Interval estimation of the overall treatment effect in random-effects meta-analyses: Recommendations from a simulation study comparing frequentist, Bayesian, and bootstrap methods

Frank Weber | Guido Knapp | Änne Glass | Günther Kundt | Katja Ickstadt

1Institute for Biostatistics and Informatics in Medicine and Ageing Research, Rostock University Medical Center, Rostock, Germany
2Department of Statistics, TU Dortmund University, Dortmund, Germany

Correspondence
Frank Weber, Institute for Biostatistics and Informatics in Medicine and Ageing Research, Rostock University Medical Center, Rostock, Germany.
Email: frank.weber@uni-rostock.de

Abstract
There exists a variety of interval estimators for the overall treatment effect in a random-effects meta-analysis. A recent literature review summarizing existing methods suggested that in most situations, the Hartung-Knapp/Sidik-Jonkman (HKSJ) method was preferable. However, a quantitative comparison of those methods in a common simulation study is still lacking. Thus, we conduct such a simulation study for continuous and binary outcomes, focusing on the medical field for application. Based on the literature review and some new theoretical considerations, a practicable number of interval estimators is selected for this comparison: the classical normal-approximation interval using the DerSimonian-Laird heterogeneity estimator, the HKSJ interval using either the Paule-Mandel or the Sidik-Jonkman heterogeneity estimator, the Skovgaard higher-order profile likelihood interval, a parametric bootstrap interval, and a Bayesian interval using different priors. We evaluate the performance measures (coverage and interval length) at specific points in the parameter space, that is, not averaging over a prior distribution. In this sense, our study is conducted from a frequentist point of view. We confirm the main finding of the literature review, the general recommendation of the HKSJ method (here with the Sidik-Jonkman heterogeneity estimator). For meta-analyses including only two studies, the high length of the HKSJ interval limits its practical usage. In this case, the Bayesian interval using a weakly informative prior for the heterogeneity may help. Our recommendations are illustrated using a real-world meta-analysis dealing with the efficacy of an intramyocardial bone marrow stem cell transplantation during coronary artery bypass grafting.

Abbreviations: BCa, bias-corrected and accelerated; biN, bimodal normal mixture distribution; BNHM, binomial-normal hierarchical model; BT, Biggerstaff-Tweedie; CDSR, Cochrane Database of Systematic Reviews; CI, confidence interval; CrI, credible interval; DIRECT, divergence restricting conditional tessellation; DL, DerSimonian-Laird; EB, empirical Bayes; FE, fixed-effect; FP, Follmann-Proschan; HC, Henmi-Copas; hc, half-Cauchy; HKSJ, Hartung-Knapp/Sidik-Jonkman; MCMC, Markov chain Monte Carlo; MCSE, Monte Carlo standard error; mHKSJ, modified Hartung-Knapp/Sidik-Jonkman; ML, maximum likelihood; N, normal distribution; NNHM, normal-normal hierarchical model; npBoot, nonparametric bootstrap; OR, odds ratio; pBoot, parametric bootstrap; PM, Paule-Mandel; PL, profile likelihood; RE, random-effects; REML, restricted maximum likelihood; Rhodes15, informative prior for \( \tau \) in case of the SMD; SE, standard error; SJ, Sidik-Jonkman; SMD, standardized mean difference; sN, skew-normal distribution; Turner12, informative prior for \( \tau \) in case of the log OR; UMVUE, uniformly minimum-variance unbiased estimator; WTq, Wald-type with quantile approximation; WTt, Wald-type \( t \); WTz, Wald-type \( z \); ZL, Zeng-Lin.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2020 The Authors. Research Synthesis Methods published by John Wiley & Sons Ltd.
1 | INTRODUCTION

Over the years, a variety of interval estimators has been developed for the overall treatment effect in random-effects (RE) meta-analyses, including frequentist, Bayesian, and bootstrap methods (see Veroniki et al. and the references therein). Correspondingly, there is also a variety of simulation studies comparing these methods. In their literature review, Veroniki et al. summarize those simulation studies (as well as real-life data studies). They “tentatively” recommend the Hartung-Knapp/Sidik-Jonkman (HKSJ) method for most situations and list situations where this choice might not be optimal (see section 2.3.2). However, they note that a comparison of all methods in a common simulation study is necessary for making reliable recommendations, but still lacking.

Thus, in this article, we want to compare a broad range of interval estimators under a broad range of scenarios. As it is computationally demanding to compare all methods under all scenarios, we confine ourselves to a sensible selection of methods and scenarios. Our choice is mainly guided by the results from Veroniki et al., but we also present some new theoretical considerations relevant for the choice of methods.

We note that our work is motivated by meta-analyses in the medical field. The results generalize beyond this field, but for the choice of the simulation parameter values and the priors used in the Bayesian interval, this does have a minor impact.

In section 2, we introduce the RE model and the interval estimators considered by Veroniki et al. In that section, we also describe the setup of the simulation study. In section 3, we present the results of the simulation study and summarize them through general recommendations. The application of those recommendations is illustrated in section 4 using a real-world meta-analysis. Our work is discussed in section 5.

For our computations, we used R. The R code used for this article, its output, as well as information about the computing environment is available online.

2 | METHODS

Let $K \geq 2$ be the number of studies included in the meta-analysis, indexed by $k \in \{1, ..., K\}$. The studies are assumed to be independent. The sizes of the treatment and the control group of study $k$ will be denoted by $N_{T, k}$ and $N_{C, k}$, respectively. We assume that $N_{T, k}$ and $N_{C, k}$ satisfy $N_k := N_{T, k} + N_{C, k} > 4$ (necessary e.g., for section 2.4.1).

Let $\theta_k \in \mathbb{R}$ be the true treatment effect in study $k$ whose estimator is $\hat{\theta}_k$. The standard error (SE) of $\hat{\theta}_k$ (given $\theta_k$) will be denoted by $\nu_k \in (0, \infty)$ whose estimator is $\hat{\nu}_k$. All methods investigated here were derived under the assumption that the $\nu_k$ are known. Currently, this assumption is standard practice for meta-analytic methods even though it might not be optimal. Due to this assumption, we describe all methods in terms of the true $\nu_k$ instead of their estimators $\hat{\nu}_k$. The feasible counterparts of the methods are obtained by substituting $\nu_k$ by $\hat{\nu}_k$. In contrast to the theoretical...
formulation of the methods, we use these feasible counterparts when applying the methods in our simulation study (see section 2.5). This is done to have a simulation study which is as realistic as possible. In section 2.4, we define \( \theta_k \), \( \theta \), and \( \nu_k \) for the standardized mean difference (SMD) as well as for the log odds ratio (log OR).

## 2.1 Random-effects model

The most popular RE model is a normal-normal hierarchical model (NNHM).\(^4,5\) Other RE models are also possible: For example in the case of binary data, a binomial-normal hierarchical model (BNHM) might be a better option (see section 5). However, our work will be restricted to the NNM, so that in the following, the term “RE model” will refer to the NNM.

Let \( N(\cdot, \cdot) \) denote the normal distribution. The RE model then consists of the following assumptions\(^4,5\):

- \( (\theta_k | \theta, r^2) \sim N(\theta, r^2) \) with \( \theta \in \mathbb{R} \) denoting the overall treatment effect and \( r^2 \in [0, \infty) \) denoting the between-study variance (heterogeneity),
- \( (\theta_k | \theta_k) \sim N(\theta_k, \nu_k^2) \).

Marginally, that is, by integrating out the study-specific effect \( \theta_k \), the RE model yields\(^4,5\)

\[
(\hat{\theta} | \theta, r^2) \sim N(\theta, \nu_k^2 + r^2).
\]

In the degenerate case \( r^2 = 0 \), the RE model simplifies to the fixed-effect (FE) model (also called the common-effect model\(^6\)), which should not be confused with the fixed-effects model\(^7\) where the \( \theta_k \) are not pooled at all, at least not by the model itself.\(^6\) The FE model is quite restrictive and therefore not recommended in general.\(^8,9\) The fixed-effects (as well as the FE) model only allows for conditional inference, meaning that its results cannot be generalized beyond the studies at hand (in contrast to the RE model which allows for unconditional inference).\(^6,10-12\) Thus, we only consider methods for the RE model.

If \( r^2 \) was known, \( \theta \) could be estimated using the uniformly minimum-variance unbiased estimator (UMVUE)\(^9\)

\[
\hat{\theta}(r^2) := \frac{1}{w(r^2)} \sum_{k=1}^{K} w_k(r^2) \hat{\theta}_k
\]

with the unnormalized precision weights \( w_k(r^2) := 1/(r^2 + \nu_k^2) \) and their sum \( w(r^2) := \sum_{k=1}^{K} w_k(r^2) \). In practice, \( r^2 \) needs to be estimated. In general, plugging in an estimator and not taking this into account leads to invalid inference.\(^8,9\) We will come back to this problem in section 2.3.1.

## 2.2 Estimators for the between-study variance \( r^2 \)

Several interval estimators for \( \theta \) require the specification of an estimator for \( r^2 \). Reviews of estimators for \( r^2 \) are given, for example, by Veroniki et al.\(^13\) and Langan et al.\(^14\) Here, we only present those estimators used in our simulation study.

### 2.2.1 DerSimonian-Laird

The DerSimonian-Laird (DL) estimator\(^15\) is the default \( r^2 \) estimator in multiple software packages and—presumably therefore—probably the most common \( r^2 \) estimator.\(^13\) For its definition, several FE quantities are needed: the weights \( u_k := 1/\nu_k^2 \), their sum \( u := \sum_{k=1}^{K} u_k \), the FE estimator \( \hat{\theta}_{FE} := \frac{1}{u} \sum_{k=1}^{K} u_k \hat{\theta}_k \), and the Q statistic by Cochran,\(^16\) given by

\[
Q := \sum_{k=1}^{K} u_k (\hat{\theta}_k - \hat{\theta}_{FE})^2.
\]

By the method of moments (i.e., equating \( Q \) to its expected value under the RE model, given \( \theta \) and \( r^2 \)) and cutting off at zero, the DL estimator \( \hat{r}^2_{DL} \) results as

\[
\hat{r}^2_{DL} := \max \left\{ 0, \frac{Q - (K-1)}{u - \frac{1}{u} \sum_{k=1}^{K} u_k^2} \right\}.
\]

### 2.2.2 Paule-Mandel

The Paule-Mandel (PM) estimator\(^17\) is recommended by Veroniki et al.\(^13\) and Langan et al.\(^14\) It is equivalent to the empirical Bayes (EB) estimator by Morris\(^18\) and Berkey et al.\(^19\) (see Veroniki et al.\(^13\)). For continuous data, Veroniki et al.\(^13\) also recommend the restricted maximum likelihood (REML) estimator, but the PM—or rather the EB—estimator and the REML estimator are closely related.\(^5,19,20\)

The PM estimator is a method of moments estimator based on the generalized \( Q \) statistic

\[
Q_{gen}(r^2) := \sum_{k=1}^{K} w_k(r^2) (\hat{\theta}_k - \hat{\theta}(r^2))^2
\]

which—given \( \theta \) and \( r^2 \)—follows a \( \chi^2 \) distribution with \( K - 1 \) degrees of freedom.\(^21\) Thus, the PM estimator \( \hat{r}^2_{PM} \) is defined as the solution to the equation
\[ Q_{gen}(\hat{r}^2_{PM}) = K - 1. \] (6)

As noted by DerSimonian and Kacker,\(^{22}\) \(Q_{gen}(r^2)\) is a strictly monotonically decreasing function of \(r^2\). Thus, if a solution to (6) exists, it is unique.\(^{22}\) However, the case \(Q_{gen}(0) < K - 1\) may also occur.\(^{22}\) In this case, there is no solution to (6) on \([0, \infty)\) and the PM estimator is set to zero.\(^{22}\)

For our simulation study, we use the \(R^2\) package metfor\(^{23}\) to calculate the PM estimate. In fact, we replaced squares and assuming a known scedastic errors. Applying the method of weighted least squares and introducing the included methods.

