Received: 2 April 2021 | Revised: 18 February 2022 | Accepted: 25 February 2022

DOI: 10.1002/bimj.202100108

Biometrical Journal

RESEARCH ARTICLE

Jointly pooling aggregated effect sizes and their standard errors from studies with continuous clinical outcomes

Osama Almalik[1] | **Zhuozhao Zhan**[1] | **Edwin R. van den Heuvel**[1,2]

[1] Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands

[2] Preventive Medicine and Epidemiology, Department of Medicine, Boston University, USA

Correspondence
Osama Almalik, Department of Mathematics and Computer Science, Eindhoven University of Technology, Postbus 513, 5600 MB Eindhoven, The Netherlands.
Email: O.Almalik@tue.nl

Funding information
Netherlands Organization for Scientific Research, Grant/Award Number: 023.005.087

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

Abstract

The DerSimonian–Laird (DL) weighted average method for aggregated data meta-analysis has been widely used for the estimation of overall effect sizes. It is criticized for its underestimation of the standard error of the overall effect size in the presence of heterogeneous effect sizes. Due to this negative property, many alternative estimation approaches have been proposed in the literature. One of the earliest alternative approaches was developed by Hardy and Thompson (HT), who implemented a profile likelihood instead of the moment-based approach of DL. Others have further extended this likelihood approach and proposed higher-order likelihood inferences (e.g., Bartlett-type corrections). In addition, corrections factors for the estimated DL standard error, like the Hartung–Knapp–Sidik–Jonkman (HKSJ) adjustment, and the restricted maximum likelihood (REML) estimation have been suggested too. Although these improvements address the uncertainty in estimating the between-study variance better than the DL method, they all assume that the true within-study standard errors are known and equal to the observed standard errors of the effect sizes. Here, we will treat the observed standard errors as estimators for the within-study variability and we propose a bivariate likelihood approach that jointly estimates the overall effect size, the between-study variance, and the potentially heteroskedastic within-study variances. We study the performance of the proposed method by means of simulation, and compare it to DL (with and without HKSJ), HT, their higher-order likelihood methods, and REML. Our proposed approach seems to have better or similar coverages compared to the other approaches and it appears to be less biased in the case of heteroskedastic within-study variances when this heteroskedasticty is correlated with the effect size.

KEYWORDS

DerSimonian–Laird, Hartung–Knapp–Sidik–Jonkman, heteroskedasticity, meta-analysis, profile likelihood

1 | INTRODUCTION

The DerSimonian–Laird (DL) method (DerSimonian & Laird, 1986) has been and still is widely used to estimate an overall effect size from aggregated data (AD) meta-analysis studies. Their pooled effect size is a weighted average of the observed effect sizes, where the weights are the inverse variances of the effect sizes (including both the within- and between-study variability). The between-study variance component is estimated with a moment estimator. The DL method was shown to be negatively biased for standardized differences when the number of studies is small (Malzahn et al., 2000) and it does not account for the uncertainty in estimating the between-study variability (Hardy & Thompson, 1996), potentially leading to liberal confidence intervals for the overall effect size (Veroniki et al., 2019).

Alternative methods have been proposed in the literature to improve the DL method (DerSimonian & Kacker, 2007; Langan et al., 2019; Petropoulou and Mavridis, 2017; Veroniki et al., 2019; Viechtbauer, 2005). One of the first approaches is the profile likelihood approach of Hardy and Thompson (HT) in 1996, where the effect sizes are assumed normally distributed and potentially heterogeneous, but with known within-study variances. The overall effect size and the between-study variance component are then estimated jointly. The authors constructed a confidence interval for the overall effect size that is based on the chi-square distribution of a likelihood ratio statistic (Hardy & Thompson, 1996). It has been shown that this profile likelihood approach has a closer to nominal coverage probability than the DL method (Tanizaki, 2004; Veroniki et al., 2019).

However, the likelihood ratio statistic is only asymptotically chi-square distributed, and for small sample sizes the approximation might be poor (Barndorff-Nielsen & Hall, 1988). For this reason, Noma (2011) proposed a Bartlett-type correction for the likelihood ratio statistic (Noma, 2011). In addition, the author proposed constructing confidence limits using the efficient score statistic and a Bartlett-corrected efficient score function (Cox & Hinkley, 1974; Guolo, 2012). These three methods for confidence intervals of the overall effect size showed conservative coverage probabilities, especially when the number of studies is small, while the DL and the HT methods had liberal coverage probabilities (Cox & Hinkley, 1974).

