in 98% and in 88% of scenarios with normal population effects, respectively. This exceeded the 80% and 71% seen for HKSJ intervals and ML-profile intervals, respectively. In individual scenarios, Jeffreys1-shortest and Jeffreys2-shortest intervals both typically exhibited over-coverage or nominal coverage with one exception: Jeffreys2-shortest intervals exhibited mild under-coverage ( $\sim 89\text{--}93\%$ ) for very small meta-analyses ( $k \leq 5$ ) that also had a continuous outcome and high heterogeneity ( $\tau = 0.50$ ).

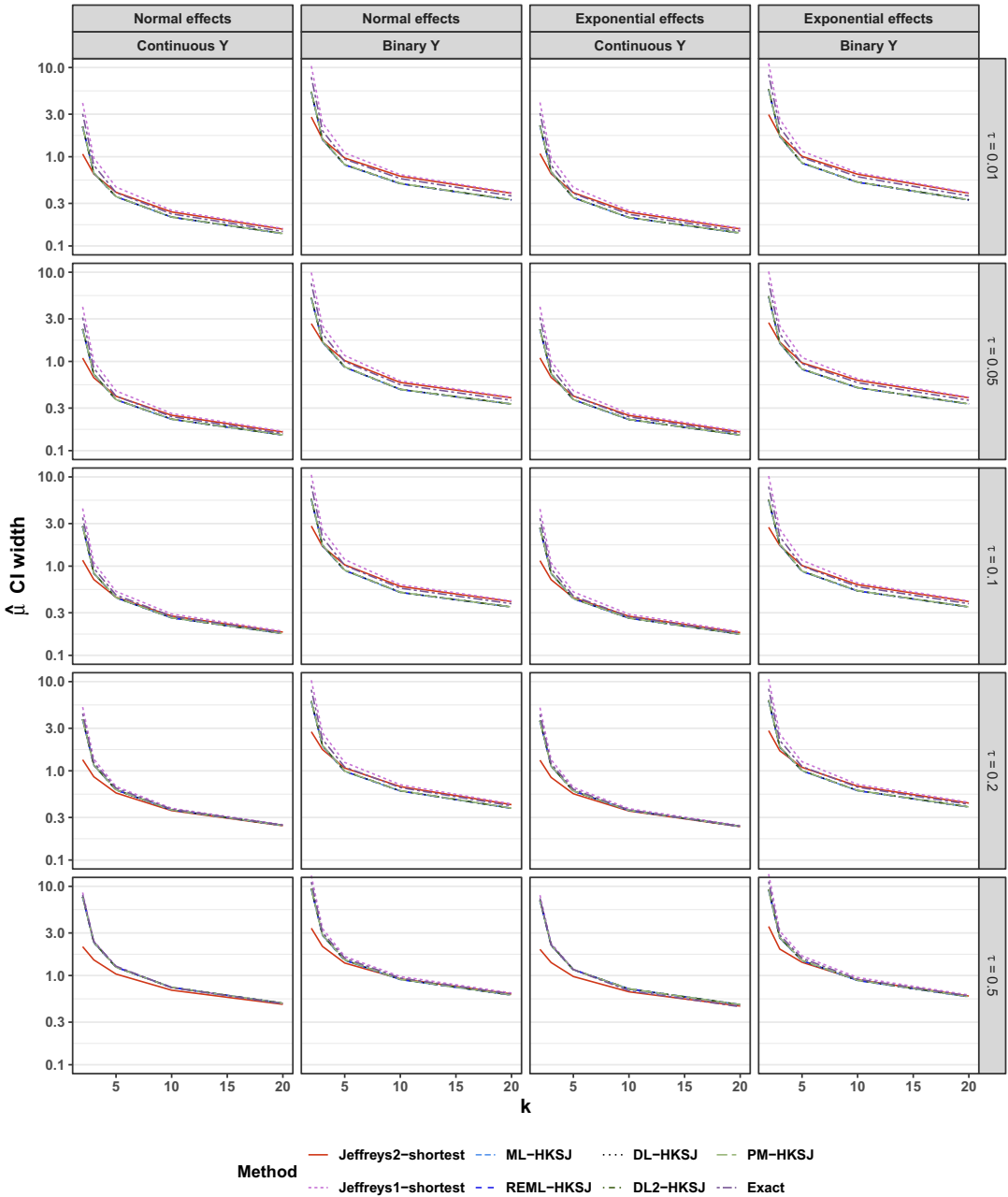
Figure 8 shows the width of 95% intervals. For  $k < 10$ , the different intervals' widths varied, sometimes substantially. In these scenarios, the ML-profile interval was consistently the narrowest, and was often considerably so for very small meta-analyses. The Jeffreys1-shortest interval was typically the widest of all, especially for very small meta-analyses. On the other hand, the Jeffreys2-shortest interval was typically the second-narrowest after ML-profile, and was substantially narrower than all HKSJ intervals for very small meta-analyses. It may be counterintuitive that the Jeffreys2-shortest interval was narrower than the HKSJ intervals while more consistently achieving at least nominal coverage; we explain this finding in Section 4.4.3 below. For  $k \geq 10$  and continuous outcomes, all types of intervals had nearly identical widths. For  $k \geq 10$  and binary outcomes, both Jeffreys intervals and the exact interval were slightly wider than those of the HKSJ methods, although this should be interpreted in light of the frequentist methods' slight under-coverage in these scenarios (Figure 7).

In scenarios with exponential population effects, all methods' relative performances were similar, although coverages declined somewhat when heterogeneity was high ( $\tau = 0.50$ ). This, too, is consistent



**Figure 7.** Coverage of CI for  $\hat{\mu}$ . Lines are slightly staggered horizontally for visibility. Lines are mean performances across scenarios, conditional on  $k$ ,  $\tau$ , the distribution of population effects, and the outcome type. All HKSJ methods performed very similarly, so their overlapping lines look like a single grey line.

with previous simulation studies.<sup>10</sup> Section 3 of the Supplementary Material provides additional results stratified by outcome type. First, results are shown for scenarios with  $k = 100$ , since these scenarios are excluded from all results in the main text. In those scenarios, as expected from theory, all point estimates performed very similarly regardless of outcome type. For binary outcomes, most methods' coverage probabilities declined somewhat at  $k = 100$ . This finding is consistent with previous simulation results



**Figure 8.** Width of CI for  $\hat{\mu}$ . Lines are slightly staggered horizontally for visibility. Lines are mean performances across scenarios, conditional on  $k$ ,  $\tau$ , the distribution of population effects, and the outcome type. Y-axis is on log scale.

involving rare binary outcomes (Langan et al. (2019)<sup>7</sup>; Appendix Figure 4) and likely reflects known two sources of misspecification when meta-analyzing log-odds ratios. In particular: (1) estimated log-odds ratios are correlated with their estimated standard errors; and (2) the conventional variance estimate is an imperfect approximation especially when there are zero cell counts, a problem that occurs even when adding positive constants to each cell.<sup>65,66</sup> We return to these issues in the Discussion.

In these scenarios, Jeffreys methods retained closer to nominal coverage than did the frequentist methods. Additional supplementary tables stratify the results in the main text (i.e., scenarios with  $k \leq 20$ ) into those with fixed versus varying  $N$  across studies. In all of these strata, the relative rankings of methods' performances were quite similar to those in the aggregate analyses.