### 2.2.3 Sidik-Jonkman

The Sidik-Jonkman (SJ) estimator\(^{24}\) is derived by writing the marginal RE model as a linear model with heteroscedastic errors. Applying the method of weighted least squares and assuming a known \(r^2\) leads to

\[ \hat{r}^2_{SJ}(r^2) := \frac{\hat{r}^2}{K-1} Q_{gen}(r^2). \] (7)

By replacing \(r^2\) with the crude estimator

**TABLE 1** Interval estimators for the overall treatment effect \(\theta\) included in the literature review by Veroniki et al\(^{1}\): 5 are included in our simulation study, 10 are excluded

| Abbreviation | Full name                                | Included/Excluded |
|--------------|-----------------------------------------|-------------------|
| WTz          | Wald-type \(z\)                         | Included          |
| WTt          | Wald-type \(t\)                         | Excluded          |
| WTqa         | Wald-type with quantile approximation   | Excluded          |
| HKSJ         | Hartung-Knapp/Sidik-Jonkman             | Included          |
| mHKSJ        | Modified Hartung-Knapp/Sidik-Jonkman    | Excluded          |
| PL           | Profile likelihood                      | Excluded          |
| Bartlett     | Profile likelihood with higher-order Bartlett correction | Excluded          |
| Skovgaard    | Profile likelihood with higher-order Skovgaard correction | Included          |
| HC           | Henmi-Copas                             | Excluded          |
| BT           | Biggerstaff-Tweedie                     | Excluded          |
| ZL           | Zeng-Lin                                | Excluded          |
| npBoot       | Nonparametric bootstrap                 | Excluded          |
| pBoot        | Parametric bootstrap                    | Included          |
| FP           | Follmann-Proschan                      | Excluded          |
| Bayesian CrI | Bayesian credible interval              | Included          |

\[ \hat{r}^2_{SJ} = \frac{1}{K} \sum_{k=1}^{K} \left( \hat{\theta}_k - \frac{1}{K} \sum_{j=1}^{K} \theta_j \right)^2, \] (8)

Sidik and Jonkman\(^{24}\) define the SJ estimator through

\[ \hat{r}^2_{SJ} = \frac{\hat{r}^2_{0}}{K-1} Q_{gen}(\hat{r}^2_{0}). \] (9)

Note that \(\hat{r}^2_{SJ}\) is almost surely positive since \(\hat{\theta}_1 = \ldots = \hat{\theta}_K\) with probability zero.

#### 2.3 Interval estimators for the overall treatment effect \(\theta\)

In their literature review, Veroniki et al\(^{1}\) consider 15 different interval estimators for \(\theta\). These are listed in Table 1 together with their abbreviations used here. Since a comparison of all 15 methods is impractical—especially when covering a wide range of scenarios—we use the information given by Veroniki et al\(^{1}\) to select only those methods which promise a favorable behavior (with the exception of the widespread Wald-type \(z\) confidence interval, see the following section). In Table 1, methods that are excluded from our simulation study are colored in gray. We explain why these methods are excluded after introducing the included methods.

The nominal coverage level will be denoted by \((1 - \alpha)\) with \(\alpha \in (0, 1)\). For the simulation, we set \(\alpha = 0.05\).

#### 2.3.1 Wald-type \(z\)

Using the fact \((\hat{\theta}(\hat{r}^2)|\theta, r^2) \sim N(\theta, 1/w(r^2))\) and by plugging in an estimator \(\hat{r}^2\) for \(r^2\), the Wald-type \(z\) (WTz) confidence interval (CI) is defined through\(^{21}\)

\[ \left[ \hat{\theta}(\hat{r}^2) \pm z_{1-\frac{\alpha}{2}} \frac{1}{\sqrt{w(\hat{r}^2)}} \right] \] (10)

with \(\hat{r}^2\) denoting the chosen \(r^2\) estimator and \(z_{1-\frac{\alpha}{2}}\) the \((1 - \frac{\alpha}{2})\) quantile of the standard normal distribution. It is well-known that the WTz CI generally has a coverage considerably below the nominal level, especially for few \(r^2\) values.\(^{1}\) This is caused by ignoring the additional uncertainty due to the estimation of \(r^2\) and by keeping the assumption of a normal distribution even though \(r^2\) is estimated.\(^{8}\) Nevertheless, we include the WTz CI in our simulation study to show its inferiority compared to the other methods.
For this, we use the DL estimator for \( r^2 \), resulting in the typical “WTz-DL” combination, that is, used as the default interval estimator for \( \theta \) in many statistical software packages.\(^1\)

### 2.3.2 Hartung-Knapp/Sidak-Jonkman

As mentioned in section 2.2.2, \( Q_{\text{gen}}(r^2) \) follows a \( \chi^2 \) distribution with \( K - 1 \) degrees of freedom (given \( \theta \) and \( r^2 \)). Thus, if \( r^2 \) was known, then the variance \( 1/w(r^2) \) of \( \theta(r^2) \) could be estimated unbiasedly by \( q(r^2)/w(r^2) \) with

\[
q(r^2) = \frac{Q_{\text{gen}}(r^2)}{K-1}.
\]

Furthermore, since \( \hat{\theta}(r^2) \) and \( Q_{\text{gen}}(r^2) \) are stochastically independent (given \( \theta \) and \( r^2 \)),\(^25\) we have

\[
\left( \frac{\hat{\theta}(r^2) - \theta}{\sqrt{q(r^2)/w(r^2)}} \right) \sim t_{K-1} \tag{12}
\]

with \( t_{K-1} \) denoting the distribution with \( K - 1 \) degrees of freedom. Thus, the HKSJ CI\(^25-27\) is defined as

\[
\hat{\theta}(r^2) \pm t_{K-1;1-\frac{\alpha}{2}} \sqrt{V_{\text{HKSJ}}} \tag{13}
\]

with \( t_{K-1;1-\frac{\alpha}{2}} \) denoting the (1 - \( \frac{\alpha}{2} \)) quantile of the \( t_{K-1} \) distribution and \( V_{\text{HKSJ}} = q(r^2)/w(r^2) \) denoting the squared “HKSJ SE estimator”.

The HKSJ CI shows favorable properties in multiple simulation studies:\(^1\) It usually maintains the nominal coverage level, even for small \( K \) and independently of the magnitude of \( r^2 \). Additionally, the coverage of the HKSJ CI is only slightly influenced by the chosen \( r^2 \) estimator.\(^1\) However, according to Veroniki et al,\(^1\) caution is required with the HKSJ CI if few (less than 5) or small studies are included in the meta-analysis, if there is no (or little) between-study heterogeneity, if analyzing binary data with rare events, or if the studies are of strongly varying precision (e.g., due to different study sizes).

In our simulation study, we will compare the PM and the SJ estimator for \( r^2 \) when calculating the HKSJ CI. The PM estimator is chosen since it is generally recommended by Veroniki et al\(^1\) and Langan et al.\(^14\) The SJ estimator is chosen since it shows good properties in the simulation study by Sánchez-Meca and Marín-Martínez.\(^28\) Depending on the results, one of those two \( r^2 \) estimators will be chosen for the final comparison with the other methods.

#### 2.3.3 Skovgaard

The Skovgaard CI\(^29,30\) arises from the ordinary profile likelihood (PL) CI\(^31\) through a higher-order correction for the likelihood ratio statistic. Let

\[
L(\theta, r^2) := \prod_{k=1}^{K} \frac{1}{\sqrt{2\pi (r^2 + \nu_k^2)}} \exp \left( -\frac{1}{2} \left( \frac{\hat{\theta}_k - \theta}{r^2 + \nu_k^2} \right)^2 \right) \tag{14}
\]

denote the likelihood of the marginal RE model and

\[
L_{pr}(\theta_0) := L(\theta_0, \hat{r}_{\text{ML}}^2(\theta_0)) \tag{15}
\]

the profile likelihood (profiled over the nuisance parameter \( r^2 \)) where \( \hat{r}_{\text{ML}}^2(\theta_0) \) denotes the maximum likelihood (ML) estimator for \( r^2 \) given a specific \( \theta_0 \in \mathbb{R} \), that is, \( \hat{r}_{\text{ML}}^2(\theta_0) \) is the solution to\(^31\)

\[
\hat{r}_{\text{ML}}^2(\theta_0) = \max \left\{ 0, \frac{\sum_{k=1}^{K} w_k^2 (\hat{r}_{\text{ML}}^2(\theta_0)) \left( (\hat{\theta}_k - \theta_0)^2 - \nu_k^2 \right)}{\sum_{k=1}^{K} w_k^2 (\hat{r}_{\text{ML}}^2(\theta_0))} \right\}. \tag{16}
\]

The ML estimator for \( \theta \), here denoted by \( \hat{\theta}_{\text{ML}} \), is then given by the solution to

\[
\hat{\theta}_{\text{ML}} = \hat{\theta}(\hat{r}_{\text{ML}}^2) \tag{17}
\]

with \( \hat{r}_{\text{ML}}^2 := \hat{r}_{\text{ML}}^2(\hat{\theta}_{\text{ML}}) \) denoting the ML estimator for \( r^2 \). Under \( H_0 : \theta = \theta_0 \), the signed square root of the likelihood ratio statistic, that is

\[
Z(\theta_0) := \text{sign}(\hat{\theta}_{\text{ML}} - \theta_0) \sqrt{-2 \log \frac{L_{pr}(\theta_0)}{L_{pr}(\hat{\theta}_{\text{ML}})}} \tag{18}
\]

(with \( \text{sign}(\cdot) \) denoting the sign function), is asymptotically \( \mathcal{N}(0, 1) \)-distributed. Instead of using \( Z(\theta_0) \) as done in the ordinary PL CI, the Skovgaard CI uses the corrected statistic

\[
\tilde{Z}(\theta_0) := Z(\theta_0) + \frac{1}{Z(\theta_0)} \log \frac{U(\theta_0)}{Z(\theta_0)} \tag{19}
\]
with \( U(\theta_0) \) given \( \hat{r}_{ML}^2(\theta_0) \) and \( \hat{r}_{ML}^2 \), having a closed-form, yet complex expression that may be found in Guolo.\(^{30}\) Note that \( \tilde{Z}(\theta_0) \) is only defined for \( \theta_0 \neq \theta_{ML} \). To our knowledge, the behavior of \( \tilde{Z}(\theta_0) \) for \( \theta_0 \to \theta_{ML} \) has not been investigated yet. Because of \( \tilde{Z}(\theta_{ML}) = 0 \), we assume that \( \tilde{Z}(\theta_0) \) is supposed to converge to 0 for \( \theta_0 \to \theta_{ML} \) and we therefore define

\[
\tilde{Z}(\theta_{ML}) = 0. \tag{20}
\]

The Skovgaard CI is then given by the set

\[
\{ \theta_0 \in \mathbb{R} : \tilde{Z}(\theta_0) \in \left( z_{2, \alpha/2} \right) \}. \tag{21}
\]

In \( \mathbb{R}^2 \), we calculate the Skovgaard CI using the function profile.metaLik() from the package metaLik,\(^{32}\) version 0.41.0. There are newer versions of this package available (at the time of writing: up to version 0.43.0) where the calculation of the Skovgaard CI is still possible using internal results (see e.g., Appendix A.8 of Veroniki et al\(^{35}\)), but version 0.41.0 is used here because it performs a “monotonicity check” when using the function profile. metaLik(). In this monotonicity check, several supporting points are used to check that \( \tilde{Z}(\cdot) \) is strictly monotonically decreasing. For example, a pole of \( \tilde{Z}(\cdot) \) at \( \theta_{ML} \) could prevent the set in (21) from forming an interval. If the monotonicity test is not passed, the ordinary PL CI is calculated. As this happened quite frequently during our simulations, we report on this issue in Section S1. In order to monitor this and other issues encountered during the simulation and to resolve some bugs, we slightly modified the source code of some functions from the package metaLik,\(^{32}\) as may be seen in the online repository for the software code\(^3\) (with the modified parts highlighted there).

### 2.3.4 Bootstrap

#### Generation of the bootstrap samples

Van Den Noortgate and Onghena\(^ {33} \) discuss four bootstrap types for the RE model: two nonparametric bootstraps (cases bootstrap, error bootstrap) and two parametric bootstraps (effect size bootstrap, raw data bootstrap). They conclude that the nonparametric cases bootstrap should be avoided.\(^ {33} \) The other three bootstrap types lead to similar results.\(^ {33} \) We therefore confine ourselves to one of those three types. As the error bootstrap and the raw data bootstrap require additional assumptions for effect measures like the SMD and the log OR, we choose the effect size bootstrap. It is performed by generating the \( R^* \in \mathbb{N} \) “bootstrap-world” equivalents to \( \hat{\theta}_k \) according to the

\[
N(\hat{\theta}(\hat{z}^2), \hat{z}^2 + \nu_k^2) \]

distribution and by using \( \nu_k \) as the bootstrap-world equivalent to \( \nu_k \) (so no new values are generated for \( \nu_k \)).