The Bartlett-type correction of the likelihood ratio statistic is only appropriate for exponential families (Guolo, 2012). The commonly assumed random effects meta-analysis model is a member of the exponential family in the unlikely case of equal within-study variances (Guolo, 2012). Guolo (2012) therefore applied an approximation to the Bartlett-type correction introduced by Skovgaard (2001). This Guolo–Skovgaard (GS) approximation produced conservative coverage probabilities in the case of a small number of studies, but its performance improved when the number of studies increases (Guolo, 2012). In one comparative study, the Bartlett-type correction method and the GS correction method were found to produce similar results (Veroniki et al., 2019).

Alternatively, several correction factors for the estimated standard error of the DL pooled estimator has been suggested (Hartung & Knapp, 2001; Sidik & Jonkman, 2005). These factors attempt to increase the standard error, although in special cases the correction factor can be less than one (Jackson et al., 2017; Partlett and Riley, 2017). The most familiar correction factor uses a weighted sums of squares for the DL pooled estimator, which has been referred to as the Hartung–Knapp–Sidik–Jonkman (HKSJ) correction. This factor has been frequently recommended over the DL approach (e.g., In 't Hout et al., 2014; Langan et al., 2019).

Finally, many alternative estimators for the between-study variance have been suggested as an alternative to the choice of DL to increase the coverage of the asymptotic confidence intervals. These variance estimators include, among others, Cochran's ANOVA or Hedges–Olkin estimator, Paule–Mandel iterative moment-based estimator, and the restricted maximum likelihood (REML) estimators (e.g., DerSimonian & Kacker, 2007; Langan et al., 2019; Petropoulou and Mavridis, 2017). These alternatives often perform better than DL estimator, with the REML estimator frequently preferred over other estimators (Langan et al., 2019; Veroniki et al., 2016), in particular when the clinical outcome for participants in the studies is continuous.

All of the methods discussed so far, assume that the within-study standard deviation is given by the observed standard error of the effect size, while the true within-study variability is unknown in practice. This may lead to bias when some form of correlation between the effect size and the standard error exists (Malzahn et al., 2000). We will assume that the observed standard error is an estimator of the true within-study variability having a chi-square distribution function in line with Cochran (1937). We will introduce a bivariate likelihood approach for estimation of the overall effect size, the between-study variance, and the within-study variances for studies with heteroskedastic continuous clinical outcomes. Using two case studies and a simulation study, we compare our method to DL, HT, the Bartlett-type correction method, the GS correction method, HKSJ correction, and REML.

In Section 2, we describe the different approaches from the literature and our proposed bivariate likelihood approach. The approaches are illustrated on two real case studies that were published in the literature before. Section 3 describes the simulation model we used. It simulates meta-analysis studies with both heterogeneous effect sizes and heteroskedastic within-study standard errors. There is evidence that heteroskedastic errors are common in practice (e.g., Caron et al., 2020; Nilsson et al., 2019; Stevens et al., 2018), but are seldom simulated (Schmidt et al., 2019; van den Heuvel et al., 2021). The results of the simulation study are provided in Section 4 and a discussion is provided in Section 5.

2 | STATISTICAL METHODS

An AD meta-analysis from studies with continuous clinical outcomes usually consists of a set of m effect sizes (e.g., mean differences, correlation coefficients), accompanied with their standard errors and their degrees of freedom (Cochran, 1937, 1954), that is, we observe the triplet (Y_i, S_i, df_i) for study $i = 1, 2, \dots, m$. It is typically assumed that the effect size Y_i is distributed according to the meta-analysis model

$$Y_i = \theta + U_i + \varepsilon_i, \tag{1}$$

with θ the true or overall effect size, $U_i \sim N(0, \tau^2)$ a random effect that is making the effect sizes heterogeneous, $\varepsilon_i \sim N(0, \sigma_i^2)$ a residual, and all random effects mutually independently distributed. The τ^2 is the variance component for the between-study variability and σ_i^2 is the variance component of the within-study variability. In the literature, it is commonly assumed that the within-study variability σ_i^2 is known and given by S_i^2, but we believe that S_i^2 is at best an estimator of σ_i^2. We will therefore treat S_i^2 as an estimator for σ_i^2 and provide a joint likelihood for Y_i and S_i^2 for the estimation of all model parameters.