#### 4.4.3. Discussion of results regarding $\mu$

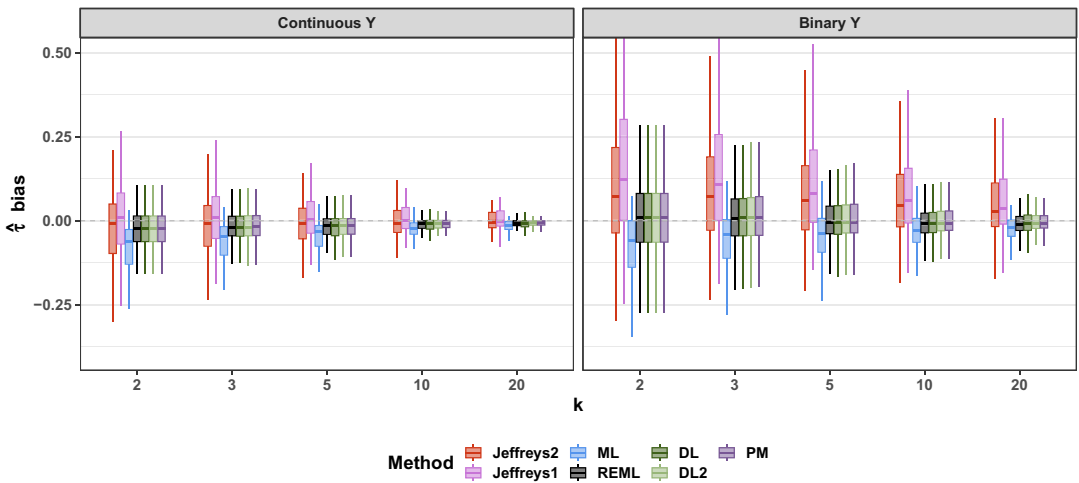
For small meta-analyses ( $k \leq 20$ ) with binary outcomes, Jeffreys2-shortest may be a useful method since its intervals had at least nominal coverage (with normal effects) and yet were often considerably narrower than all others except ML-profile, whose coverage was inconsistent across scenarios. To illustrate, we provide some numerical comparisons between Jeffreys2-shortest and REML-HKSJ intervals for meta-analyses of binary outcomes. We compare to a single type of frequentist interval for simplicity. In scenarios with binary outcomes and normal population effects, Jeffreys2-shortest had coverage  $>94\%$  in 90% of scenarios, whereas REML-HKSJ did so in 80% of scenarios. Accordingly, Jeffreys2-shortest had coverage at least equal to that of REML-HKSJ in 85% of scenarios. At the same time, the Jeffreys2-shortest interval was on average 27% narrower than the REML-HKSJ interval; and in meta-analyses with  $k \leq 5$ , this efficiency improvement increased to 51%. For binary outcomes, Jeffreys1-shortest did not appear to have clear advantages over Jeffreys2-shortest or the other methods, since the Jeffreys1-shortest interval was wider than even that of the exact method.

For small meta-analyses with continuous outcomes, more caution is warranted when using Jeffreys2-shortest intervals, since they exhibit mild under-coverage ( $\sim 89\text{--}93\%$ ) for very small meta-analyses ( $k \leq 5$ ) that also had high heterogeneity. Since Jeffreys2-shortest provided only modest improvements in efficiency for meta-analyses of continuous outcomes once  $k > 5$ , it may be preferable to conservatively use a frequentist method with an HKSJ interval for continuous outcomes, regardless of  $k$ . Although the Jeffreys1-shortest interval did, in general, retain at least nominal coverage for continuous outcomes, this interval was again wider than the exact interval, and was considerably wider than the HKSJ intervals.

As noted above, it may be counterintuitive that the Jeffreys2-shortest interval was typically narrower than the HKSJ intervals while more consistently achieving at least nominal coverage. There are two reasons for this finding. First, whereas HKSJ intervals for  $\mu$  are always symmetric on the analyzed effect scale (i.e., Hedges'  $g$  for continuous outcomes and log-odds ratio for binary outcomes), the Jeffreys1-shortest and Jeffreys2-shortest intervals can be symmetric or asymmetric depending on the shape of the posterior (Section 2.3 of the Supplementary Material). Second, within a given scenario, the width of the Jeffreys2-shortest interval was typically much less variable across repeated samples than the HKSJ intervals. Thus, in many scenarios in which the Jeffreys2-shortest interval exhibited over-coverage but comparison methods exhibited nominal or less than nominal coverage, this was because the HKSJ methods often yielded extremely wide intervals under repeated sampling, whereas the Jeffreys2-shortest intervals were bounded within a narrower range (Section 2.3 of the Supplementary Material).

#### 4.4.4. Point and interval estimation for $\tau$

For both continuous and binary outcomes, results for point and interval estimation depended on whether  $\tau$  was near the boundary value of zero, especially for the Jeffreys methods. Regarding point estimation, the frequentist methods, especially ML, typically showed a slight negative bias (Figure 9). Point estimates from Jeffreys1 and Jeffreys2 varied more in the sign and magnitude of bias than did the frequentist point estimates (Figure 9). Regarding MAE and RMSE, the frequentist methods DL, DL2, REML, and PM were comparable to one another. In contrast, ML often performed slightly better on these metrics (Figures 10 and 11). Jeffreys1 and Jeffreys2 had comparable MAE and RMSE to one another. Relative to the frequentist methods, Jeffreys1 and Jeffreys2 typically showed comparable MAE and RMSE at midrange values of  $\tau$  (e.g.,  $\tau = 0.10$ ), showed better MAE and RMSE for  $\tau > 0.10$ , and showed worse MAE and RMSE for  $\tau < 0.10$ . These patterns were more pronounced for binary



**Figure 9.** Bias of  $\hat{\tau}$ ; all scenarios. Hinges of each boxplot are the 25th, 50th, and 75th percentiles. The upper and lower whiskers extend from the hinge to the minimum or maximum value that is no more than  $1.5 \times$  (interquartile range) from the nearest hinge.