#### Bootstrap confidence intervals

The R\(^2 \) package boot,\(^ {34, 35} \) which is used here for performing the bootstrap and for calculating the corresponding CI for \( \theta \), offers five options for the CI type: the normal approximation CI,\(^ {35} \) the basic bootstrap CI\(^ {36} \) (also known as nonstudentized pivotal bootstrap CI\(^ {37} \)), the studentized bootstrap CI\(^ {36} \) (also known as studentized pivotal bootstrap CI or bootstrap-t CI\(^ {37} \)), the bootstrap percentile CI,\(^ {38} \) and the adjusted bootstrap percentile CI\(^ {39} \) (also known as bias-corrected and accelerated [BCa] bootstrap CI\(^ {37} \)).

In order to choose a CI type for our simulation study, we first consider the coverage accuracy. A first-order accuracy means that the asymptotic behavior of the coverage error (the true coverage minus the nominal coverage) is \( O(1/\sqrt{K}) \), in contrast to the second-order accuracy where this behavior is \( O(1/K) \).\(^ {40} \) Among the above-mentioned CI types, only the studentized bootstrap CI and the BCa bootstrap CI have a second-order accuracy.\(^ {35} \) In contrast to the BCa bootstrap CI, the studentized bootstrap CI may be easily combined with a parametric bootstrap. Thus, for our simulation study, we choose the studentized bootstrap CI. It uses the pivot\(^ {35} \)

\[
\hat{\xi} = \xi(\hat{z}^2) = \frac{\hat{\theta}(\hat{z}^2) - \theta}{\hat{V}} \tag{22}
\]

where \( \hat{V} \) is a squared SE estimator for \( \hat{\theta}(\hat{z}^2) \). Here, we will use the squared HKSJ SE estimator for \( \hat{\theta}(\hat{z}^2) \) (i.e., \( \hat{V} = \hat{V}_{HKSJ} \)) instead of the squared WTz SE estimator \( 1/w(\hat{z}^2) \) since the HKSJ CI was shown to perform slightly better than the Wald-type \( t \) (WT) CI\(^ {41} \) in terms of coverage\(^ {28} \) (and they only differ in the SE estimator).

Let \( \hat{\xi}_{(1)} \leq \ldots \leq \hat{\xi}_{(R^*)} \) denote the ordered bootstrap-world equivalents to \( \xi \) (which implies that for \( \hat{\xi}_{(1)} \leq \ldots \leq \hat{\xi}_{(R^*)} \), the true \( \theta \) is replaced by its bootstrap-world equivalent \( \hat{\theta}(\hat{z}^2) \)) and assume the number \( R^* \) of bootstrap samples to be chosen such that

\[
r_{lo} = (R^* + 1) \frac{\alpha}{2} \quad \text{and} \quad r_{up} = (R^* + 1) \left( 1 - \frac{\alpha}{2} \right) \tag{23}
\]

are integer-valued. Then the \( \frac{\alpha}{2} \) and the \( 1 - \frac{\alpha}{2} \) quantiles of the distribution of the pivot are estimated by

\[
\hat{\xi}_{(r_{lo})} \quad \text{and} \quad \hat{\xi}_{(r_{up})} \tag{24}
\]
respective. Thus, the studentized bootstrap CI (henceforth the bootstrap CI) is defined here through
\[ \tilde{\theta}(\hat{\tau}^2) - \zeta_{(\alpha)}^* \sqrt{V_{\text{HKSJ}}} \tilde{\theta}(\hat{\tau}^2) - \zeta_{(\alpha)}^* \sqrt{V_{\text{HKSJ}}}. \] 
(25)

For the number of bootstrap samples, we choose the value \( R^* := 999 \) at which the Monte Carlo error should be low enough to allow for reliable comparisons.\(^{35,42}\)

For the \( \hat{\tau}^2 \) estimator, we will choose the superior one from the comparison of \( \hat{\tau}^2 \) estimators for the HKSJ CI.

### 2.3.5 Bayesian credible interval

In our article, the Bayesian credible interval (CrI) is only considered as a technical device in a frequentist meta-analysis. For the Bayesian CrI, the heterogeneity will be parameterized in terms of \( \tau \) instead of \( \hat{\tau}^2 \). Since our focus is not on the study-specific effects \( \theta_k \), we define the likelihood through \( p(x|\theta, \tau) = L(\theta, \tau^2) \) with \( L(\theta, \tau^2) \) denoting the likelihood of the marginal RE model from Equation (14) and \( x = (\theta_1, ..., \theta_K) \) denoting the data.\(^{43}\)

* A priori, \( \theta \) and \( \tau \) are assumed to be independent, so that their joint prior factorizes into their respective parts:

\[ p(\theta, \tau) = p(\theta) \cdot p(\tau). \] 
(26)

### Prior for \( \theta \)

For our simulation study, we will use the \( R^2 \) package bayesmeta.\(^{43}\)

This package offers two options for the prior distribution for \( \theta \): an improper uniform distribution on the real line (denoted here by \( U(-\infty, \infty) \)) and a normal distribution \( N(\mu_\theta, \sigma^2_\theta) \) whose mean \( \mu_\theta \) and variance \( \sigma^2_\theta \) may be specified by the user.\(^{43}\)

The \( N(\mu_\theta, \sigma^2_\theta) \) prior is conjugate conditionally on \( \tau \).\(^{43}\)

The \( U(-\infty, \infty) \) prior also results in a normal conditional posterior for \( \theta \).\(^{4,5,43}\)

The mean and the variance of this latter normal conditional posterior are the limiting cases of those resulting from using the normal prior when \( \sigma^2_\theta \) from the normal prior approaches infinity.\(^{43}\)

Additionally, they are identical to the frequentist UMVUE introduced in section 2.1 (where we also conditioned on \( \tau \)) and its variance, respectively.\(^{4,5,43}\)

However, as noted by other authors,\(^{4,5,43}\) this relationship between frequentist and Bayesian inference does not hold for unknown \( \tau \).

In our simulation study, we will use the \( U(-\infty, \infty) \) prior as done e.g. by Friede et al.\(^4\)

The uniform prior may be deemed a noninformative prior, but in general, we have to warn about the downsides of noninformative priors as they may be a lot more informative than intended.\(^{44}\)

For the informative prior, we choose the following priors which are both defined for \( \log \tau^2 \):

- for the SMD: nonstandardized \( t \) distribution with 5 degrees of freedom, location \(-3.44\), and scale 2.59 by Rhodes et al.\(^{47}\);
- for the log OR: \( N(-2.56, 1.74^2) \) distribution by Turner et al.\(^{48}\)

These priors are based on data from the Cochrane Database of Systematic Reviews (CDSR) and are thus tailored to an application in the medical field. For both priors, we choose the general (marginal) setting, that is, we don’t specify any characteristics such as the outcome type or the type of intervention.

For the weakly informative prior, we mainly follow the arguments of Gelman\(^{49}\) and Polson and Scott.\(^{50}\)

Gelman\(^{49}\) derives the family of half-\( t \) distributions as a family of weakly informative priors for \( \tau \) based on Bayesian arguments. This family includes the half-Cauchy distribution (degrees of freedom equal to 1) and—although not in a strict sense—the normal distribution (degrees of freedom approaching infinity).\(^{43}\)

Polson and Scott\(^{50}\) show that the half-Cauchy distribution is also favorable from a frequentist point of view: Its (frequentist) risk—using the quadratic loss function—is a good compromise between an improvement compared to the James-Stein estimator\(^{51}\) for smaller Euclidean norms of \( \theta = (\theta_1, ..., \theta_K)^T \) and a concession for larger Euclidean norms of \( \theta \).\(^{50}\)

Because of the arguments given by Polson and Scott\(^{50}\) and because of its heavy tail which—rarely—allows for large heterogeneities, we choose the half-Cauchy prior among the family of half-\( t \) distributions. This is also supported by the work of Williams et al.\(^{12}\)

Concerning the scale of the half-Cauchy prior, Röver\(^{43}\) states that 0.5 is a reasonable choice for the log OR and similar endpoints. For their half-normal priors, Friede et al.\(^3\) and Friede et al.\(^4\) use scales of 0.5 and 1.0. As explained by Friede et al.,\(^4\) the scale of 1.0 may be used for the case that “very large between-trial heterogeneity is deemed possible.”\(^4\)

However, as we choose a half-Cauchy prior which has a much...
heavier tail than the half-normal, the scale of 1.0 shouldn’t be necessary here: The 0.95 quantile of the half-Cauchy with scale 0.5 is approximately 6.35 which corresponds to 40.32 on the scale of $\tau_2$. Assuming $\theta = 0$, this quantile gives a 2$\tau_2$ environment of $[-12.7, 12.7]$ for the study-specific effects $\theta_k$ which is already huge for effect measures like the log OR (consider this 2$\tau_2$ environment transformed on OR scale which gives $[3.05/C_{10}, 3.28/C_{10^5}]$), but for the scale of 1.0, that quantile would even double on the scale of $\tau$ and quadruple on the scale of $\tau_2$. Thus, we only choose the scale of 0.5 for our half-Cauchy prior. As our second effect measure—the SMD—may be regarded as being similar to the log OR,47,52 this applies to both, the SMD and the log OR.

In the case of the “noninformative” prior for $\tau$, we choose the $U(0, \infty)$ distribution as it is the limiting case of the half-t distributions when the scale approaches infinity (and the degrees of freedom are fixed).49 It is important to note that the $U(0, \infty)$ prior only results in a proper posterior if $K > 2$.49 Thus, in the special case $K = 2$, we will use a $U(0, 10^3)$ prior instead of the $U(0, \infty)$ prior. This is a heuristic workaround and only intended for illustrative purposes.

All of the $\tau$ priors chosen for our simulation study are illustrated on the common scale of $\tau$ in Figure 1 using the following abbreviations:

- Rhodes15: informative prior by Rhodes et al47 in case of the SMD,
- Turner12: informative prior by Turner et al48 in case of the log OR,
- hC(0.5): half-Cauchy prior with scale 0.5,
- U(0, Inf): $U(0, \infty)$ prior.

### Marginal posterior for $\theta$

The two options for the prior for $\theta$ allow the package bayesmeta43 to use analytical and numerical procedures (instead of Markov chain Monte Carlo [MCMC] procedures) for approximating the marginal posterior distribution of $\theta$, that is, the distribution with density $p(\theta | x)$. The key idea is to write $p(\theta | x)$ as a mixture of Gaussian densities (mixed over the marginal posterior of $\tau$), that is

$$p(\theta | x) = \int_0^\infty p(\theta | \tau, x) p(\tau | x) d\tau,$$

and to discretize in $\tau$.43 This way, the integral in Equation (27) is approximated by a weighted sum of Gaussian densities with weights given by the marginal posterior probability masses of $\tau$ lying in the discretization intervals.43 The marginal posterior probability masses of $\tau$ are calculated using a combination of analytic simplifications and numerical integration.43 In order to control the approximation of $p(\theta | x)$, the divergence restricting conditional tessellation (DIRECT) algorithm53 is used. It restricts the symmetrized Kullback-Leibler divergence between $p(\theta | \tau, x)$ and its approximation. This also guarantees a restriction of the symmetrized Kullback-Leibler divergence between $p(\theta | x)$ and its approximation (the weighted sum).53

After approximating $p(\theta | x)$, a numerical integration yields the (approximate) cumulative distribution function which may be used for deriving quantiles through a numerical root-finding algorithm.43 The $\frac{1}{2}$ and $(1 - \frac{1}{2})$ quantiles are used here to give the central $(1 - \alpha)$ CrI for $\theta$. 

![FIGURE 1 Priors for $\tau$ chosen for our simulation study. The constant level of the $U(0, \infty)$ prior was (arbitrarily) set to 1](wileyonlinelibrary.com)
2.3.6 | Excluded methods

The WTt CI\(^{41}\) arises from the WTz CI when substituting the standard normal quantile \(z_{1-\frac{\alpha}{2}}\) by the \(t_{k-1,1-\frac{\alpha}{2}}\) quantile. The Wald-type CI with quantile approximation (WTqa CI)\(^{54}\) arises from the WTz CI when substituting the standard normal quantile by a special quantile which was derived by Brockwell and Gordon\(^{54}\) and depends on \(K\). The WTt CI and the WTqa CI are excluded from our simulation study as they were shown to perform worse than the HKSJ CI in terms of coverage.\(^1\)

The modified Hartung-Knapp/Sidik-Jonkman (mHKSJ) CI\(^{55}\) corresponds to the HKSJ CI, except that it uses \(\tilde{q}(\tau^2) = \max\{1, q(\tau^2)\}\) instead of \(q(\tau^2)\) in the definition of the HKSJ SE estimator. The mHKSJ CI paired with the PM estimator for \(\tau^2\) is identical to the WTt CI (by definition of the PM estimator which always gives \(\tilde{q}(\tau^2_{PM}) = 1\)).\(^{56}\) Thus, with the exclusion of the WTt CI, the mHKSJ CI is also excluded from our simulation study. A further justification for this exclusion is that the mHKSJ CI has indeed been recommended instead of the HKSJ CI in some very specific cases, but we could not observe strong deviations of the HKSJ CI’s coverage from the nominal level in those cases. This is something we will come back to later in the discussion (section 5).