In Section 2.1, we describe the original DL method and the HKSJ correction. In Section 2.2, the existing (restricted) maximum likelihood-based methods with their finite-sample corrections are presented. Finally, our bivariate method for estimating the overall effect size θ is presented in Section 2.3. For all these methods, we also describe how the 95% confidence intervals are constructed. Section 2.4 presents two case studies from literature where all methods are being demonstrated. One study investigates a mean difference, while the other study combines Spearman's correlation coefficients.

2.1 | The DL method

The DL method first estimates the between-study variance component τ^2 with the moment estimator given by

$$\hat{\tau}_{DL}^2 = \max\left[0, \frac{Q - (m-1)}{\sum_{i=1}^m w_i - \sum_{i=1}^m w_i^2 / \sum_{i=1}^m w_i}\right], \tag{2}$$

where $w_i = 1/S_i^2$, Q is Cochran's Q-statistic given by $Q = \sum_{i=1}^m [(Y_i - \bar{Y})^2 / S_i^2]$ (DerSimonian & Laird, 1986), and $\bar{Y}$ is the weighted average given by $\bar{Y} = \sum_{i=1}^m (Y_i/S_i^2) / \sum_{i=1}^m (1/S_i^2)$. Then the pooled estimator $\hat{\theta}_{DL}$ of the overall effect size θ is calculated using the estimator $\hat{\tau}_{DL}^2$. The DL pooled estimator is given by

$$\hat{\theta}_{DL} = \left[\sum_{i=1}^m Y_i (\hat{\tau}_{DL}^2 + S_i^2)^{-1}\right] \Big/ \left[\sum_{i=1}^m (\hat{\tau}_{DL}^2 + S_i^2)^{-1}\right]. \tag{3}$$

A $(1-\alpha) \times 100\%$ confidence interval on θ may be determined by $\hat{\theta}_{DL} \pm t_{m-1,\alpha/2} S_{DL}$, with $t_{d,q}$ the qth upper quantile of the t-distribution with d degrees of freedom and $S_{DL}^2 = [\sum_{i=1}^m 1/(\hat{\tau}_{DL}^2 + S_i^2)]^{-1}$ the estimated variance of the pooled estimator $\hat{\theta}_{DL}$ having $m-1$ degrees of freedom. Note that it has been more common in the literature to use the normal quantile instead of the quantile of the t-distribution for the DL approach (Brockwell & Gordon, 2007; Jackson et al., 2010; Thorlund et al., 2011), but we believe that DerSimonian and Laird were not explicit on this topic (DerSimonian & Laird, 1986) and therefore did not rule out our preferred choice. In the presence of heterogeneous effect sizes ($\tau^2 > 0$), the precision of the pooled estimator $\hat{\theta}_{DL}$ depends on the number of studies and confidence intervals based on normal quantiles lead to an undercoverage. This has been well-established in the work of Cochran (Cochran, 1954), who proposed to use the t-distribution with $m-1$ degrees of freedom instead of the normal distribution, in particular in the presence of heterogeneity (see also Mzolo et al., 2013). Higher degrees of freedom should be used when no heterogeneity in the effect sizes is present (Cochran, 1954).

The use of a t-distribution is common when the corrected standard error of HKSJ is used (Hartung & Knapp, 2001; Sidik & Jonkman, 2005). The standard error S_{DL} is then multiplied with a data-driven scaling factor $CF = [\sum_{i=1}^{m}(Y_i - \hat{\theta}_{DL})^2/((\hat{\tau}^2_{DL} + S_i^2)(n-1))]^{1/2}$ (see Sidik & Jonkman, 2005). This scaling factor invokes the t-distribution for construction of confidence intervals on θ, although the distribution of $CF \cdot S_{DL}$ is not well understood (Jackson et al., 2017). We will also study this corrected standard error since it has been proposed as the preferred method in the literature (e.g., In 't Hout et al., 2014), but we will use the maximum value of one and this correction factor instead ($\max\{1, CF\}$). Using a correction factor that could potentially be lower than one has been criticized (Jackson et al., 2017; Partlett and Riley, 2017). Note that a comparison of coverages of confidence intervals between the use of this corrected standard error and the traditional DL method with the quantile of the t-distribution has never been investigated, due to the preference for the normal-based confidence interval for DL.