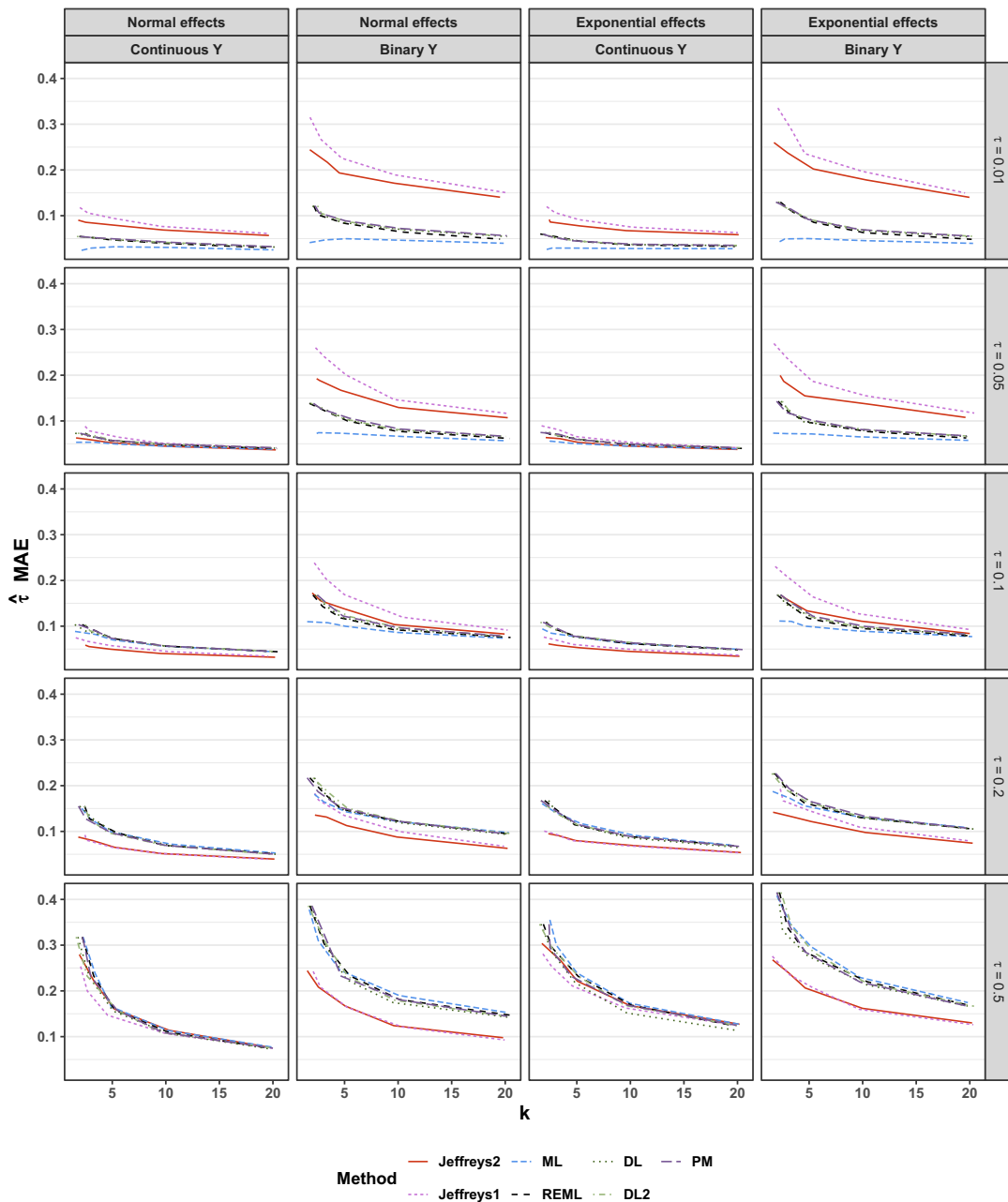
outcomes, though the relative rankings of methods were similar for both outcome types. The patterns were similar for both normal and exponential population effects.

Regarding interval estimation, pilot tests with the bootstrap methods again suggested that these methods performed relatively poorly compared to the other methods (Sections 3.7 and 3.8 of the Supplementary Material), so we again omit the bootstrap methods from the main text. Figure 12 shows the coverage of 95% intervals. With normal population effects, all Q-profile intervals performed comparably to one another and showed close to nominal coverage ( $>94\%$  in 83% of scenarios). ML-profile typically showed nominal coverage or over-coverage in the large majority of scenarios; in scenarios with normal effects, the coverage of these intervals was  $>94\%$  in 82% of scenarios, similar to the Q-profile methods. However, ML-profile did show under-coverage for smaller meta-analyses that also had high heterogeneity. This under-coverage was minimal for binary outcomes ( $\sim 90\%$  at minimum), but could be substantial for continuous outcomes ( $\sim 75\%$  at minimum).

Jeffreys1-shortest showed at least nominal coverage for  $\tau > 0.01$ , but showed substantial under-coverage when  $\tau = 0.01$ . Jeffreys2-shortest behaved similarly, except additionally showed under-coverage for meta-analyses of continuous outcomes with high heterogeneity ( $\tau = 0.50$ ), especially for  $k \leq 5$ . The coverage of Jeffreys1-shortest and Jeffreys2-shortest was  $>94\%$  in, respectively, 83% and 74% of scenarios. Both Jeffreys1-central and Jeffreys2-central performed considerably worse (i.e., showed more severe under-coverage) than Jeffreys1-shortest and Jeffreys2-shortest for smaller values of  $\tau$ : in scenarios with normal population effects, coverage of Jeffreys1-central and Jeffreys2-central was  $>94\%$  in, respectively, 54% and 56% of scenarios. This under-coverage reflects overestimation of  $\tau$  when it was near the boundary of the parameter space.

Figure 13 shows the width of 95% intervals. We now discuss only the methods that had the highest rates of at least nominal coverage, so exclude discussion of Jeffreys2-shortest, Jeffreys1-central, and Jeffreys2-central. The widths of the various Q-profile intervals the Jeffreys1-shortest intervals were comparable, but the ML-profile intervals were typically considerably narrower, especially for very small meta-analyses.

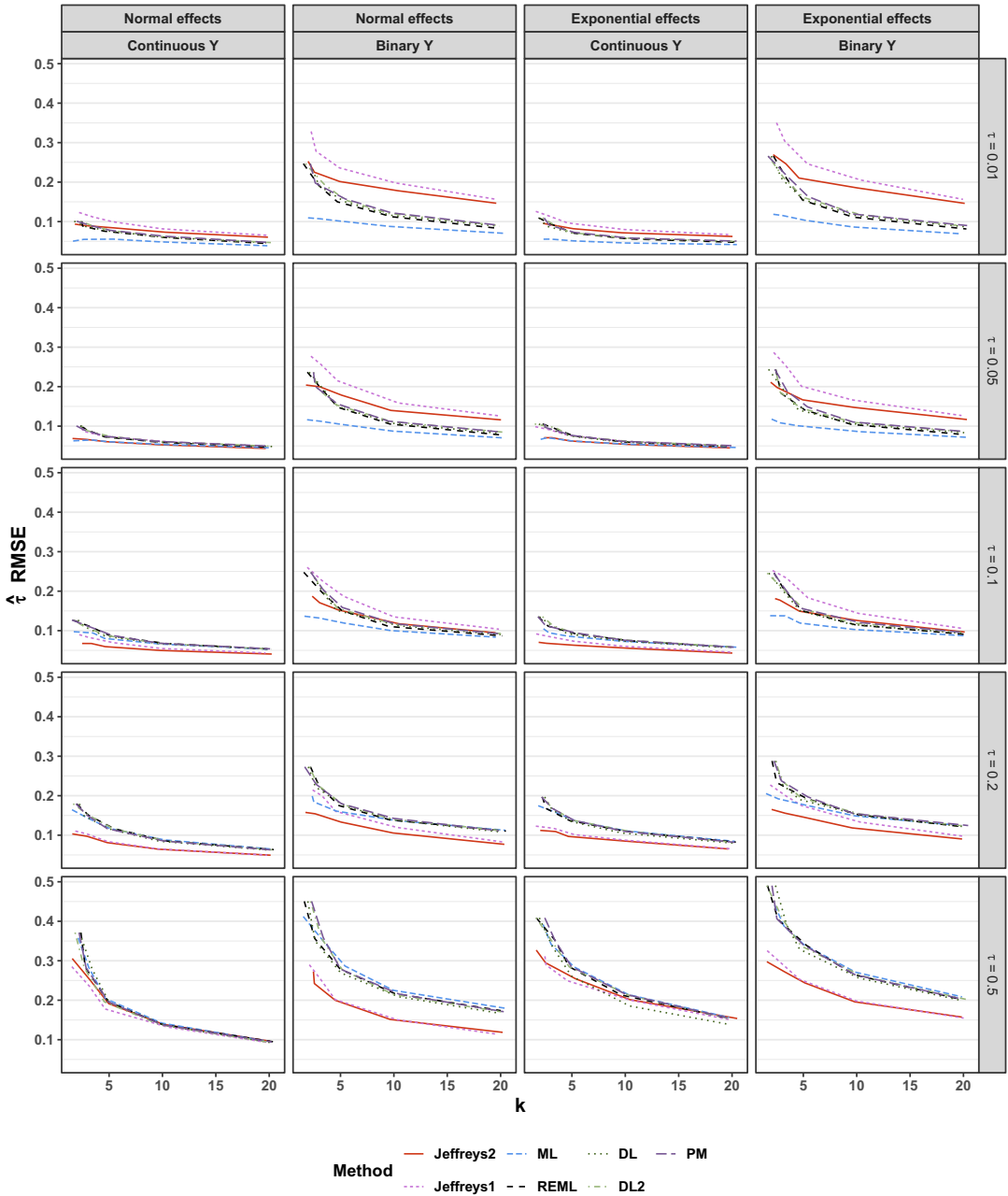
In scenarios with exponential population effects, all methods' relative performances for estimation and interval estimation on  $\tau$  were similar, although coverages declined for all methods. Additional stratified results (Section 3 of the Supplementary Material) suggest that patterns of performance were also comparable for  $k = 100$  and for fixed versus varying  $N$  across studies.