The ordinary PL CI, which uses \(Z(\theta_0)\) instead of \(\tilde{Z}(\theta_0)\) in Equation (21), has been shown to perform worse than the Skovgaard and the Bartlett method, especially for \(\theta\) in Equation (21), has been shown to perform worse than the Skovgaard CI in terms of coverage.\(^1\)

The modified Hartung-Knapp/Sidik-Jonkman (mHKSJ) CI\(^{55}\) corresponds to the HKSJ CI, except that it uses \(\tilde{q}(\tau^2) = \max\{1, q(\tau^2)\}\) instead of \(q(\tau^2)\) in the definition of the HKSJ SE estimator. The mHKSJ CI paired with the PM estimator for \(\tau^2\) is identical to the WTt CI (by definition of the PM estimator which always gives \(\tilde{q}(\tau^2_{PM}) = 1\)).\(^{56}\) Thus, with the exclusion of the WTt CI, the mHKSJ CI is also excluded from our simulation study. A further justification for this exclusion is that the mHKSJ CI has indeed been recommended instead of the HKSJ CI in some very specific cases, but we could not observe strong deviations of the HKSJ CI’s coverage from the nominal level in those cases. This is something we will come back to later in the discussion (section 5).

The Henmi-Copas (HC) CI\(^{61}\) was developed to guard against publication bias by using the FE point estimator, paired with an appropriate SE estimator and quantile for building a CI. Since the investigation of publication bias is outside of the scope of this article, we exclude the HC CI.

The Biggerstaff-Tweedie (BT) CI\(^{62}\) replaces the weight estimators \(w_k(\hat{\tau}^2_{DL})\) in \(\hat{\theta}(\hat{\tau}^2_{DL})\) by their (approximated) expected values, estimates the SE of the resulting point estimator for \(\theta\), and thereby builds a CI using a standard normal quantile. The BT CI was shown to have a performance similar to the WTz CI,\(^{1,54,63,64}\) so we exclude it here. We note that Preuß and Ziegler\(^{64}\) proposed a modification of the BT CI which uses exact weights.

The Zeng-Lin (ZL) CI\(^{65}\) is based on a two-step resampling procedure which uses the DL estimator for \(\tau^2\) in its first step. The ZL CI is excluded here since it was shown to perform worse than the Skovgaard CI in terms of coverage.\(^{66}\)

For the nonparametric bootstraps, we already explained in section 2.3.4 why we exclude them.

The Follmann-Prochans (FP) CI\(^{41}\) is composed of those values for \(\theta\) for which a permutation test cannot reject the corresponding null hypothesis. It has led to promising results in existing simulation studies\(^1,59,60,63\) but it has a crucial downside\(^60\): At a significance level of \(\alpha = 0.05\), the permutation test can reject the null hypothesis only if \(K \geq 6\) so that the FP CI equals \((-\infty, \infty)\) if \(K < 6\). As will be seen in section 2.5, real-world meta-analyses usually consist of a small number of studies so that the FP CI is useless in most practically relevant cases.

2.4 | Study-specific effect measures

In the following, the individual observations (i.e., on patient level) will be denoted by \(X_{T, k, n}\) \((n \in \{1, ..., N_{T, k}\})\) and \(X_{C, k, n}\) \((n \in \{1, ..., N_{C, k}\})\) for the treatment and the control group of study \(k\), respectively. In addition to the between-study independence, we also assume that the individual observations are stochastically independent within each study.

2.4.1 | Standardized mean difference

For the SMD, we assume a normal distribution (and variance homogeneity) for the patient-level observations:

\[
X_{T, k, n} \sim \mathcal{N}(\mu_{T, k, \tau^2_k}, \sigma_k^2) \quad \text{and} \\
X_{C, k, n} \sim \mathcal{N}(\mu_{C, k, \tau^2_k}, \sigma_k^2)
\]

with \(\mu_{T, k} \in \mathbb{R}\) and \(\mu_{C, k} \in \mathbb{R}\) denoting the means in the treatment and the control group (respectively) and \(\sigma_k^2 \in (0, \infty)\) denoting the variance in both treatment groups. The study-specific treatment effect (the SMD of study \(k\)) is then defined through\(^{67}\)

\[
\theta_k = \frac{\mu_{T, k} - \mu_{C, k}}{\sigma_k}.
\]

Its unbiased estimator (Hedges’s \(g\)) is defined through\(^{68}\)

\[
\hat{\theta}_k = \frac{X_{T, k} - X_{C, k}}{\sigma_k}.
\]
with
\[
\begin{align*}
\sigma_k^2 &= \frac{1}{N_k-2} \left( (N_k-1) \hat{\text{OR}}^2 - (N_k-1) \hat{\sigma}_{C,k}^2 \right), \\
\hat{\text{OR}}_{T,k} &= \frac{1}{NT,k} \sum_{n=1}^{N_{T,k}} X_{T,k,n}, \quad \hat{\text{OR}}_{C,k} = \frac{1}{NC,k} \sum_{n=1}^{N_{C,k}} X_{C,k,n}, \\
\hat{\sigma}_{T,k}^2 &= \frac{1}{N_{T,k}-1} \sum_{n=1}^{N_{T,k}} (X_{T,k,n} - \hat{\text{OR}}_{T,k})^2, \\
\hat{\sigma}_{C,k}^2 &= \frac{1}{N_{C,k}-1} \sum_{n=1}^{N_{C,k}} (X_{C,k,n} - \hat{\text{OR}}_{C,k})^2,
\end{align*}
\]
where \(\Gamma(\cdot)\) denotes the gamma function. The variance of \(\hat{\theta}_k\) (given \(\theta_k\)) is given by
\[
\nu_k^2 = \sigma_k^2 + \left( \frac{N_k-2}{N_{T,k}N_{C,k}} + \left( \frac{N_k-2}{N_{C,k}} \right)^2 \right) \hat{\theta}_k^2
\]
which may be estimated unbiasedly by
\[
\hat{\nu}_k^2 = \frac{N_{T,k}N_{C,k}}{N_k} + \left( 1 - \frac{N_k-4}{c_k^2(N_k-2)} \right) \hat{\theta}_k^2.
\]

\[
\hat{\nu}_k^2 = \begin{cases} 
\frac{1}{E_{T,k} + \frac{1}{2}} + \frac{1}{N_{T,k}-E_{T,k} + \frac{1}{2}} + \frac{1}{E_{C,k} + \frac{1}{2}} + \frac{1}{N_{C,k}-E_{C,k} + \frac{1}{2}} & \text{if } E_{T,k} \in \{0, N_{T,k}\} \text{ and } E_{C,k} \in \{0, N_{C,k}\}, \\
\frac{1}{E_{T,k} + \frac{1}{2}} + \frac{1}{N_{T,k}-E_{T,k} + \frac{1}{2}} + \frac{1}{E_{C,k} + \frac{1}{2}} + \frac{1}{N_{C,k}-E_{C,k} + \frac{1}{2}} & \text{else.}
\end{cases}
\]

\[
\theta_k = \logit(p_{T,k}) - \logit(p_{C,k})
\]

using the logit function \(\logit(p) = \log \frac{p}{1-p}\) for \(p \in (0, 1)\). Let
\[
E_{T,k} = \sum_{n=1}^{N_{T,k}} X_{T,k,n} \quad \text{and} \quad E_{C,k} = \sum_{n=1}^{N_{C,k}} X_{C,k,n}
\]
denote the number of events in the treatment and control group of study \(k\), respectively. We then estimate the log OR of study \(k\) by
\[
\hat{\theta}_k = \begin{cases} 
\log \frac{E_{T,k} + E_{C,k} - E_{C,k} - E_{T,k} E_{C,k}}{N_{T,k} - E_{T,k} - E_{C,k} + \frac{1}{2}} & \text{if } E_{T,k} \in \{0, N_{T,k}\} \text{ and } E_{C,k} \in \{0, N_{C,k}\}, \\
\log \frac{E_{T,k} + E_{C,k} - E_{C,k} - E_{T,k} E_{C,k} + \frac{1}{2}}{N_{T,k} - E_{T,k} + \frac{1}{2}} & \text{else.}
\end{cases}
\]

This type of zero-cell correction is used, for example, by Pateras et al.\(^69\) It is also the default zero-cell correction in the R\(^2\) package metafor.\(^23\) Accordingly, the variance of \(\hat{\theta}_k\) (given \(\theta_k\)) is estimated by

\[
\frac{1}{E_{T,k} + \frac{1}{2}} + \frac{1}{N_{T,k}-E_{T,k} + \frac{1}{2}} + \frac{1}{E_{C,k} + \frac{1}{2}} + \frac{1}{N_{C,k}-E_{C,k} + \frac{1}{2}}
\]

2.4.2 Log odds ratio

For the log OR, we assume a Bernoulli distribution for the binary patient-level observations:
\[
X_{T,k,n} \sim \text{Bin}(1, p_{T,k}) \quad \text{and} \quad X_{C,k,n} \sim \text{Bin}(1, p_{C,k})
\]
with \(p_{T,k} \in (0, 1)\) and \(p_{C,k} \in (0, 1)\) denoting the event probabilities in the treatment and control group of study \(k\) (respectively) and Bin(N, p) denoting the binomial distribution with parameters \(N \in \mathbb{N}\) and \(p \in (0, 1)\). The log OR of study \(k\) is defined through

\[
\theta_k = \logit(p_{T,k}) - \logit(p_{C,k})
\]

2.5 Simulation setup

For each scenario of our simulation study, we perform 5000 replications. The parameters defining the scenarios are listed in Table 2. The choice of those parameter values is described in detail in sections 2.5.1–2.5.6. As a full factorial design would have increased the number of scenarios considerably, we split our simulation study up into four analyses (see Table 2). In the location shift analysis, we investigate the sensitivity of the interval estimators with respect to a shift of magnitude of \(\theta\) (see section 2.5.3). In the robustness analysis, we assert the sensitivity
with respect to deviations from the normal distribution of the $\theta_k$ (see section 2.5.5).

Given a fixed scenario and a fixed replication of this scenario, we generate the study-specific estimates $\hat{\theta}_k$ and $\hat{\tau}_k^2$ as follows:

1. Generate the $N_{T,k}$ according to the chosen distribution (see section 2.5.2) and set $N_{C,k} := N_{T,k}$.
2. Generate the true study-specific parameters:
   - For the SMD:
     a. Generate the $\theta_k$ according to the chosen distribution, usually $N(\theta, \tau^2)$ (see section 2.5.5).
     b. Calculate $\mu_{T,k} := \mu_{C,k} + \theta_k \cdot \tau_k$.
   - For the log OR (Pateras et al.\textsuperscript{69} call this procedure used by Hartung and Knapp\textsuperscript{25} pRandom):
     a. Calculate $\ell_C := \logit(p_C)$ with a value for $p_C$ from section 2.5.6.
     b. Generate $\ell_{T,k} \sim N(\ell_C + \theta, \tau^2/2)$ and $\ell_{C,k} \sim N(\ell_C, \tau^2/2)$ independently.
     c. Calculate $p_{T,k} := \logit^{-1}(\ell_{T,k})$ and $p_{C,k} := \logit^{-1}(\ell_{C,k})$ with $\logit^{-1}$ denoting the inverse of the logit function.
3. Generate the patient-level observations:
   - For the SMD: Generate $X_{T,k,n} \sim N(\mu_{T,k}, \sigma_k^2)$ and $X_{C,k,n} \sim N(\mu_{C,k}, \sigma_k^2)$.
   - For the log OR: Generate directly $E_{T,k} \sim \text{Bin}(N_{T,k}, p_{T,k})$ and $E_{C,k} \sim \text{Bin}(N_{C,k}, p_{C,k})$.
4. Calculate the study-specific estimates $\hat{\theta}_k$ and $\hat{\tau}_k^2$ as explained in section 2.4.