To obtain the estimates $\hat{\tau}^2_{DL}$ and $\hat{\theta}_{DL}$ and the confidence limits on θ from data in our simulation study, we programmed the method in SAS, since most [R] packages seem to have incorporated a normal quantile for the traditional DL, for example, "meta" (Schwarzer, 2007) and "metafor" (Viechtbauer, 2010).

2.2 | Existing likelihood-based methods

Three likelihood-based approaches for parameter estimation and confidence intervals have been proposed in the literature. They all make use of the same maximum likelihood estimators for the parameters θ and τ^2, which is based on the procedure of Hardy and Thompson (1996), but they differ in the construction of confidence intervals.

2.2.1 | The HT method

The log-likelihood function that was proposed in Hardy and Thompson (1996) is given by

$$l(\theta, \tau^2) = -\frac{1}{2} m \log(2\pi) - \frac{1}{2}\sum_{i=1}^{m} \log(\tau^2 + S_i^2) - \frac{1}{2}\sum_{i=1}^{m}(Y_i - \theta)^2/(\tau^2 + S_i^2). \tag{4}$$

It shows that the within-study variances σ_i^2 are assumed known and equal to S_i^2. Maximizing (4) with respect to θ and τ^2 results in solving the following two equations iteratively:

$$\begin{aligned} \theta &= \left[\textstyle\sum_{i=1}^{m} Y_i(\tau^2 + S_i^2)^{-1}\right] / \left[\textstyle\sum_{i=1}^{m} (\tau^2 + S_i^2)^{-1}\right], \\ \tau^2 &= \left[\textstyle\sum_{i=1}^{m} ((Y_i - \theta)^2 - S_i^2)(\tau^2 + S_i^2)^{-2}\right] / \left[\textstyle\sum_{i=1}^{m} (\tau^2 + S_i^2)^{-2}\right]. \end{aligned} \tag{5}$$

The two solutions are HT maximum likelihood estimators $\hat{\theta}_{HT}$ and $\hat{\tau}^2_{HT}$. For the construction of confidence regions on (θ, τ^2) a kind of log-likelihood ratio statistic $T_{HT}(\theta, \tau^2)$ was proposed:

$$T_{HT}(\theta, \tau^2) = -2[l(\theta, \tau^2) - l(\hat{\theta}_{HT}, \hat{\tau}^2_{HT})]. \tag{6}$$

It is assumed that $T_{HT}(\theta, \tau^2)$ is chi-square distributed with 2 degrees of freedom. All pairs of values (θ, τ^2) that would satisfy $T_{HT}(\theta, \tau^2) < \chi^2_2(1-\alpha)$, with $\tau^2 \geq 0$ and $\chi^2_d(q)$ the qth upper quantile of the chi-square distribution with d degrees of freedom, form the $(1-\alpha) \times 100\%$ confidence region on (θ, τ^2) (Hardy & Thompson, 1996).

To obtain confidence intervals for θ and τ^2 separately, a profile likelihood function was considered. Here, we focus on the $(1-\alpha) \times 100\%$ confidence interval for θ, but a similar approach can be applied to τ. If we assume that θ is given, we could maximize the log-likelihood function in (4) for τ first, resulting in the constrained maximum likelihood estimator $\hat{\tau}^2(\theta)$.

Substituting this estimator in (4) results in the profile log-likelihood function $\tilde{l}(\theta) \equiv l(\theta, \hat{\tau}^2(\theta))$. The profile log-likelihood ratio statistic for θ is then defined as

$$\tilde{T}_{HT}(\theta) = -2[\tilde{l}(\theta) - \tilde{l}(\hat{\theta}_{HT})]. \tag{7}$$

All values of θ that would satisfy inequality $\tilde{T}_{HT}(\theta) < \chi^2_1(1-\alpha)$ would form the $(1-\alpha) \times 100\%$ confidence interval for θ.

There may exists several [R] packages that can calculate the profile likelihood-based confidence interval on θ, for example, the [R] package "metaplus" (Beath, 2016), but we used the [R] package "pimeta" (Nagashima et al., 2019). This package will also be used for the higher-order likelihood method in Section 2.2.2. It can determine the maximum likelihood estimators $\hat{\theta}_{HT}$ and $\hat{\tau}^2_{HT}$, the confidence region for (θ, τ^2), and the two confidence intervals for θ and τ^2 from real data. We have used this package for both the case studies and the simulation study.