**Figure 10.** MAE of  $\hat{\tau}$ . Lines are slightly staggered horizontally for visibility. Lines are mean performances across scenarios, conditional on  $k$ ,  $\tau$ , the distribution of population effects, and the outcome type.

4.4.5. Discussion of results regarding  $\tau$

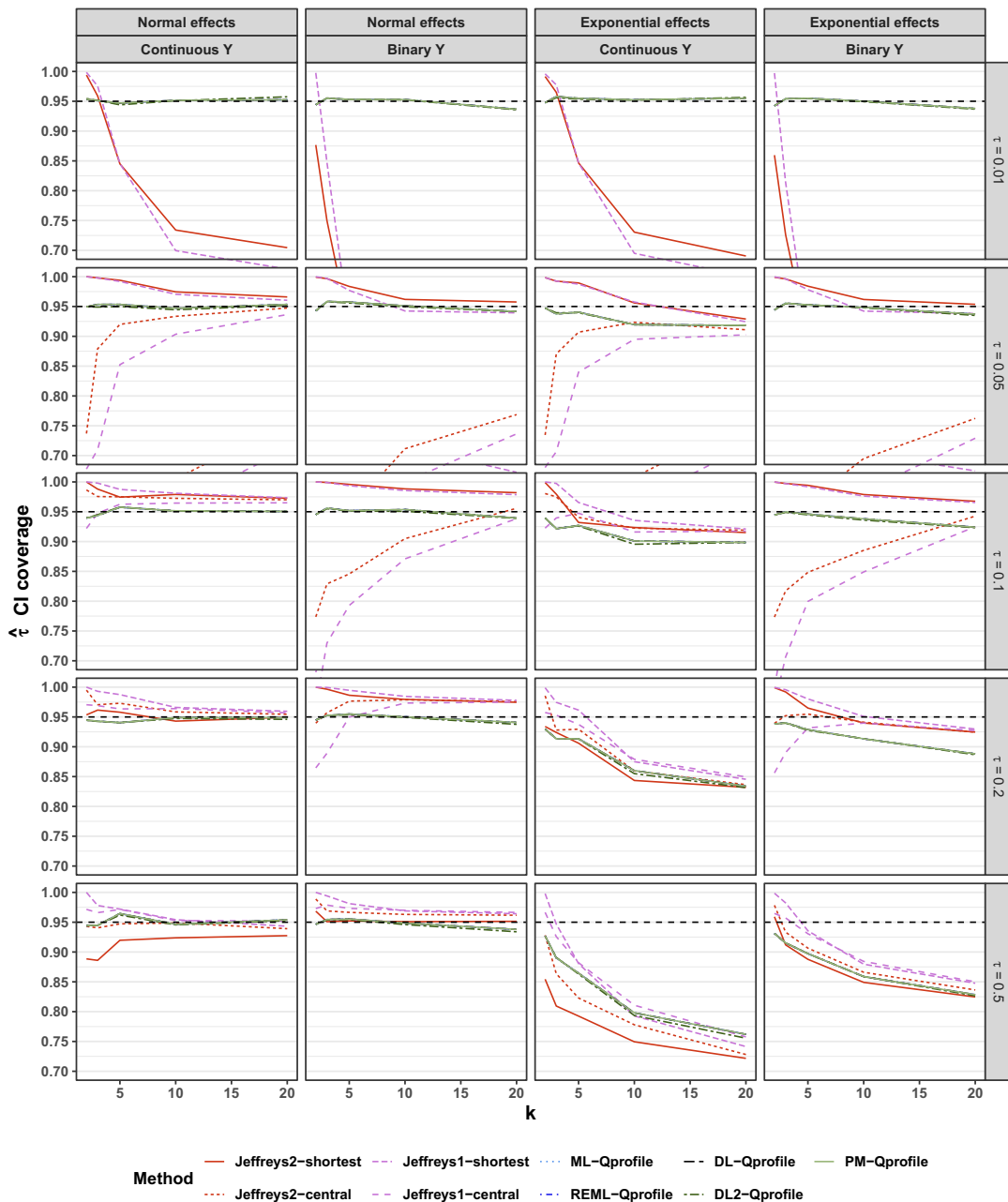
For point estimation of  $\tau$ , no method emerged as clearly optimal, since methods' performances depended strongly on  $\tau$  itself. The low coverage of the Jeffreys methods occurred when  $\tau$  was near zero (the boundary of the parameter space). This reflects overestimation of  $\tau$ , which is often viewed as conservative in the context of random-effects meta-analysis. Regarding interval estimation for  $\tau$ , the frequentist estimators with Q-profile or ML-profile intervals appear preferable to the



**Figure 11.** RMSE of  $\hat{\tau}$ . Lines are slightly staggered horizontally for visibility. Lines are mean performances across scenarios, conditional on  $k$ ,  $\tau$ , the distribution of population effects, and the outcome type.

Jeffreys methods. Of the two Jeffreys priors and two types of intervals, Jeffreys1-shortest is the only one whose coverage was competitive with that of the frequentist methods. However, since Jeffreys1-shortest intervals were somewhat wider than those of the frequentist intervals, this method does not seem to offer an overall advantage over the frequentist intervals. The Q-profile intervals performed slightly more consistently across scenarios than did ML-profile, although average performances were similar. ML-profile intervals were, however, considerably narrower than the Q-profile intervals.