Afterwards, all the 5 interval estimators (8 when taking into account the variants arising from different $\tau^2$ estimators or $\tau$ priors) from sections 2.3.1–2.3.5 are applied to this simulated meta-analysis. By averaging over the 5 000 replications, we estimate the primary and secondary performance measures, the coverage and the interval length, respectively. When plotting those performance measures, we also plot the Monte Carlo standard error (MCSE).

Note that in contrast to the frequentist and the bootstrap methods, the Bayesian CrI is not designed to maintain the nominal coverage when given a fixed point in parameter space (here, the two-dimensional space of $\theta$ and $\tau$). It is rather supposed to maintain the nominal coverage when averaging over the prior distribution.\textsuperscript{43} By estimating the coverage at fixed points in parameter space, we take a frequentist point of view. This is necessary, however, to see which interval estimators are affected by the magnitude of $\tau^2$. Williams et al.\textsuperscript{12} also argue why comparing methods according to their frequentist properties makes sense, even for Bayesian methods.

### 2.5.1 Number of studies ($K$)

For the number of studies, $K$, we choose the Fibonacci numbers 2, 3, 5, 8, 13, 21, 34, and 55. They reflect the right-skewed empirical distribution of $K$ from the CDSR quite well\textsuperscript{70} and cover the range of $K$ values used in most existing simulation studies.\textsuperscript{1,8,14,71}

### 2.5.2 Treatment group sizes ($N_{T,k}$)

For generating the treatment group sizes $N_{T,k}$, we use three settings: small, large, and mixed. The small group sizes are drawn from a uniform distribution over {10, 11, ..., 30} and the large group sizes from a uniform distribution over {100, 101, ..., 300}. Those two settings are also used by Paniyakul et al.\textsuperscript{72} For the mixed group sizes, we follow Doi et al.\textsuperscript{73} and Doi et al.\textsuperscript{74} in using a Delaporte distribution (with differing parameter values, however). The Delaporte distribution (denoted here by Delap($\lambda, \kappa, \xi$)) arises from a Poisson distribution whose mean is the sum of a fixed component ($\lambda \in (0, \infty)$) and a gammadistributed component (shape: $\kappa \in (0, \infty)$), scale:
Thus, it also arises as the convolution of a Poisson distribution and a negative binomial distribution. Here, we use a Delap(15,0.5,60) distribution and truncate it by setting values lower than 10 to 10. The parameters of the Delaporte distribution are chosen so that approximately 50% of the values lie within \{10, …, 30\}, 10% within \{100, …, 300\} and 40% in between. These are the proportions used by Trikalinos et al\textsuperscript{77} for their mixed study sizes. The probability mass function for the Delap(15,0.5,60) distribution is illustrated in Figure 2.

For the location shift analysis and the robustness analysis, we only use small group sizes. Mixed group sizes would have hindered a clean interpretation and large group sizes are less common in practice.\textsuperscript{70}

### 2.5.3 Overall treatment effect (θ)

For the main analyses as well as for the robustness analysis, we use θ = 0.5 as this is a reasonable value for the SMD\textsuperscript{78} as well as for the log OR\textsuperscript{79} that is also used in many existing simulation studies.\textsuperscript{1,8,14,71} In the location shift analysis, we will compare the results for θ = 0.5 to those for θ ∈ \{0,0.8\} in selected scenarios (see Table 2).

### 2.5.4 Heterogeneity (τ²)

For τ², we will use the values 0, 0.05, 0.15, 0.5, and 1.5. On the scale of τ, these values correspond to 0, 0.22, 0.39,
0.71, and 1.22. They are plausible in real-world meta-analyses\textsuperscript{47,48} (see also Figure 1) and cover the range of \(\tau^2\) values used in most existing simulation studies.\textsuperscript{1,8,14,71} In the location shift analysis and in the robustness analysis, we will only use \(\tau^2 = 0.15\) (the median of the chosen values).
2.5.5 Study-specific treatment effects ($\theta_k$)

The $\theta_k$ will be generated according to the $\mathcal{N}(\theta, \tau^2)$ distribution (abbreviated in tables and figures by “N”), except for the robustness analysis where we will investigate a skewed-normal distribution (“sN”) and a bimodal 50:50 mixture of two normal distributions of equal variance (“biN”). To maintain comparability, we ensured that the mean and the variance of the skewed-normal distribution and the bimodal normal mixture are also equal to $\theta$ and $\tau^2$, respectively.

For drawing from the skewed-normal distribution, we use the function rsnorm() from the R2 package fGarch with argument xi (determining the skewness) set to $\frac{1}{2}$.

For the bimodal normal mixture, we need to specify the distance $\delta_{\text{mix}}$ between the means of the two normal distributions. In order to have a positive common variance and an actual bimodal distribution, this distance needs to fulfill $\delta_{\text{mix}} \in (\sqrt{2\tau}, \sqrt{4\tau})$. Thus, we set $\delta_{\text{mix}} = \sqrt{3\tau}$ (more or less heuristically). For the $\tau^2$ value used in the robustness analysis ($\tau^2 = 0.15$), we get $\delta_{\text{mix}} = \sqrt{3\cdot0.15} \approx 0.67$.

The densities of this skewed-normal distribution and this bimodal normal mixture are illustrated in Figure 3, together with the corresponding normal distribution.

2.5.6 Parameters specific to the effect measures

In case of the SMD, we choose $\mu_{C,k} = 0$ and $\sigma^2_k = 1$ as done in other simulation studies. In case of the log OR, we choose an “overall control group event probability” of $p_C \in [0.1, 0.5]$ as done, for example, by Hartung and Knapp. The value of $p_C$ is varied since the coverage generally depends on this parameter.

3 RESULTS

Issues that occurred during the simulation are listed in Section S1. The plots for the interval length as well as some other plots may be found in Section S4. Henceforth, all references to Supporting Information 1 may be recognized by the prefix “S”, for example, Section S1 or Figure S1.

The interval length as well as $K$ are always plotted on logarithmic scale. In situations where there are many scenario parameters to visualize, we use a nested loop plot (see e.g., Figure 5).

3.1 Standardized mean difference

3.1.1 Main analysis

We first performed the following pre-selections:

1. pre-selection of the $\tau^2$ estimator (PM or SJ) in the HKSJ CI,
2. pre-selection of the prior for $\tau$ (informative, weakly informative, or “noninformative”) in the Bayesian CrI.

The corresponding results are given in Section S2. The SJ estimator is selected for the HKSJ CI and the hC(0.5) prior for the Bayesian CrI.

We now focus on comparing all methods chosen a priori or in the respective pre-selection. Figure 4 and Figure S10 show the coverage and the interval length for these methods. Figure S9 corresponds to Figure 4, however zoomed into the relevant part of the coverage.
Interestingly, the HKSJ-SJ CI and the bootstrap CI building upon it (pBoot-HKSJ-SJ) have a very similar coverage and interval length. Since the formulas of these CIs only differ in their quantiles, this similarity suggests that the \( t_K - 1 \) distribution underlying the HKSJ CI is a good approximation to the true sampling distribution of \( \hat{\theta}^2 \).

This might be the reason why, in terms of coverage, the HKSJ-SJ and the pBoot-HKSJ-SJ CI perform best in most scenarios. They also yield the longest intervals, but in most scenarios, this additional length is not too large. For \( K = 2 \) however, the length of these CIs increases extremely. This is due to \( t_{1.075} \approx 12.7 \). The huge length is also associated with a decrease in power for the corresponding test. The HKSJ-SJ and the pBoot-HKSJ-SJ CI cannot be recommended in the case \( K = 2 \).

As was shown before, the WTz-DL CI often has a coverage well below the nominal level. Its short length is not an advantage but rather associated with this inappropriate behavior.

The coverage of the Skovgaard CI is also below the nominal level if \( r^2 \) is large and \( K \) is very small. In contrast, for \( r^2 = 0 \), the coverage of the Skovgaard CI is above the nominal level. The behavioral pattern of the Skovgaard CI is to some degree similar to the Bayes-hC (0.5) CI's. However, when looking closer at the problematic case \( K = 2 \), the coverage of the Bayes-hC(0.5) CI does not decrease as heavily as the Skovgaard CI's when \( r^2 \) increases. Apart from that, the length of the Bayes-hC (0.5) CI is usually even lower than the Skovgaard CIs. This suggests a preference of the Bayes-hC(0.5) CI over the Skovgaard CI for the case \( K = 2 \).

### 3.1.2 Location shift analysis

The coverage and the interval length for the location shift analysis are shown in Figure 5 and Figure S11, respectively. Here, we omit the pre-selection steps for the HKSJ CI and the Bayesian CI (as they do not provide any new insights additionally to the results presented here), but they may be found in the online repository for the software code. The same applies to the robustness analysis following in section 3.1.3.

The interval length increases slightly for increasing \( \theta \) (Figure S11). This may be due to the fact that \( \nu_k \) increases with increasing \( \theta \) (see Equation (36) and use the fact that \( c_k \in (0.957,1) \) for the study sizes \( N_k \geq 20 \) used here) so that on average, an increasing \( \theta \) leads to:

- an increasing SE \( \frac{1}{\sqrt{\nu_k}} \) of \( \hat{\theta}(\tau^2) \) given a fixed \( \tau^2 \) (relevant for the WTz-DL, the HKSJ-SJ, and the pBoot-HKSJ-SJ CI)

- a wider likelihood for \( \theta \) given \( \tau^2 \) (relevant for the Skovgaard CI and the Bayes-hC(0.5) CI; for the latter, this yields a wider marginal posterior for \( \theta \)).

Note that in the following, we often have similar observations for the frequentist and the Bayesian methods which is not surprising due to their relationship under the \( \mathcal{U}(-\infty,\infty) \) prior for \( \theta \) when conditioning on \( \tau^2 \) (see section 2.3.5).

For very large \( K \) (\( K = 55 \)), the coverage decreases with increasing \( \theta \). This has already been observed by Sánchez-Meca and Marin-Martínez who explain this by an increasingly negative bias of \( \hat{\theta}(\tau^2) \) for increasing \( \theta \) (see also Viechtbauer). The negative bias itself might be again due to the monotonic relationship between \( \nu_k \) and \( |\theta_k| \). Note that Sánchez-Meca and Marin-Martínez use \( K \in \{5,10,20,40,100\} \) and average the coverage over their scenarios defined by \( K, N_k, \) and \( \tau^2 \). Thus, it seems indeed plausible that Sánchez-Meca and Marin-Martínez observe the same effect as we do for \( K = 55 \). For \( K < 55 \), we cannot observe this effect consistently. This might be due to the Monte Carlo error being too large to detect such small differences.
3.1.3 Robustness analysis

When taking into account the Monte Carlo error, the main conclusion from the plots for the coverage and the interval length in the robustness analysis (Figure 6 and Figure S12, respectively) is that all methods are relatively robust against deviations from the normal distribution of the study-specific effects. The only exceptions might be

FIGURE 7 Coverage of all chosen interval estimators. Effect measure: log odds ratio [Colour figure can be viewed at wileyonlinelibrary.com]
the coverages of the WTz-DL and the Skovgaard CI when the study-specific effects follow the “biN” distribution in the case \( K = 2 \). All other differences in coverage are rather small compared to the Monte Carlo error and not consistent across the values for \( K \). The missing consistency also holds for the interval length: There might be a slight increase in the interval length for the “sN” distribution and an additional slight increase for the “biN” distribution, but not consistently across the values for \( K \).

The performance of the HKSJ-SJ and the pBoot-HKSJ-SJ CI is again very similar and they usually provide the coverage closest to the nominal level. In particular, we may conclude that the pBoot-HKSJ-SJ CI is not needed to ensure robustness.

For a subset of the methods investigated here, this robustness has already been observed—also under additional non-normal distributions.63,84

### 3.2 Log odds ratio

The pre-selection steps (\( r^2 \) estimator for the HKSJ CI, \( \tau \) prior for the Bayesian CrI) are also performed for the log OR, but shown in Section S3. We draw the same conclusions as for the SMD, that is, the SJ estimator is selected for the HKSJ CI and the hC(0.5) prior for the Bayesian CrI.

Here, we will focus on the comparison of all chosen methods. The coverage and the interval length are illustrated in Figure 7 and Figure S13, respectively. Essentially, the results are very similar to those from the SMD. Most importantly, when looking at the overall performance and taking into account the Monte Carlo error, the HKSJ-SJ and the pBoot-HKSJ-SJ CI perform the best in terms of coverage. They are strikingly similar in terms of coverage and interval length. Besides, they are usually the longest intervals, but the difference to the other methods is already small for \( K = 3 \) and it decreases even further for increasing \( K \). For \( K = 2 \) however, the length of these two CIs is extreme compared to the other methods.