2.2.2 | The Noma–Bartlett (NB) method

The profile likelihood approach for θ mentioned in Section 2.2.1 is considered a first-order likelihood inference method (Guolo, 2012). Higher-order asymptotic methods for the proposed profile likelihood ratio statistic will provide more accurate inference (Barndorff-Nielsen & Hall, 1988; Cox & Hinkley, 1974), in particular for smaller values of m. Noma (2011) applied a Bartlett-type correction (Barndorff-Nielsen & Hall, 1988) to the profile likelihood ratio statistic $\tilde{T}_{HT}(\theta)$ in (7) by normalizing it with a constant that depends on the constrained maximum likelihood estimator $\hat{\tau}^2(\theta)$. The NB method uses this corrected likelihood ratio statistic, which is given by $\tilde{T}_{NB}(\theta) = \tilde{T}_{HT}(\theta)/[1 + 2C(\hat{\tau}^2(\theta))]$, with

$$C(\tau^2) = \left[\sum_{i=1}^{m}(S_i^2 + \tau^2)^{-3}\right] \Big/ \left[\sum_{i=1}^{m}(S_i^2 + \tau^2)^{-1} \sum_{i=1}^{m}(S_i^2 + \tau^2)^{-2}\right]. \tag{8}$$

The $(1-\alpha) \times 100\%$ confidence interval for θ is formed by all θ's satisfying $\tilde{T}_{NB}(\theta) < \chi^2_1(1-\alpha)$. For the case studies and our simulation study, we obtained the estimates of the overall effect size θ and the NB confidence interval with the [R] package "pimeta" (Nagashima et al., 2019). Note that the NB method uses the estimators $\hat{\theta}_{HT}$ and $\hat{\tau}^2_{HT}$ of HT, and therefore provides only an alternative confidence interval for θ.

2.2.3 | The GS method

Instead of using the profile likelihood ratio statistic $\tilde{T}_{HT}(\theta)$ in (7), a signed profile likelihood ratio statistic can be used:

$$\tilde{r}_G(\theta) = \text{sign}(\hat{\theta}_{HT} - \theta)\sqrt{l(\hat{\theta}_{HT}, \hat{\tau}^2_{HT}) - l(\theta, \hat{\tau}^2(\theta))}. \tag{9}$$

The statistic $\tilde{r}_G(\theta)$ is approximately normally distributed (Guolo, 2012). Thus the set of values θ for which inequalities $z_{\alpha/2} \leq \tilde{r}_G(\theta) \leq z_{1-\alpha/2}$ hold true, with z_q the qth quantile of a standard normal distribution, provides a $(1-\alpha) \times 100\%$ confidence interval for θ.

Alternatively, a Skovgaard correction to the signed profile likelihood ratio statistic in (9) can be applied in a random-effects meta-analysis. This GS corrected statistic is given by

$$\tilde{r}_{GS}(\theta) = \tilde{r}_G(\theta) + [\tilde{r}_G(\theta)]^{-1}\log(\tilde{u}(\theta)/\tilde{r}_G(\theta)), \tag{10}$$

with $\tilde{u}(\theta) = [S^{-1}(\theta)q(\theta)]_1 |I(\hat{\theta}_{HT}, \hat{\tau}^2_{HT})|^{1/2} |J(\hat{\theta}_{HT}, \hat{\tau}^2_{HT})|^{-1} |S(\theta)| |I_{22}(\theta, \hat{\tau}^2(\theta))|^{-1/2}$, $S(\theta)$ the 2×2 matrix given by

$$S(\theta) = \begin{pmatrix} \sum_{i=1}^{m}(S_i^2 + \hat{\tau}^2(\theta))^{-1} & \sum_{i=1}^{m}(\hat{\theta}_{HT} - \theta)(S_i^2 + \hat{\tau}^2(\theta))^{-2} \\ 0 & \left[\sum_{i=1}^{m}(S_i^2 + \hat{\tau}^2(\theta))^{-2}\right]/2 \end{pmatrix}, \tag{11}$$