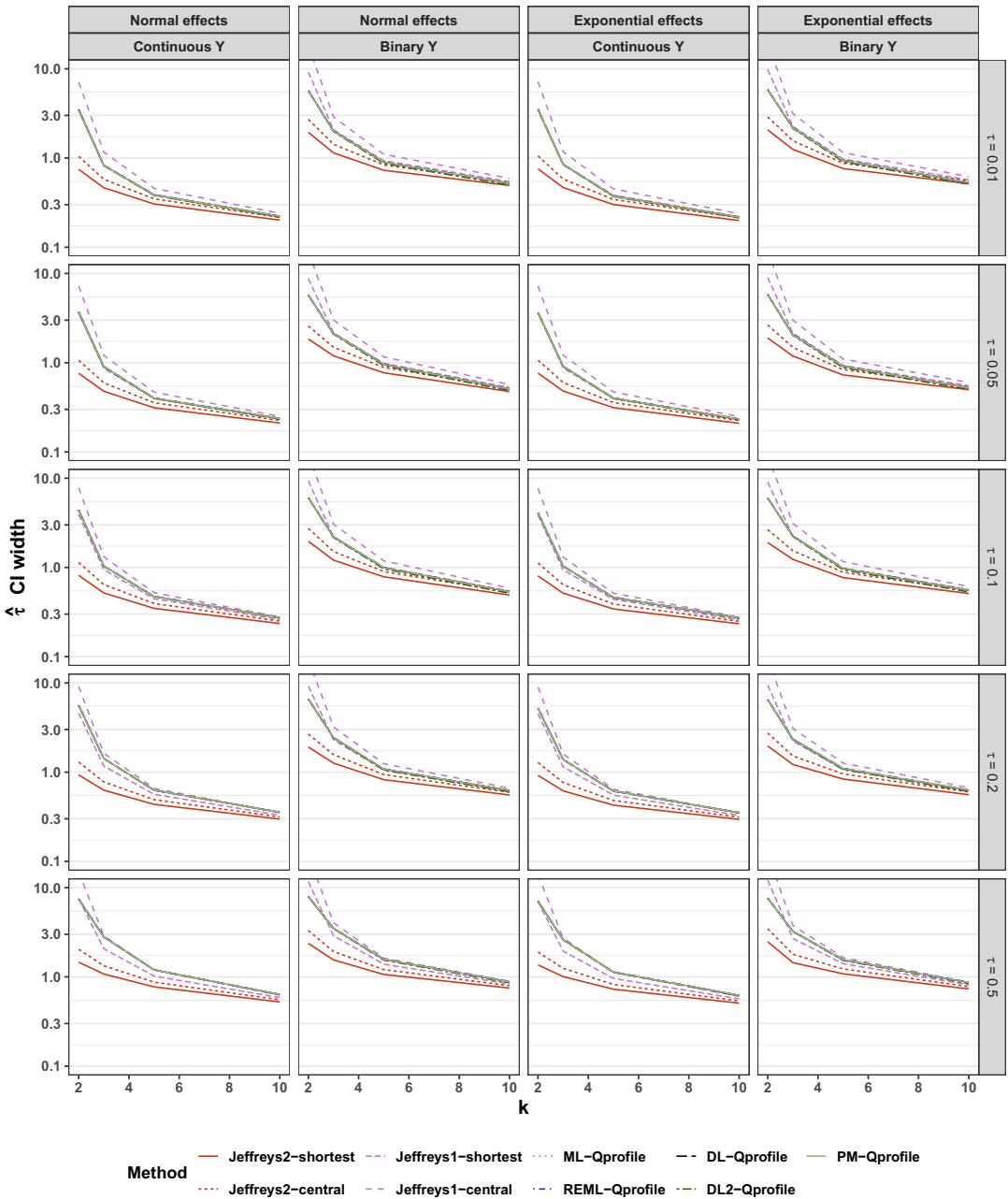




**Figure 12.** Coverage of CI for  $\hat{\tau}$ . Lines are slightly staggered horizontally for visibility. Lines are mean performances across scenarios, conditional on  $k$ ,  $\tau$ , the distribution of population effects, and the outcome type.

#### 4.5. Overall conclusions

All methods performed similarly for point estimation of  $\mu$ . In general, standard frequentist methods with HKSJ intervals for  $\mu$  and Q-profile intervals for  $\tau$  performed the most consistently across outcome types. Jeffreys2-shortest also showed consistently strong performance for meta-analyses of binary outcomes, and yielded substantially narrower intervals than frequentist methods. However,



**Figure 13.** Width of CI for  $\hat{\tau}$ . Lines are slightly staggered horizontally for visibility. Lines are mean performances across scenarios, conditional on  $k$ ,  $\tau$ , the distribution of population effects, and the outcome type. Y-axis is on log scale.

the Jeffreys2-shortest interval did not perform as consistently for continuous outcomes; this method exhibited mild under-coverage for very small meta-analyses with high heterogeneity. Regarding point estimation of  $\tau$ , all methods again performed comparably on average, though the optimal method depended on the value of  $\tau$  itself. Regarding interval estimation for  $\tau$ , the Q-profile method arguably performed the best and behaved consistently across scenarios.

Overall, for small meta-analyses of continuous outcomes, we would recommend standard frequentist methods with an HKSJ interval for  $\mu$  and a Q-profile interval for  $\tau$ , consistent with previous recommendations. However, for small meta-analyses of binary outcomes, Jeffreys2 may be preferable over standard frequentist methods if the meta-analyst is primarily interested in point and interval estimation for  $\mu$ , potentially along with point estimation of  $\tau$  (although, again, the best-performing method for estimating  $\tau$  depended on the value of  $\tau$  itself). This is because the Jeffreys2-shortest interval more frequently had at least nominal coverage, yet was substantially more precise. If the meta-analyst is also interested in obtaining an interval for  $\tau$ , then using a frequentist method with a Q-profile interval for  $\tau$  would likely provide closer to nominal coverage for  $\tau$  than would Jeffrey2-shortest; however, this would likely sacrifice a substantial amount of precision for  $\mu$ .

## 5. Applied example

Zito et al.<sup>67</sup> meta-analyzed randomized trials that compared various diagnostic strategies for detecting coronary artery disease (CAD) in patients experiencing CAD-related symptoms. The authors conducted meta-analyses for each pairwise comparison between multiple diagnostic methods; for simplicity, we focus on studies comparing coronary computed tomography angiography (CCTA) to stress single-photon emission computed tomography myocardial perfusion imaging (SPECT-MPI). We replicated the authors' meta-analyses for each of six outcomes: cardiovascular death and myocardial infarction ( $k = 2$ ), all-cause death ( $k = 3$ ), myocardial infarction ( $k = 2$ ), index invasive coronary angiography (ICA) ( $k = 4$ ), index revascularization ( $k = 4$ ), and downstream testing ( $k = 4$ ). The authors' meta-analyses<sup>67</sup> used the DL method and used Wald rather than HKSJ confidence intervals.<sup>iii</sup> We scraped study-level summary statistics from the published forest plots and re-analyzed the studies assessing each outcome using DL, REML, the exact method, Jeffreys1-shortest, and Jeffreys2-shortest. For DL and REML, we used HKSJ intervals, following established recommendations.<sup>7,12–14</sup> Since our simulation study suggested relatively minor differences between the various frequentist methods with HKSJ intervals, we focus only on DL and REML for brevity. All code and data required to reproduce the applied example are publicly available and documented (<https://osf.io/9qfah>).

Figure 2 shows the Jeffreys1 and Jeffreys2 priors for a single outcome (all-cause death), and Figure 3 shows the resulting joint posterior under the Jeffreys2 prior. Figure 5 shows all methods' point estimates and intervals for  $\hat{\mu}$  for all outcomes; a similar forest plot for heterogeneity estimates appears in Section 4 of the Supplementary Material. As in the simulation study, all point estimates were nearly identical, but the Jeffreys2-shortest interval was often considerably narrower than those from Jeffreys1-shortest, REML, DL, and the exact method. Across all six outcomes, the Jeffreys2-shortest interval on the log-odds scale was on average 45% narrower than the narrowest interval from the other methods. For the meta-analyses of only two studies, this improvement in precision increased to 112%.