As for the SMD, the Skovgaard CI has a coverage below the nominal level if \( r^2 \) is large and \( K \) is very small. To some degree, the Skovgaard CI and the Bayes-hC(0.5) CrI behave again similarly in terms of coverage. In particular, for large \( r^2 \), the Bayes-hC(0.5) CrI sometimes also has a coverage below the nominal level. However, when looking only at \( K = 2 \) (and large \( r^2 \)), the coverage of the Bayes-hC(0.5) CrI is still higher than the Skovgaard CI’s. The inferiority of the WTz-DL CI in terms of coverage is proven here again.

Concerning the behavior of the coverage with respect to \( p_C \), we observe that for the WTz-DL CI, the Skovgaard CI, and the Bayes-hC(0.5) CrI, the coverage usually decreases for common events \( (p_C = 0.5) \) compared to rare events \( (p_C = 0.1) \). If the coverage is already above the nominal level for both \( p_C \) values (usually for small to moderate \( r^2 \)), this yields a “common events” coverage closer to the nominal level than the “rare events” coverage. In contrast, if the coverage is already below the nominal level for both \( p_C \) values (especially for large \( r^2 \)), this yields a “common events” coverage farther away from the nominal level. Interestingly, for small \( K \) \( (K \in \{2, 3\}) \), the HKSJ-SJ and the pBoot-HKSJ-SJ CI usually behave the opposite way (see also Figure S5 where this is also the case for the HKSJ-PM CI).

The ranking of the interval estimators with respect to their length is the same as for the SMD. The influence of \( p_C \) on the interval length is very consistent: Rare events lead to longer intervals than common events and this is more pronounced as \( r^2 \) decreases. The longer intervals for rare events might be due to increasing SEs \( \nu_k \) of the log OR estimators \( \hat{\theta}_k \) in case of rare events. For the WTz-DL, the HKSJ-SJ, and the pBoot-HKSJ-SJ CI, this leads to an increasing SE \( \sqrt{1/w(r^2)} \) of \( \theta(r^2) \) given a fixed \( r^2 \) and this also explains why this phenomenon is more pronounced as \( r^2 \) decreases: For small \( r^2 \), the \( \nu_k^2 \) dominate the precision weights \( w_k(r^2) = 1/(r^2 + \nu_k^2) \). For the Skovgaard CI and the Bayes-hC(0.5) CrI, the increasing \( \nu_k \) lead to a wider likelihood for \( \theta \) given \( r^2 \) (yielding a wider marginal posterior for \( \theta \) in case of the Bayes-hC(0.5) CrI). Again, this also explains why this phenomenon is more pronounced as \( r^2 \) decreases: For small \( r^2 \), the \( \nu_k^2 \) dominate the term \( (r^2 + \nu_k^2) \) occurring in the likelihood (see Equation (14)).

### 3.3 Recommendations for applied researchers

We want to keep our summarizing recommendations as simple as possible. Thus, even though that for increasing \( K \), the performance of all methods becomes more and more similar, we generally recommend the HKSJ-SJ CI as it performs best in most scenarios.

### Table 3

| Study          | MD   | SMD  |
|----------------|------|------|
| Stamm et al    | 6.30 | 0.84 |
| Hendrikx et al | 2.50 | 0.27 |
| Ahmadi et al   | -1.40| -0.26|
| Mocini et al   | 4.00 | 0.58 |
| Patel et al    | 10.20| 3.76 |
| Zhao et al     | 9.40 | 1.25 |

Abbreviations: MD, mean difference; SMD, standardized mean difference.
estimate of $θ^*$ to the MD (an assumption supported by Veroniki et al.
1) that our recommendations from section 3.3 also apply
LVEF difference significantly. Under the assumption
the additional stem cell transplantation improves the
considered here is the left-ventricular ejection fraction
bypass grafting while the control group only underwent
the coronary artery bypass grafting. The outcome
performed best in terms of coverage and still
better than the Skovgaard CI in terms of interval length.
Thus, for $K = 2$, we recommend the Bayes-hC(0.5) CrI.

Note that the above recommendations hold for the
SMD as well as for the log OR.

4 | EXAMPLE FROM CARDIAC SURGERY

We illustrate our recommendations using a meta-analysis
conducted by Donndorf et al.
85 where the treatment
group received an autologous intramyocardial bone mar-
row stem cell transplantation during coronary artery
bypass grafting while the control group only underwent
the coronary artery bypass grafting. The outcome
considered here is the left-ventricular ejection fraction
(LVEF) difference, calculated as the LVEF after the
intervention (in %) minus the LVEF before the interven-
tion (in %). For this outcome, $K = 6$ studies were included
in the meta-analysis. The effect measure used by
Donndorf et al.
85 is the mean difference (MD), that is,
$θ_k := μ_T, k − μ_C, k$, $δ_k := X_{T, k} − X_{C, k}$, and
$τ^2_k := (\frac{1}{N_{T, k}} + \frac{1}{N_{C, k}}) \hat{σ}^2_k$.
The results from the 6 studies are shown in Table 3.
Using the WTz-DL method, Donndorf et al.
85 obtain a
0.95 CI for $θ$ of $[1.36, 9.44]$ with a corresponding point
estimate of $\hat{θ}(τ^2_{DL}) \approx 0.90$. We wish to emphasize again
that the WTz-DL CI is not recommended, though. The
HKSJ-SJ method, which is recommended in this case,
would have yielded a 0.95 CI of $[-0.40, 2.34]$ with a point
estimate of $\hat{θ}(τ^2_{SJ}) \approx 0.97$. Thus, the WTz-DL CI would
have resulted in an overall treatment effect significantly
different from 0 and the HKSJ-SJ CI not. We assume that
these different conclusions arise because of the failure of
the WTz-DL CI to maintain the nominal coverage level.
For this specific example with small study sizes
85 and for all investigated $τ^2$ values greater than 0, this can be seen in
the 4 rightmost subplots in the first row of Figure 4.

This example also demonstrates that the conclusions
drawn from MD and SMD analyses do not need to coinc-
cide. Possible reasons are:

1. The SMD includes the standard deviation estimator $\hat{σ}_k$
in its denominator which might introduce additional
variability (seen with respect to the scale, e.g., by
looking at the variation coefficient).

2. The true distribution of the SMD estimator $\hat{θ}_k$ (given
$θ_k$) is a scaled noncentral $t$ distribution,
so the normal distribution is just an approximation, even when
assuming normally distributed patient-level observations.
Under the same assumption, the MD estimator
exactly follows a normal distribution.

3. For the SMD, $ν_k$ and $|θ_k|$ follow a positive monotonic
relationship (see section 3.1.2).

As all studies in the meta-analysis by Donndorf et al.
85 measure the LVEF difference in the same unit, we consider
the MD to be a legitimate effect measure
26,92,93 which allows for an easier interpretation.
Nevertheless, we point out that even when using the MD, the significance of the
difference of the overall treatment effect from 0 is only
weak (HKSJ-SJ CI: $[0.71, 10.19]$). If the null hypothesis
would not have been a point hypothesis ($θ = 0$), but
rather an interval hypothesis ($θ ∈ [−δ, δ]$ with a pre-
specified $δ > 0$), the significance would probably not have persisted. An interval hypothesis might be more
helpful, especially for a medical interpretation.

5 | DISCUSSION

The approach for correcting “zero cells” when analyzing
binary data has a crucial impact on the coverage.
94 The
approach chosen here (add $\frac{1}{2}$ to all cells of studies having at least one zero cell) was shown to be preferable over the approach of adding $\frac{1}{2}$ to all cells of all studies (no matter if they have a zero cell or not), especially for rare events. A zero-cell correction may be avoided in the first place, for example, by using the arcsine difference as effect measure, or by using the BNHM—mentioned in section 2.1—instead of the NNHM, either in a Bayesian framework or in a frequentist framework.

In case of a meta-analysis of only $K = 2$ studies, the RE model might not be the optimal model in the first place since it basically requires estimating a variance (the between-study variance $\tau^2$) with only two data points and since a power gain (compared to the average power of the individual studies) is not to be expected for $K < 5$. Bender et al. investigate the case $K \in \{2,3,4\}$ more deeply than we do. Their conclusions are very detailed (also including the FE model and a qualitative evidence synthesis, apart from the HKSJ and the Bayesian approach) and thus, we refer readers interested in meta-analyses with small $K$ to their paper.

### 5.1 Interval estimators

For the estimation of $\tau^2$ in the HKSJ CI, we did not consider the REML estimator as it is closely related to the PM estimator. We could confirm this similarity for the HKSJ CI in the SMD main analysis of our simulation (results not shown).

Our simulation results provide an additional justification for excluding the mHKSJ CI (cf. section 2.3.6): First, the mHKSJ CI would not have helped for $K = 2$ since it is always at least as long as the HKSJ CI. Second, for general $K$, the coverage of the HKSJ CI is already close to the nominal level so that a substantial improvement is not to be expected by the mHKSJ CI. However, note that the mHKSJ CI has been recommended if:

1. $K$ is small and the study-specific SEs $\nu_k$ show high variability (the latter being reflected by the mixed study sizes in our simulation study) or
2. binary data is analyzed.

In those situations, we could not observe a remarkable failure of the HKSJ CI. But—as will be discussed in section 5.2—our simulation setup might not be extreme enough in those cases. Furthermore, the mHKSJ CI has been recommended if the estimated between-study variance is zero. However, for the SJ estimator recommended here, this is not possible: It gives a positive estimate for $\tau^2$ almost surely.

In a personal communication, the authors of the metaLik package pointed out that the Skovgaard method was mainly intended to correct the likelihood ratio test, not the corresponding CI (the ordinary PL CI). Thus, when comparing different tests especially in terms of power, the Skovgaard method is still worth to be included.

Concerning the bootstrap CI investigated here (the pBoot-HKSJ-SJ CI), our simulation results show that its additional computational and theoretical complexity (compared to the HKSJ-SJ CI) does not provide any remarkable advantage for applied researchers. This is why we favor the simpler HKSJ-SJ CI instead.

Because of the general downsides of so-called “non-informative” priors in a Bayesian analysis (see section 2.3.5), we also had a look at a weakly informative prior for the overall treatment effect $\theta$ instead of the $U(-\infty, \infty)$ prior: For the log OR, Günhan et al. derived the $\mathcal{N}(0,2.82^2)$ distribution as a weakly informative prior for $\theta$. We tested this prior (together with the hC(0.5) prior for $\tau$), but found little differences in coverage and interval length compared to the $U(-\infty, \infty)$ prior, at least for our preset value of $\theta = 0.5$ (see Figures S14 and S15). This in accordance with the results from Günhan et al. The only minor systematic difference between the two priors for $\theta$ observable here is that for small $K$, the weakly informative prior yields a slightly shorten CrI (see Figure S15), which is also in accordance with Günhan et al. This shorter length for small $K$ may be directly explained by the narrower prior for $\theta$, carrying forward its narrowness to the marginal posterior for $\theta$ in case of sparse data.

Our choice of the half-Cauchy distribution as a weakly informative prior for $\tau$ was mainly based on theoretical considerations (see section 2.3.5). However, it may be interesting to include other priors in the comparison: Under the assumption of a “constant rate penalisation,” Simpson et al. derive the exponential distribution as a weakly informative prior for $\tau$ (even though not in a meta-analytic context). In future studies, it might be interesting to compare the exponential prior to the half-Cauchy prior. In this case, the gamma distribution (as a generalization of the exponential distribution) could also be investigated. Hamaguchi et al. investigate various “noninformative” priors for $\tau$ (and for transformations of $\tau$) by simulation. They focus on prediction intervals, but also investigate CrIs for $\theta$, with very similar results for these two types of intervals. Like in our work, they also observe that the performance of the Bayesian CrI largely depends on $\tau^2$ (apart from the dependence on other simulation parameters like $K$), so our general recommendation of the HKSJ-SJ CI for $K > 2$ is not challenged by the additional priors investigated by Hamaguchi et al. However, the special case $K = 2$ is not investigated by Hamaguchi et al., so we ran parts of our simulation study (mixed study sizes) with $K = 2$ again for the priors performing best in the simulation study by Hamaguchi
et al. The best performing priors were “Jeffreys’ prior”,43,105,106 and the “conventional prior”.43,107 For Jeffreys’ prior, the results by Hamaguchi et al46 already indicated that the Bayesian CI might be longer than the HKSJ-SJ CI and our additional simulation confirmed this (results not shown). Thus, we did not investigate Jeffreys’ prior further. For the conventional prior, we received a lot of warnings from the R2 package bayesmeta43 (about 10% of all scenario replications). Thus, before being able to interpret the results for the conventional prior reliably, future investigations first need to clarify where these warnings come from. The remaining 90% of the replications (results not shown) indicate that the conventional prior might perform slightly better than the hC(0.5) prior. Note that Hamaguchi et al46 also mention the “Berger-Deely prior”.43,107 The U(0,10) prior for $\tau$ and the U(0, $\infty$) prior for $\alpha$ among the better performing priors, but the Berger-Deely prior leads to an even longer CrI than Jeffreys’ prior,46 the U(0,10) prior performs similarly to the U(0,$\infty$) prior,46 and the U(0,$\infty$) prior was already included in our simulation study.