$q(\theta)$ the vector given by

$$q(\theta) = \begin{pmatrix} \sum_{i=1}^{m} (\hat{\theta}_{HT} - \theta)(S_i^2 + \hat{\tau}^2(\theta))^{-1} \\ -\sum_{i=1}^{m} \left[(S_i^2 + \hat{\tau}^2_{HT})^{-1} - (S_i^2 + \hat{\tau}^2(\theta))^{-1}\right]/2 \end{pmatrix}, \tag{12}$$

$I(\theta, \tau^2)$ the 2×2 Fisher information matrix, $I_{22}(\theta, \tau^2)$ the second diagonal element of $I(\theta, \tau^2)$, $J(\theta, \tau^2)$ the Hessian matrix (i.e., $I(\theta, \tau^2) = -\mathbb{E}J(\theta, \tau^2)$), and $[S^{-1}(\theta)q(\theta)]_1$ the first element of the vector $S^{-1}(\theta)q(\theta)$. The GS $(1-\alpha) \times 100\%$ confidence interval for θ is obtained by the set of values of θ that satisfies $z_{\alpha/2} \leq \tilde{r}_{GS}(\theta) \leq z_{1-\alpha/2}$. These confidence limits will be calculated from data using [R] package "metaLik" (Guolo & Varin, 2012). Also the GS method uses the estimators $\hat{\theta}_{HT}$ and $\hat{\tau}^2_{HT}$ of HT, and constructs only an alternative confidence interval for θ.

2.2.4 | The REML estimation

The restricted log-likelihood function for estimation of the variance component τ^2 is given by (Kontopantelis & Reeves, 2012)

$$\begin{aligned} \ell_{REML}(\tau^2) = &-\frac{1}{2} \textstyle\sum_{i=1}^{m} \log\left(2\pi[\tau^2 + S_i^2]\right) \\ &-\frac{1}{2}\left[\textstyle\sum_{i=1}^{m}[Y_i - \theta_{REML}][\tau^2 + S_i^2]^{-1} + \log\left(\textstyle\sum_{i=1}^{m}[\tau^2 + S_i^2]^{-1}\right)\right] \end{aligned} \tag{13}$$

with $\theta_{REML} = \sum_{i=1}^{m} Y_i[\tau^2 + S_i^2]^{-1} / \sum_{i=1}^{m}[\tau^2 + S_i^2]^{-1}$. A general formulation of the restricted likelihood function for linear mixed models is given by Gurka (2006). Maximizing the log likelihood in (13) with respect to τ^2 can be conducted with procedure MIXED of SAS (see the Appendix). The REML estimator $\hat{\tau}^2_{REML}$ is constrained to nonnegative values. Then, the REML estimator for the overall effect size θ is determined by substituting $\hat{\tau}^2_{REML}$ in θ_{REML} (see also Equation 5), leading to

$$\hat{\theta}_{REML} = \sum_{i=1}^{m} Y_i[\hat{\tau}^2_{REML} + S_i^2]^{-1} / \sum_{i=1}^{m}[\hat{\tau}^2_{REML} + S_i^2]^{-1}. \tag{14}$$

Note that this estimator is the maximum likelihood estimator for the likelihood function in (4) when the between-study variance τ^2 is replaced by $\hat{\tau}^2_{REML}$. Thus the variance of $\hat{\theta}_{REML}$ in (14) is based on the Fisher information matrix and can now be estimated by

$$V(\hat{\theta}_{REML}) = \left[\sum_{i=1}^{m}[\hat{\tau}^2_{REML} + S_i^2]^{-1}\right]^{-1}. \tag{15}$$

Here, we use the $(1-\alpha) \times 100\%$ asymptotic confidence interval $\hat{\theta}_{DL} \pm t_{m-1,\alpha/2-1}S_{REML}$, with $t_{d,q}$ the qth upper quantile of the t-distribution with d degrees of freedom and $S_{REML} = V^{1/2}(\hat{\theta}_{REML})$ the estimated standard error.