## 6. Discussion

To the best of our knowledge, this paper provides the first empirical assessment of the Jeffreys2 prior in meta-analysis. We compared point estimates and intervals from the Jeffreys2 prior to those of the Jeffreys1 prior and to several of the best-performing parametric, semiparametric, and nonparametric frequentist methods. Extending previous simulation studies on the Jeffreys1 prior, we additionally considered different types of Bayesian point estimates and intervals, and we considered point and interval estimation for both  $\mu$  and  $\tau$ . As summarized in Section 4.5, for small meta-analyses of binary outcomes, Jeffreys2 may be preferable over standard frequentist methods for point and interval estimation for  $\mu$ , providing improvements in efficiency that can be substantial. However, for small meta-analyses of continuous outcomes, standard frequentist methods with HKSJ intervals for  $\mu$  and Q-profile CIs for  $\tau$  seem to be the best choices, avoiding the mild under-coverage that Jeffreys2-shortest intervals can sometimes exhibit for very small meta-analyses with high heterogeneity. For both outcome types, the best-performing method for point estimation of  $\tau$  varied according to  $\tau$  itself. When  $\tau$  is very

small, the Jeffreys methods performed conservatively in that they typically overestimate  $\tau$ . Finally, we showed that the Jeffreys2 prior has a straightforward generalization to the case of meta-regression (Section 1 of the Supplementary Material).

Given our interest in the frequentist properties of Jeffreys priors as the Firth correction to ML estimates, we have treated point and interval estimation from a frequentist perspective. For example, our simulation study considered the coverage of 95% intervals estimated from repeated samples that were generated from fixed values of the parameters. In contrast, in Bayesian inference, the parameters are viewed as random draws from the prior, rather than as fixed quantities. The Bayesian framework does permit empirical assessment of certain analogs to coverage, but doing so involves drawing repeated samples from parameters sampled from the prior, rather than from parameters held constant.<sup>9,68,69</sup> As an additional complication, performing these Bayesian calibration checks requires a proper prior from which to sample, yet both Jeffreys priors are improper.<sup>68</sup> Cook et al. (2006) argued that this difficulty in assessing calibration with improper priors is a disadvantage of using such priors in the first place.<sup>68</sup> Given our interest in methods' frequentist motivations and their frequentist empirical properties, we did not consider any of the numerous other Bayesian priors that have been proposed for meta-analysis (e.g., as reviewed by Röver (2020)<sup>9</sup>). It is somewhat difficult to compare the performance of standard frequentist methods to Bayesian methods that lack a frequentist interpretation, which is perhaps why many previous simulation studies have not included any Bayesian methods<sup>7,11</sup> (but with exceptions<sup>15–17</sup>).

Our simulation study had other limitations. First, we considered only one form of model misspecification, namely exponentially distributed population effects, and found that methods' relative rankings were largely unaffected. However, we did not assess any other forms of misspecification, such as more severe departures from normality<sup>10</sup> or clustered population effects. Second, for meta-analyses of binary outcomes, we considered only standard inverse-variance weighted meta-analysis, but arm-based approaches may have better statistical properties.<sup>66</sup> On the other hand, arm-based methods can introduce bias due to non-exchangeability across trials,<sup>70,71</sup> and inverse-variance meta-analysis more readily accommodates the possibility that studies adjust for covariates, and may be more feasible when original papers reported only limited summary statistics. Additionally, assessing inverse-variance meta-analysis provides a more direct comparison to previous simulation studies.<sup>11</sup> Third, the two within-study estimators we used, namely log-odds ratios and Hedges'  $g$ , both involve approximations which may introduce slight finite-sample biases of their own. Such decisions can nontrivially affect the results of simulation studies,<sup>72</sup> and we used these estimators to ensure direct comparability with previous simulation studies.<sup>7</sup> Additionally, these two measures are among the most commonly used in meta-analyses.<sup>73</sup> Future work could explore relative performances with effect measures that do not require approximations, such as raw mean differences, although these are not frequently used in practice.<sup>73</sup> Fourth, we considered only two estimands,  $\mu$  and  $\tau$ , but these alone provide a limited summary of the random-effects distribution. Additional metrics that can be informative include the percentage of population effects exceeding a chosen threshold for a meaningful effect size<sup>47,55,74</sup>; the prediction interval for a new population effect<sup>54,75</sup>; and shrinkage estimates for each study's population effect.<sup>75,76</sup> An advantage of Bayesian estimation is that such metrics can be obtained readily from the posterior; several are implemented in the R package `bayesmeta`.<sup>9</sup> Future simulation studies could consider these estimands and intervals as well. Fifth, we made the usual assumption that any estimation error in the within-study standard errors is negligible. We did not assess the extent to which this approximation compromised interval estimation. A number of approaches have been proposed to accommodate this form of estimation error; perhaps future work could incorporate these developments into the Jeffreys priors.<sup>77–80</sup>

Our work remains a preliminary investigation of Jeffreys1 and Jeffreys2 priors. We would particularly encourage future work to consider other generalizations to these priors, besides our generalization to meta-regression. For example, as noted in Introduction, we recently found that a Jeffreys prior on  $\mu$  and  $\tau$  performed well for an estimation problem involving severe  $p$ -hacking, which required estimating the parameters of a truncated distribution.<sup>34</sup> Certain selection models for publication bias

lead to related distributions that involve step functions in the publication probability.<sup>81</sup> These models can perform poorly for small meta-analyses, often exhibiting extremely wide intervals for parameters related to publication bias severity.<sup>82,83</sup> Might using a Jeffreys prior on  $\mu$ ,  $\tau$ , and the bias parameters also improve these models' performance for small meta-analyses? Additional extensions could include accommodating clustered population effects. We look forward to future research along these lines.

**Acknowledgements.** Christian Röver provided helpful discussions and implemented the Jeffreys2 prior in his R package, `bayesmeta`. Dean Langan shared code from his simulation study. Annamaria Guolo provided advice on the use of her R package, `metaLik`.

**Author contributions.** M.B.M. conducted this research.

**Competing interest statement.** The author declares that no competing interests exist.

**Data availability statement.** All code and data required to reproduce the simulation study and applied example are publicly available and documented (<https://osf.io/9qfah>).

**Funding statement.** This research was supported by National Institutes of Health grants R01 LM013866, UL1TR003142, P30CA124435, and P30DK116074. The funders had no role in the design, conduct, or reporting.

**Supplementary material.** The supplementary material for this article can be found at <https://doi.org/10.1017/rsm.2024.2>.

## Notes

- i. The invariance property does not mean, however, that the posterior mode of  $\tau$  under a  $\tau$ -parameterization is equal to the square-root of the posterior mode of  $\tau^2$  under a  $\tau^2$ -parameterization, for example. Rather, the usual change-of-variables expression would apply.
- ii. Specifically, in the context of bias correction for  $p$ -hacking, Mathur (2024)<sup>34</sup> described meta-analyzing only studies that are nonsignificant or that have negative point estimates. In that paper, we derived the two-parameter Jeffreys prior for a version of the likelihood that is appropriately right-truncated to reflect the inclusion of only a subset of studies. The present Jeffreys2 prior is the special case in which there is no truncation, i.e., the limit as the truncation threshold approaches  $\infty$ .
- iii. The paper states that HKSJ intervals were used, but our re-analyses replicated their results only when we used Wald intervals. Through correspondence with the authors and journal editors, we confirmed that Wald intervals were indeed used (S. Chang and J. Cornell, 14 March 2024).