Note that the Bayesian CI may always be used in a fully Bayesian meta-analysis. Apart from the typical advantages of Bayesian analyses in general108 (most importantly, a full posterior distribution for $\theta$ and a more intuitive interpretation of the interval estimate), this has the advantage that posteriors for the $\theta_k$ and posterior predictive distributions (for the $\theta_k$ as well as for the $\theta_0$) are easily obtained.43 In a fully Bayesian meta-analysis, using a weakly informative prior for $\theta$ instead of the U($-\infty$,$\infty$) prior might be preferred.12,44,98,109 In that case, if a weakly informative prior with heavier tails than the normal distribution is desired (e.g., a Cauchy distribution109), the R2 package bayesmeta43 cannot be used anymore. The computational flexibility required for such an analysis is offered, for example, by the R2 package brms110,111 which is based on Stan112 and its R2 interface package rstan.113 Note that in a fully Bayesian meta-analysis, the U(0,$\infty$) prior for $\tau$ clearly does not make sense as authors of real-world meta-analyses usually only include studies with comparable designs (which restricts their heterogeneity). Thus, the choice of a prior for $\tau$ is then rather between a weakly informative (e.g., half-Cauchy with scale 0.5) and an informative prior. For the latter, a context-specific setting of the parameters offered by Turner et al48 and Rhodes et al47 may be used.

The following CIs have not been further investigated by Veroniki et al1 and we also excluded them here (as explained below):

1. a quantile-based mHKSJ CI which constrains

$$q(\hat{z}^2) > \left( \frac{z_{1-\alpha/2}}{\sqrt{k-1.1}} \right)^2,$$ 114-116

2. modifications of the HKSJ CI using alternative SE estimators for $\hat{\theta}^2$ 27,117

3. the Kenward-Roger CI,118

4. several exact CIs based on inverting hypothesis tests,119,120

5. an exact CI based on Monte Carlo sampling and inverting a hypothesis test,121

6. a CI based on a conditional Monte Carlo approach and the inversion of a hypothesis test.122

We excluded the quantile-based mHKSJ CI since it is very similar to the mHKSJ CI. This does not only refer to their conceptual similarity: In the SMD main analysis of our simulation study (see section 2.5), we observed that in general, the quantile-based mHKSJ CI is overly conservative (results not shown) which has already been observed—to a larger extent—for the mHKSJ CI.56,123 However, like the mHKSJ CI, the quantile-based mHKSJ CI may be more appropriate than the HKSJ CI in case of extremely varying within-study precisions. Our study sizes might not be mixed extremely enough to detect this (see section 5.2).

We excluded the alternative SE estimators27,117 for the HKSJ CI since the simulation study by Partlett and Riley123 showed that they lead to undercoverage and therefore perform worse than the HKSJ CI.

We excluded the Kenward-Roger CI since the simulation study by Partlett and Riley123 showed that in general, it is overly conservative. For extremely mixed study sizes however, the Kenward-Roger CI might perform better than the HKSJ CI.123 Again, our study sizes might not be mixed extremely enough (see section 5.2).

The exact CIs by Liu et al119 and Wang and Tian120 were excluded here since they suffer from the same problem as the FP CI: They are not appropriate for $K < 6$ and $r = 0.05$.

Michael et al121 state that their exact CI performs similarly to the Bayesian CI with “a typical” noninformative prior. This was the main reason why it was excluded here, but when trying to apply their method in our simulation study (via the R2 package rma.exact124), we also ran into a lot of errors (more than with any included method).

The conditional Monte Carlo CI by Sugawasa and NoMA122 might be interesting for future investigations but we had to exclude it here for technical reasons (the runtime being about 11 times higher than for the Bayesian CI which is already the most time-consuming method of our simulation study).

### 5.2 Simulation settings

Our simulation scenarios were chosen with a focus on medical research. Even though our simulation study is...
extensive, it cannot cover all possible settings. Our conclusions are restricted to the settings we have investigated.

In a real-world meta-analysis, the assumption of variance homogeneity (i.e., the same $\sigma_k^2$ in both treatment groups) is most probably violated. However, extending our simulation study to include scenarios with heterogeneous variances would be outside of the scope of our article.

When simulating log ORs, the procedure for generating the event probabilities $p_{T,k}$ and $p_{C,k}$ has a non-negligible impact, especially for small study sizes $N_k$. For the HKSJ CI however, this influence is rather small so that our general recommendation of the HKSJ-SJ CI should remain valid.

The probability of $p_C = 0.1$ might be considered too large for actually rare events as they occur, for example, for safety endpoints of a clinical trial.85,98

In our “mixed” setting, the study sizes $N_k$ could have been varied more extremely and not as smooth as implied by the Delaporte distribution. This might lead to conclusions comparable to Röver et al56 and Partlett and Riley123 who observed a decreasing coverage of the HKSJ CI if the studies are of varying precision.

For the bimodal normal mixture in the robustness analysis, we note that a mixture ratio other than 50:50 (in particular, an extreme mixture with one component being much more probable than the other) might have altered our results and would be interesting for future investigations.

Finally, our simulation study does not answer the question which method to choose in meta-regressions or in meta-analyses where a publication bias is deemed possible. For the latter, the HC CI (see section 2.3.6) or the IVhet CI might be methods to include in a comparison.

ACKNOWLEDGMENTS

The authors are grateful for the helpful comments made by the Editor, the Associate Editor, and two anonymous reviewers during the revision process. Open Access funding enabled and organized by ProjektDEAL. Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The $R^2$ code used for this article as well as its output is available online3.

ORCID

Frank Weber © https://orcid.org/0000-0002-4842-7922

ENDNOTES

* Gelman49 investigates the special case $\nu_1^2 = \ldots = \nu_k^2$, but we consider his main results to be transferable to the RE model anyhow.

† Poison and Scott50 investigate the special case $\nu_1^2 = \ldots = \nu_k^2$ and $\theta = 0$. Besides, they impose the half-Cauchy prior with a fixed scale of $1 \times 0.5$ instead of a free-scaled half-Cauchy prior on $\tau$. Nevertheless, we consider their main results to be transferable to the RE model.

‡ Günhan et al99 investigate rare events in the BNHM and as a “noninformative” prior for $\theta$, they use a very vague $N(0,100^2)$ prior for $\theta$ instead of a $\mathcal{U}(-\infty,\infty)$ prior. They also investigate a wider range of values for $\theta$.

REFERENCES

1. 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. https://doi.org/10.1002/jrsm.1319.

2. R Core Team. R: A language and environment for statistical computing, Vienna, Austria: R Foundation for Statistical Computing; 2020. https://www.R-project.org/.

3. Weber F, Knapp G, Glass A, Kundt G, Ickstadt K. Code and Output for "Interval estimation of the overall treatment effect in random-effects meta-analyses: Recommendations from a simulation study comparing frequentist, Bayesian, and bootstrap methods". OSF Home; 2020. https://doi.org/10.17605/OSF.IO/86VUW

4. Friede T, Röver C, Wandel S, Neuenschwander B. Meta-analysis of two studies in the presence of heterogeneity with applications in rare diseases. Biom J. 2017;59(4):658-671. https://doi.org/10.1002/bimj.201500236.

5. Friede T, Röver C, Wandel S, Neuenschwander B. Meta-analysis of few small studies in orphan diseases. Res Synth Methods. 2017;8(1):79-91. https://doi.org/10.1002/jrsm.1217.

6. Bender R, Friede T, Koch A, et al. Methods for evidence synthesis in the case of very few studies. Res Synth Methods. 2018;9(3):382-392. https://doi.org/10.1002/jrsm.1297.

7. Laird NM, Mosteller F. Some statistical methods for combining experimental results. Int J Technol Assess Health Care. 1990;6(1):5-30. https://doi.org/10.1017/S0266462300008916.

8. Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. Stat Med. 2001;20(6):825-840. https://doi.org/10.1002/sim.650.

9. Viechtbauer W. Bias and efficiency of meta-analytic variance estimators in the random-effects model. J Educ Behav Stat. 2005;30(3):261-293. https://doi.org/10.3102/1076998603003261.

10. Hedges LV, Vevea JL. Fixed- and random-effects models in meta-analysis. Psychol Methods. 1998;3(4):486-504. https://doi.org/10.1037/1082-989X.3.4.486.

11. Schmidt F. Meta-analysis: A constantly evolving research integration tool. Organ Res Methods. 2008;11(1):96-113. https://doi.org/10.1177/1094428107303161.

12. Williams DR, Rast P, Bürkner PC. Bayesian meta-analysis with weakly informative prior distributions. PsyArXiv. 2018. https://doi.org/10.31234/osf.io/7tbrm.
13. Veroniki AA, Jackson D, Viechtbauer W, et al. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Res Synth Methods*. 2016;7(1):55-79. https://doi.org/10.1002/jsm.1164.

14. 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. https://doi.org/10.1002/jsm.1198.

15. Paule RC, Mandel J. Consensus values and weighting factors. *A simple confidence interval for meta-analysis*. *J Stat Soc Ser C Appl Stat*. 1992;41:395-411. https://doi.org/10.1002/sim.4780140406.

16. Sidik K, Jonkman JN. A comparison of heterogeneity variance estimators in meta-analysis. *Control Clin Trials*. 1986;7(3):177-188. https://doi.org/10.1016/0197-2456(86)90046-2.

17. Berkey CS, Hoaglin DC, Mosteller F, Colditz GA. A random-effects regression model for meta-analysis. *Stat Med*. 1995;14(4):395-411. https://doi.org/10.1002/sim.4780140406.

18. DerSimonian R, Kacker R. Random-effects model for meta-analysis. *Stat Med*. 2001;20(12):1771-1782. https://doi.org/10.1002/sim.791.

19. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Contemp Clin Trials*. 2005;26(9):1964-1981. https://doi.org/10.1016/j.cct.2006.04.004.

20. Hartung J, Knapp G, Sinha BK. *Statistical Meta-Analysis with Applications*. Hoboken, NJ: Wiley; 2008.

21. Sidik K, Jonkman JN. A comparison of heterogeneity variance estimators in combining results of studies. *Stat Med*. 2007;26(9):1964-1981. https://doi.org/10.1002/sim.2688.

22. Hartung J, Knapp G, Sinha BK. *Statistical Meta-Analysis with Applications*. Hoboken, NJ: Wiley; 2008.

23. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw*. 2010;36(1):1-48. https://doi.org/10.18637/jss.v036.i03.

24. Sidik K, Jonkman JN. Simple heterogeneity variance estimation for meta-analysis. *J R Stat Soc Ser C Appl Stat*. 2005;54(2):367-384. https://doi.org/10.1111/j.1467-9876.2005.00489.x.

25. Hartung J, Knapp G. A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Stat Med*. 2001;20(24):3875-3889. https://doi.org/10.1002/sim.1009.

26. Hartung J, Knapp G. On tests of the overall treatment effect in meta-analysis with normally distributed responses. *Stat Med*. 2001;20(12):1771-1782. https://doi.org/10.1002/sim.791.

27. Sidik K, Jonkman JN. Simple heterogeneity variance estimation for meta-analysis. *J R Stat Soc Ser C Appl Stat*. 2005;54(2):367-384. https://doi.org/10.1111/j.1467-9876.2005.00489.x.

28. Sánchez-Meca J, Marin-Martínez F. Confidence intervals for the overall effect size in random-effects meta-analysis. *Psychol Methods*. 2008;13(1):31-48. https://doi.org/10.1037/1082-989X.13.1.31.

29. Skovgaard IM. An explicit large-deviation approximation to one-parameter tests. *Beroulli (Andover)*. 1996;2(2):145-165. https://doi.org/10.2307/3318548.

30. Guolo A. Higher-order likelihood inference in meta-analysis and meta-regression. *Stat Med*. 2012;31(4):313-327. https://doi.org/10.1002/sim.4451.

31. Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. *Stat Med*. 1996;15(6):619-629. https://doi.org/10.1002/sim.4780150606.