2.3 | A bivariate distribution (BD) method

The methods discussed in Sections 2.1 and 2.2 provide estimators and confidence intervals for the parameters θ and τ^2 conditionally on $\sigma_i^2 = S_i^2$. We believe that S_i^2 should be viewed as an estimator for σ_i^2 to be able to address the uncertainty of S_i^2 as an estimator. In this view, it will be unlikely that S_i^2 will be equal to σ_i^2 and therefore may impact the analysis, in particular when S_i^2 is imprecise or possibly not an ideal estimator. Treating S_i^2 as an estimator for σ_i^2, instead of assuming that $\sigma_i^2 = S_i^2$, has been acknowledged in literature (Cochran, 1937; Hardy & Thompson, 1996), but the importance of its uncertainty has been rejected, since it would not or marginally affect the calculation of confidence intervals for θ compared to an analysis where σ_i^2 is assumed equal to S_i^2 (Cochran, 1937; Hardy & Thompson, 1996). It is argued that the estimator for the between-study variance plays a more dominant role in the calculation of confidence intervals on θ than the within-study variances. For meta-analyses with only large studies, we expect that the uncertainty of S_i^2 as estimator will be small, but for smaller studies the imprecision may affect the variability of the pooled estimator $\hat{\theta}$. Thus, we propose to investigate this issue in more detail by considering a joint model for Y_i and S_i^2.

We assume that Y_i follows model (1) and S_i^2 has approximately a chi-square distribution, that is, $df_i S_i^2/\sigma_i^2 \sim \chi^2_{df_i}$, with df_i the degrees of freedom for S_i^2 in line with Cochran (1937). We assume that df_i is either observed or can be calculated from the aggregated information (e.g., from the sample sizes). The need for a degrees of freedom makes our approach most suitable for effect sizes from continuous clinical outcomes (i.e., functions of mean differences, regression parameters of linear regression analyses, and correlation coefficients) where a degrees of freedom exists naturally. Furthermore, we will assume that $\sigma_i^2 \approx \sigma^2\eta_i$, with $\eta_i > 0$ a known parameter value that would typically depend on the sample size of study i. Thus we assume in the analysis of an AD meta-analysis that the residual variances in (1) are considered heteroskedastic across studies as a consequence of different study sizes and that they share (approximately) a common within-study variance parameter σ^2. With the proposed likelihood approach below, this analysis assumption would make the pooled estimator less sensitive to potential correlations between an individual effect size and its standard error unrelated to study size. The conditional analysis methods in Sections 2.1 and 2.2 would be more sensitive to such correlations, because they may incorrectly weigh the observed effect sizes and introduce a bias. It should be noted that an assumption of $\sigma_i^2 \approx \sigma^2\eta_i$ is in line with literature on pooling estimates from biological assays (Cochran, 1954). Furthermore, it helps us maintain a parsimonious model, because estimating m variance parameters σ_i^2 may result in an overfit and will lead to numerical complexities.

In practice, $\sigma_i^2 \approx \sigma^2\eta_i$ may be a strong or unrealistic assumption, but we believe that violation of this assumption may not necessarily lead to serious problems in our analysis. First of all, $\sigma_i^2 \approx \sigma^2\eta_i$ may be viewed as a first-order approximation, since study size is typically one dominant factor that drives differences in precision across studies. Furthermore, factors that may cause violation of the approximation $\sigma_i^2 \approx \sigma^2\eta_i$ will most likely also affect the estimator S_i^2 and therefore the ratio $df_i S_i^2/\sigma_i^2$, which is relevant in the estimation of σ_i^2 through σ^2, may neutralize such violations. In our simulation study, we will introduce heteroskedastic within-study variances and therefore violate the assumption of $\sigma_i^2 \approx \sigma^2\eta_i$ in the data generation process on purpose, but we will still implement $\sigma_i^2 \approx \sigma^2\eta_i$ in the analysis to verify the robustness of our approach.

To illustrate the importance of study size and choices of η_i for the variance σ_i^2, one example is a meta-analysis of Fisher's z transformed correlation coefficients. In such a setting, $\eta_i = [n_i - 3]^{-1}$, with n_i the total sample size for study i. Estimation of the variance parameter σ^2 is then expected to be close to one for Pearson's correlation, since $\sigma_i^2 \approx [n_i - 3]^{-1}$, but close to 1.06 for Spearman's correlation (Fieller & Pearson, 1961). Another example is a meta-analysis of mean differences. In case of homoskedasticity, $\eta_i = [n_{i0}^{-1} + n_{i1}^{-1}]$, with n_{ij} the sample size for the binary exposure $j \in \{0, 1\}$. Then the variance parameter σ^2 represents the between-participant variation within studies (van den Heuvel et al., 2021) that is assumed consistent across studies. For an association of an exposure variable with a continuous clinical outcome that is corrected for confounders, the meta-analysis is concerned with the pooling of a (standardized) regression parameter from a linear regression analysis. The variance of the estimated regression parameter is directly proportional to the degrees of freedom of the residual variance of the linear regression model, that is, $\eta_i = [n_i - p_i]^{-1}$, with n_i the study size of study i and $p_i - 1$ the number of confounders. More generally, we may just consider $\eta_i = [df_i]^{-1}$ as a general approximation to the different choices and view df_i as the effective sample size of study i used to estimate σ_i^2. The variance parameter σ^2 would then become a *nuisance parameter* as a measure of within-study variability without having a direct meaning to the underlying individual data from the studies (see Section 3).