## References

- [1] DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control. Clin. Trials*. 1986;7(3): 177–188.
- [2] DerSimonian R, Kacker R. Random-effects model for meta-analysis of clinical trials: an update. *Contemp. Clin. Trials*. 2007;28(2): 105–114.
- [3] Paule RC, Mandel J. Consensus values and weighting factors. *J. Res. Nat. Bureau Stand.* 1982;87(5): 377.
- [4] van Aert RCM, Jackson D. Multistep estimators of the between-study variance: The relationship with the Paule-Mandel estimator. *Stat. Med.* 2018;37(17): 2616–2629.
- [5] Raudenbush SW, Bryk AS. Empirical Bayes meta-analysis. *J. Educ. Stat.* 1985;10(2): 75–98.
- [6] Harville DA. Maximum likelihood approaches to variance component estimation and to related problems. *J. Amer. Stat. Assoc.* 1977;72(358): 320–338.
- [7] Langan D, Higgins JPT, Jackson D, et al. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Res. Synth. Methods*. 2019;10(1): 83–98.
- [8] Van Erp S, Verhagen J, Grasman R, et al. Estimates of between-study heterogeneity for 705 meta-analyses reported in Psychological Bulletin from 1990–2013. *J. Open Psychol. Data*. 2017;5(1).
- [9] Röver C. Bayesian random-effects meta-analysis using the bayesmeta R package. *J. Stat. Softw.* 2020;93(6).
- [10] Kontopantelis E, Reeves D. Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: a simulation study. *Stat. Methods Med. Res.* 2012;21(4): 409–426.
- [11] Langan D, Higgins JPT, Simmonds M. Comparative performance of heterogeneity variance estimators in meta-analysis: a review of simulation studies. *Res. Synth. Methods*. 2017;8(2): 181–198.
- [12] Int'Hout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Med. Res. Methodol.* 2014;14(1): 1.
- [13] Sidik K, Jonkman JN. A simple confidence interval for meta-analysis. *Stat. Med.* 2002;21(21): 3153–3159.
- [14] Knapp G, Hartung J. Improved tests for a random effects meta-regression with a single covariate. *Stat. Med.* 2003;22(17): 2693–2710.

- [15] Bodnar O, Link A, Arendacká B, et al. Bayesian estimation in random effects meta-analysis using a non-informative prior. *Stat. Med.* 2017;36(2): 378–399.
- [16] Friede T, Röver C, Wandel S, et al. Meta-analysis of few small studies in orphan diseases. *Res. Synth. Methods.* 2017;8(1): 79–91.
- [17] Seide SE, Röver C, Friede T. Likelihood-based random-effects meta-analysis with few studies: empirical and simulation studies. *BMC Med. Res. Methodol.* 2019;19: 1–14.
- [18] Michael H, Thornton S, Xie M, et al. Exact inference on the random-effects model for meta-analyses with few studies. *Biom.* 2019;75(2): 485–493.
- [19] Viechtbauer W. Confidence intervals for the amount of heterogeneity in meta-analysis. *Stat. Med.* 2007;26(1): 37–52.
- [20] Jeffreys H. An invariant form for the prior probability in estimation problems. *Proc. Royal Soc. London. Series A. Math. Phys. Sci.* 1946;186(1007): 453–461.
- [21] Gelman A, Carlin JB, Stern HS, et al. *Bayesian Data Analysis*. 3rd ed. CRC Press; 2014.
- [22] Cencov NN. *Statistical Decision Rules and Optimal Inference*. Vol 53. American Mathematical Society; 2000.
- [23] Datta GS, Ghosh M. On the invariance of noninformative priors. *Ann Stat.* 1996;24(1): 141–159.
- [24] Kass RE, Wasserman L. Formal rules for selecting prior distributions: A review and annotated bibliography. *J. Amer. Stat. Assoc.* 1996;435: 1343–1370.
- [25] Bernardo JM. Reference posterior distributions for Bayesian inference. *J. Royal Stat. Soc. Series B: Stat. Methodol.* 1979;41(2): 113–128.
- [26] Firth D. Bias reduction of maximum likelihood estimates. *Biometrika.* 1993;80(1): 27–38.
- [27] Magis D. A note on weighted likelihood and Jeffreys modal estimation of proficiency levels in polytomous item response models. *Psychometrika.* 2015;80: 200–204.
- [28] Zhou X, Giacometti R, Fabozzi FJ, et al. Bayesian estimation of truncated data with applications to operational risk measurement. *Quant. Finance.* 2014;14(5): 863–888.
- [29] Sartori N. Bias prevention of maximum likelihood estimates for scalar skew normal and skew t distributions. *J. Stat. Plan. Inference.* 2006;136(12): 4259–4275.
- [30] Bodnar O, Link A, Elster C. Objective Bayesian inference for a generalized marginal random effects model. *Bayesian Anal.* 2016;11(1): 25–45.
- [31] Kosmidis I, Guolo A, Varin C. Improving the accuracy of likelihood-based inference in meta-analysis and meta-regression. *Biometrika.* 2017;104(2): 489–496.
- [32] Schünemann HJ, Higgins J, Vist GE, et al. Chapter 14: completing ‘summary of findings’ tables and grading the certainty of the evidence. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). *Cochrane Handbook for Systematic Reviews of Interventions*. 2nd Edition. Chichester: John Wiley & Sons, 2019.
- [33] Alam TF, Rahman MS, Bari W. On estimation for accelerated failure time models with small or rare event survival data. *BMC Med. Res. Methodol.* 2022;22(1): 1–15.
- [34] Mathur MB. P-hacking in meta-analyses: A formalization and new meta-analytic methods. *Res. Synth. Methods*. In press. Early view version available at <https://onlinelibrary.wiley.com/doi/full/10.1002/jrsm.1701>.
- [35] Cope EW. Penalized likelihood estimators for truncated data. *J. Stat. Plan. Inference.* 2011;141(1): 345–358.
- [36] Veroniki AA, Jackson D, Viechtbauer W, Bender R, Bowden J, Knapp G, Salanti G. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Res. Synth. Methods.* 2016;7(1): 55–79.
- [37] Viechtbauer W. Bias and efficiency of meta-analytic variance estimators in the random-effects model. *J. Educ. Behav. Stat.* 2005;30(3): 261–293.
- [38] Huizenga HM, Visser I, Dolan CV. Testing overall and moderator effects in random effects meta-regression. *Br. J. Math. Stat. Psychol.* 2011;64(1): 1–19.
- [39] Guolo A. Higher-order likelihood inference in meta-analysis and meta-regression. *Stat. Med.* 2012;31(4): 313–327.
- [40] Guolo A, Varin C. Random-effects meta-analysis: the number of studies matters. *Stat. Methods Med. Res.* 2017;26(3): 1500–1518.
- [41] Noma H. Confidence intervals for a random-effects meta-analysis based on Bartlett-type corrections. *Stat. Med.* 2011;30(28): 3304–3312.
- [42] Hilde H, Ingmar V. metatest: Fit and test metaregression models. R package version 1.0-5. 2018. <https://CRAN.R-project.org/package=metatest>.
- [43] Veroniki AA, Jackson D, Bender R, et al. Methods to calculate uncertainty in the estimated overall effect size from a random-effects meta-analysis. *Res. Synth. Methods.* 2019;10(1): 23–43.
- [44] Van Den Noortgate W, Onghena P. Parametric and nonparametric bootstrap methods for meta-analysis. *Behav. Res. Methods.* 2005;37: 11–22.
- [45] Efron B. Better bootstrap confidence intervals. *J. Amer. Stat. Assoc.* 1987;82(397): 171–185.
- [46] Carpenter J, Bithell J. Bootstrap confidence intervals: When, which, what? A practical guide for medical statisticians. *Stat. Med.* 2000;19(9): 1141–1164.
- [47] Mathur MB, VanderWeele TJ. Robust metrics and sensitivity analyses for meta-analyses of heterogeneous effects. *Epidemiology.* 2020;31(3): 356–358.
- [48] Cornell JE, Mulrow CD, Localio R, et al. Random-effects meta-analysis of inconsistent effects: a time for change. *Ann. Internal Med.* 2014;160(4): 267–270.