32. Guolo A, Varin C. The R package metaLik for likelihood inference in meta-analysis. *J Stat Softw*. 2012;50(1):1-14. https://doi.org/10.18637/jss.v050.i07.

33. Turner RM, Davey J, Clarke MJ, Thompson SG, Higgins JPT. Predictive distributions for the extent of heterogeneity in meta-analyses of continuous outcome data. *J Clin Epidemiol*. 2015;68(1):52-60. https://doi.org/10.1016/j.jclinepi.2014.08.012.

34. Gelman A. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper).
50. Polson NG, Scott JG. On the half-Cauchy prior for a global scale parameter. _Bayesian Anal._ 2012;7(4):887-902. doi:10.1214/12-BA117A.

51. James W, Stein C. Estimation with Quadratic Loss. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics. The Regents of the University of California; 1961. Retrieved from https://projecteuclid.org/euclid.bsmsp/1200512173.

52. Anzures-Cabrera J, Sarpatwari A, Higgins JPT. Expressing findings from meta-analyses of continuous outcomes in terms of risks. _Stat Med._ 2011;30(25):2967-2985. doi:10.1002/sim.4298.

53. Röver C, Friede T. Discrete approximation of a mixture distribution via restricted divergence. _J Comput Graph Stat._ 2017;26(1):217-222. doi:10.1080/10618600.2016.1276840.

54. Brockwell SE, Gordon IR. A simple method for inference on an overall effect in meta-analysis. _Stat Med._ 2007;26(25):4531-4543. doi:10.1002/sim.2883.

55. Knapp G, Hartung J. Improved tests for a random effects meta-regression with a single covariate. _Stat Med._ 2003;22(17):2693-2710. doi:10.1002/sim.1482.

56. Röver C, Knapp G, Friede T. Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies. _BMC Med Res Methodol._ 2015;15(1):99. doi:10.1186/s12874-015-0091-1.

57. Bartlett MS, Fowler RH. Properties of sufficiency and statistical tests. _Proc R Soc Lond A Math Phys Sci._ 1937;160(901):268-282. doi:10.1098/rspa.1937.0109.

58. Noma H. Confidence intervals for a random-effects meta-analysis based on Bartlett-type corrections. _Stat Med._ 2011;30(28):3304-3312. doi:10.1002/sim.4350.

59. 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. doi:10.1348/000711010X522687.

60. Guolo A, Varin C. Random-effects meta-analysis: The number of studies matters. _Stat Methods Med Res._ 2017;26(3):1500-1518. doi:10.1177/0962280215583568.

61. Henmi M, Copas JB. Confidence intervals for random effects meta-analysis and robustness to publication bias. _Stat Med._ 2010;29(29):2969-2983. doi:10.1002/sim.4029.

62. Biggerstaff BJ, Tweedie RL. Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. _Stat Med._ 1997;16(7):753-768. doi:10.1002/(SICI)1097-0258(19970415)16:7<753::AID-SIM494>3.0.CO;2-G

63. 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. doi:10.1177/0962280210392008.

64. Preuβ M, Ziegler A. A simplification and implementation of random-effects meta-analyses based on the exact distribution of Cochran’s Q. _Methods Inf Med._ 2014;53(1):54-61. doi:10.3414/ME13-01-0073.

65. Zeng D, Lin DY. On random-effects meta-analysis. _Biometrika._ 2015;102(2):281-294. doi:10.1093/biomet/asv011.

66. 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. doi:10.1093/biomet/asx001.

67. Hedges LV. Distribution theory for Glass’s estimator of effect size and related estimators. _J Educ Stat._ 1981;6(2):107-128. doi:10.3102/10769986006002107.

68. Hedges LV. A random effects model for effect sizes. _Psychol Bull._ 1983;93(2):388-395. doi:10.1037/0033-2909.93.2.388.

69. Petropoulou M, Mavridis D. A comparison of 20 heterogeneity variance estimators in statistical synthesis of results from studies: A simulation study. _Stat Med._ 2017;36(27):4266-4280. doi:10.1002/sim.7431.

70. Paniyakul T, Burumnsup C, Knapp G. On estimating residual heterogeneity in random-effects meta-regression: A comparative study. _J Stat Theory Appl._ 2013;12(3):253-265. doi:10.1002/jsta.13.12.3.4.

71. Doi SAR, Barendregt JJ, Khan S, Thalib L, Williams GM. Advances in the meta-analysis of heterogeneous clinical trials I: The inverse variance heterogeneity model. _Contemp Clin Trials._ 2015;45:130-138. doi:10.1016/j.cct.2015.05.009.

72. Doi SAR, Barendregt JJ, Khan S, Thalib L, Williams GM. Advances in the meta-analysis of heterogeneous clinical trials II: The quality effects model. _Contemp Clin Trials._ 2015;45:123-129. doi:10.1016/j.cct.2015.05.010.

73. Kahle W, Mercier S, Paroissin C. Degradation Processes in Reliability. Vol 3. London: ISTE Ltd; 2016.

74. Johnson N, Kemp AW, Kotz S. _Univariate Discrete Distributions._ Hoboken, NJ: Wiley; 2005.

75. Trikalinos TA, Trow P, Schmid CH. _Simulation-Based Comparison of Methods for Meta-Analysis of Proportions and Rates._ Research Report 13(14)-EHC084-EF. Rockville, MD: Agency for Healthcare Research and Quality (AHRQ); 2013. https://effectivehealthcare.ahrq.gov/products/metanalysis-proportion-rate-comparison/research/. doi:10.3414/ME13-01-0073.

76. Cohen J. _Statistical Power Analysis for the Behavioral Sciences._ Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.

77. Chen H, Cohen P, Chen S. How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. _Commun Stat Simul Comput._ 2010;39(4):860-864. doi:10.1080/03610910903650383.

78. Würtz D, Setz T, Chalabi Y, Boudt C, Chausse P, Miklovac M. fGarch: Rmetrics—Auto Regressive Conditional Heteroskedastic modelling. R package version 3.0.1.82.1; 2017. Retrieved from https://cran.r-project.org/package=fGarchURL.

79. Holzmann H, Vollmer S. A likelihood ratio test for bimodality in two-component mixtures with application to regional income distribution in the EU. _Adv Stat Anal._ 2008;92(1):57-69. doi:10.1007/s10182-008-0057-2.
Ahmadi H, Baharvand H, Ashtiani SK, et al. Safety analysis and improved cardiac function following local autologous transplantation of CD133+ enriched bone marrow cells after myocardial infarction. Circulation 2007;116(13):1537-1542. https://doi.org/10.1161/CIRCULATIONAHA.106.689371.

Donndorf P, Kundt G, Kaminski A, et al. Intramyocardial bone marrow cell transplantation during coronary artery bypass surgery: A meta-analysis. J Thorac Cardiovasc Surg. 2011;142(4):911-920. https://doi.org/10.1016/j.jtcvs.2010.12.013.

Stamm C, Kleine HD, Choi YH, et al. Intramyocardial delivery of CD133+ bone marrow cells and coronary artery bypass grafting for chronic ischemic heart disease: Safety and efficacy studies. J Thorac Cardiovasc Surg. 2007;133(3):717-725. https://doi.org/10.1016/j.jtcvs.2006.08.077.

Hendrikx M, Hensen K, Clijsters C, et al. Recovery of regional but not global contractile function by the direct intramyocardial autologous bone marrow transplantation. Circulation. 2006;114:1-101-107. https://doi.org/10.1161/CIRCULATIONAHA.105.000505.

Ahmadi H, Baharvand H, Ashtiani SK, et al. Safety analysis and improved cardiac function following local autologous transplantation of CD133+ enriched bone marrow cells after myocardial infarction. Curr Neurovasc Res. 2007;4(3):153-160. https://doi.org/10.2174/156720207781387141.

Mocini D, Staibano M, Mele L, et al. Autologous bone marrow mononuclear cell transplantation in patients undergoing coronary artery bypass grafting. Am Heart J. 2006;151(1):192-197. https://doi.org/10.1016/j.ahj.2005.02.001.

Patel AN, Geffner L, Vina RF, et al. Surgical treatment for congestive heart failure with autologous adult stem cell transplantation: A prospective randomized study. J Thorac Cardiovasc Surg. 2005;130(6):1631-1638. https://doi.org/10.1016/j.jtcvs.2005.07.056.

Zhao Q, Sun Y, Xia L, Chen A, Wang Z. Randomized study of mononuclear bone marrow cell transplantation in patients with coronary surgery. Ann Thorac Surg. 2008;86(6):1833-1840. https://doi.org/10.1016/j.athoracsur.2008.08.068.

Higgins JPT, Li T, Deeks JJ. The mean difference (or difference in means). In: JPT H, Thomas J, Chandler J, et al., eds. Cochrane Handbook for Systematic Reviews of Interventions. The Cochrane Collaboration; 2019 Retrieved from https://training.cochrane.org/handbook/current/chapter-06#section-6-5-1-1.

Higgins JPT, Li T, Deeks JJ. Chapter 6.5.1.2: The standardized mean difference. In: Higgins JPT, Thomas J, Chandler J, et al., eds. Cochrane Handbook for Systematic Reviews of Interventions. The Cochrane Collaboration. 2019. Version 6.0 (updated July 2019). URL: https://training.cochrane.org/handbook/current/chapter-06#section-6-5-1-2.

Weber F, Knapp G, Ickstadt K, Kundt G, Glass A. Zero-cell corrections in random-effects meta-analyses. Res Synth Methods. 2020;11(6):913-919. https://doi.org/10.1002/jrsm.1460.
112. Carpenter B, Gelman A, Hoffman MD, et al. Stan: A probabilistic programming language. *J Stat Softw.* 2017;76(1):1-32. https://doi.org/10.18637/jss.v076.i01.

113. Stan Development Team. RStan: The R interface to Stan. R package version 2.21.2; 2020. Retrieved from https://mc-stan.org/.

114. Wiksten A, Rücker G, Schwarzer G. Hartung-Knapp method is not always conservative compared with fixed-effect meta-analysis. *Stat Med.* 2016;35(15):2503-2515. https://doi.org/10.1002/sim.6879.

115. Jackson D, Law M, Rücker G, Schwarzer G. The Hartung-Knapp modification for random-effects meta-analysis: A useful refinement but are there any residual concerns? *Stat Med.* 2017;36(25):3923-3934. https://doi.org/10.1002/sim.7411.

116. Van Aert RCM, Jackson D. A new justification of the Hartung-Knapp method for random-effects meta-analysis based on weighted least squares regression. *Res Synth Methods.* 2019;10(4):515-527. https://doi.org/10.1002/jrsm.1356.

117. Sidik K, Jonkman JN. Robust variance estimation for random effects meta-analysis. *Comput Stat Data Anal.* 2006;50(12):3681-3701. https://doi.org/10.1016/j.csda.2005.07.019.

118. Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics.* 1997;53(3):983-997. https://doi.org/10.2307/2533558.

119. Liu S, Tian L, Lee S, Xie M. Exact inference on meta-analysis with generalized fixed-effects and random-effects models. *Biostat Epidemiol.* 2018;2(1):1-22. https://doi.org/10.1080/24709360.2017.1400714.

120. Wang Y, Tian L. An efficient numerical algorithm for exact inference in meta analysis. *J Stat Comput Simul.* 2018;88(4):646-656. https://doi.org/10.1080/00949655.2017.1402331.

121. Michael H, Thornton S, Xie M, Tian L. Exact inference on the random-effects model for meta-analyses with few studies. *Biometrics.* 2019;75(2):485-493. https://doi.org/10.1111/biom.12998.

122. Sugawara S, Noma H. A unified method for improved inference in random effects meta-analysis. *Biostatistics.* 2019. https://doi.org/10.1093/biostatistics/kcz020.

123. Partlett C, Riley RD. Random effects meta-analysis: Coverage performance of 95% confidence and prediction intervals following REML estimation. *Stat Med.* 2017;36(2):301-317. https://doi.org/10.1002/sim.7140.

124. Michael H, Tian L. rma.exact: Exact confidence intervals for random effects meta-analyses. R package version 0.1.0; 2017. Retrieved from https://CRAN.R-project.org/package=rma.exact.

**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section at the end of this article.

---

**How to cite this article**: Weber F, Knapp G, Glass Ä, Kundt G, Ickstadt K. Interval estimation of the overall treatment effect in random-effects meta-analyses: Recommendations from a simulation study comparing frequentist, Bayesian, and bootstrap methods. *Res Syn Meth.* 2021;12:291–315. https://doi.org/10.1002/jrsm.1471