- [49] Berger JO, Bernardo JM. Ordered group reference priors with application to the multinomial problem. *Biometrika*. 1992;79(1): 25–37.
- [50] Bodnar O, Bodnar T. Objective Bayesian meta-analysis based on generalized marginal multivariate random effects model. *Bayesian Anal.* 2024;19(2): 531–564.
- [51] Röver C, Friede T. Discrete approximation of a mixture distribution via restricted divergence. *J. Comput. Graph. Stat.* 2017;26(1): 217–222.
- [52] Greenland S. Generalized conjugate priors for Bayesian analysis of risk and survival regressions. *Biometrics* 2003;59(1): 92–99.
- [53] Jeffreys H. *Theory of Probability*. 3rd ed. Oxford University Press; 1961.
- [54] Riley RD, Higgins JPT, Deeks JJ. Interpretation of random effects meta-analyses. *BMJ*. 2011;342.
- [55] Mathur MB, VanderWeele TJ. New metrics for meta-analyses of heterogeneous effects. *Stat. Med.* 2019;38(8): 1336–1342.
- [56] Hedges LV. Distribution theory for Glass's estimator of effect size and related estimators. *J. Educ. Stat.* 1981;6(2): 107–128.
- [57] Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. *Stat. Med.* 2001;20(6): 825–840.
- [58] Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J. Stat. Softw.* 2010;36(3): 1–48.
- [59] Cauty A, Boot RB. Bootstrap functions. R package version 1.3-30. 2024. <https://CRAN.R-project.org/package=boot>.
- [60] Hedges LV. Estimation of effect size from a series of independent experiments. *Psychol. Bull.* 1982;92(2): 490.
- [61] Nemes S, Jonasson JM, Genell A, et al. Bias in odds ratios by logistic regression modelling and sample size. *BMC Med. Res. Methodol.* 2009;9: 1–5.
- [62] Rothman KJ. Disengaging from statistical significance. *Eur. J. Epidemiol.* 2016;31: 443–444.
- [63] McShane BB, Gal D, Gelman A, et al. Abandon statistical significance. *Amer. Stat.* 2019;73(sup1): 235–245.
- [64] Mathur MB, VanderWeele TJ. Finding common ground in meta-analysis “wars” on violent video games. *Perspect. Psychol. Sci.* 2019;14(4): 705–708.
- [65] Tang J-L. Weighting bias in meta-analysis of binary outcomes. *J. Clin. Epidemiol.* 2000;53(11): 1130–1136.
- [66] Chang B-H, Hoaglin DC. Meta-analysis of odds ratios: current good practices. *Med. Care* 2017;55(4): 328.
- [67] Zito A, Galli M, Biondi-Zoccai G, Abbate A, Douglas PS, Princi G, Burzotta F. Diagnostic strategies for the assessment of suspected stable coronary artery disease: a systematic review and meta-analysis. *Ann. Internal Med.* 2023;176(6): 817–826.
- [68] Cook SR, Gelman A, Rubin DB. Validation of software for Bayesian models using posterior quantiles. *J. Comput. Graph. Stat.* 2006;15(3): 675–692.
- [69] Röver C, Friede T. Dynamically borrowing strength from another study through shrinkage estimation. *Stat. Methods Med. Res.* 2020;29(1): 293–308.
- [70] White IR, Turner RM, Karahalios A, et al. A comparison of arm-based and contrast-based models for network meta-analysis. *Stat. Med.* 2019;38(27): 5197–5213.
- [71] Dias S, Ades AE. Absolute or relative effects? Arm-based synthesis of trial data. *Res. Synth. Methods* 2016;7(1): 23.
- [72] Kulinskaya E, Hoaglin DC, Bakbergenuly I. Exploring consequences of simulation design for apparent performance of methods of meta-analysis. *Stat. Methods Med. Res.* 2021;30(7): 1667–1690.
- [73] Mathur MB, VanderWeele TJ. Estimating publication bias in meta-analyses of peer-reviewed studies: A meta-meta-analysis across disciplines and journal tiers. *Res. Synth. Methods.* 2021;12(2): 176–191.
- [74] Mathur MB, VanderWeele TJ. Sensitivity analysis for unmeasured confounding in meta-analyses. *J. Amer. Stat. Assoc.* 2020;115(529): 163–170.
- [75] Wang C-C, Lee W-C. A simple method to estimate prediction intervals and predictive distributions: summarizing meta-analyses beyond means and confidence intervals. *Res. Synth. Methods* 2019;10(2): 255–266.
- [76] Röver C, Rindskopf D, Friede T. How trace plots help interpret meta-analysis results. *Res. Synth. Methods.* 2024;15(3): 413–429.
- [77] Bellio R, Guolo A. Integrated likelihood inference in small sample meta-analysis for continuous outcomes. *Scand. J. Stat.* 2016;43(1): 191–201.
- [78] Malzahn U, Böhning D, Holling H. Nonparametric estimation of heterogeneity variance for the standardised difference used in meta-analysis. *Biometrika*. 2000;87(3): 619–632.
- [79] Böhning D, Malzahn U, Dietz K, et al. Some general points in estimating heterogeneity variance with the DerSimonian–Laird estimator. *Biostatistics.* 2002(3)4 pp. 445–457.
- [80] Johnson BT, Huedo-Medina TB. Meta-Analytic Statistical Inferences for Continuous Measure Outcomes as a Function of Effect Size Metric and Other Assumptions: AHRQ Methods for Effective Health Care. *Statist. Sci.* 1992;7(2): 246–255. <https://doi.org/10.1214/ss/1177011364>.
- [81] Hedges LV. Modeling publication selection effects in meta-analysis. *Stat. Sci.* 1992: 246–255.
- [82] Carter EC, Schönbrodt FD, Gervais WM, et al. Correcting for bias in psychology: A comparison of meta-analytic methods. *Adv. Methods Pract. Psychol. Sci.* 2019;2(2): 115–144. <https://doi.org/10.1177/2515245919847196>.
- [83] Terrin N, Schmid CH, Lau J, et al. Adjusting for publication bias in the presence of heterogeneity. *Stat. Med.* 2003;22(13): 2113–2126. <https://doi.org/10.1002/sim.1461>.