Except for normalizing constants, the log-likelihood function for the BD of (Y_i, S_i^2) is given by

$$l(\theta, \tau^2, \sigma^2) \approx -\frac{1}{2}\left[\sum_{i=1}^{m}\left(\log(\tau^2 + \sigma_i^2) + (Y_i - \theta)^2/(\tau^2 + \sigma_i^2) - df_i \log\left(\chi_i^2\right) + \chi_i^2\right)\right], \tag{16}$$

with $\chi_i^2 = df_i S_i^2/\sigma_i^2$ the chi-square distributed variable and $df = \sum_{i=1}^{m} df_i$ the total number of degrees of freedom. Note that the sum $\sum_{i=1}^{m} \chi_i^2$ in the likelihood (16) is also chi-square distributed with df degrees of freedom (Moschopoulos, 1985). Calculating the likelihood equations for the estimation of the parameters θ, τ^2, and σ^2, leads to the two equations in (5) with S_i^2 replaced by $\sigma_i^2 = \sigma^2/df_i$ and additionally to a third equation

$$\sum_{i=1}^{m} \frac{(Y_i - \theta)^2 - (\tau^2 + \sigma_i^2)}{df_i(\tau^2 + \sigma_i^2)^2} = \sum_{i=1}^{m} \frac{\sigma_i^2 - S_i^2}{\sigma_i^4}. \tag{17}$$

Here, σ^2 can be obtained by applying the Newton–Raphson method (Choi and Wette, 1969) if θ and τ^2 would be given. Estimation of all three parameters θ, τ^2, and σ^2 can be obtained with the procedure NLMIXED of SAS software. The ML estimators of our BD method are referred to as $\hat{\theta}_{BD}$, $\hat{\tau}^2_{BD}$, and $\hat{\sigma}^2_{BD}$. The programming codes for procedure NLMIXED are provided in the Appendix. In the special case that there is no heterogeneity in the effect sizes ($\tau^2 = 0$), the estimator for the overall effect θ and the nuisance parameter σ^2 are given by

$$\hat{\theta}_{BD} = \sum_{i=1}^{m}[df_i Y_i] / \sum_{i=1}^{m} df_i \quad \text{and} \quad \hat{\sigma}^2_{BD} = \frac{1}{m+df}\sum_{i=1}^{m} df_i\left[(Y_i - \hat{\theta}_{BD})^2 + df_i S_i^2\right]. \tag{18}$$

This pooled estimator $\hat{\theta}_{BD}$ is normally distributed with mean θ and variance σ^2/df and the estimator $\hat{\sigma}^2_{BD}$ is directly related to the chi-square distribution with $m + df$ degrees of freedom, that is, $(df + m)\hat{\sigma}^2_{BD}/\sigma^2 \sim \chi^2_{m+df}$.

An asymptotic $(1-\alpha)\times 100\%$ confidence interval on θ can also be provided by the SAS procedure NLMIXED and it is given by $\hat{\theta}_{BD} \pm t_{m-1,\alpha/2}\hat{SE}(\hat{\theta}_{BD})$, with $t_{d,q}$ the qth upper quantile of the t-distribution with d degrees of freedom, and $\hat{SE}(\hat{\theta}_{BD})$ the estimated asymptotic standard error of the estimator $\hat{\theta}_{BD}$ (SAS Institute, 1996). SAS uses the number of random effects minus one ($m-1$) as the default number of degrees of freedom when a RANDOM statement is used, but it requires the DF option to use the appropriate number when the marginal likelihood in (16) is implemented in NLMIXED (see the Appendix) to obtain the correct degrees of freedom $m-1$.

We do realize that the proposed bivariate method requires more input than the other described methods, since the number of degrees of freedom df_i associated with the within-study variance estimate S_i^2 is required in our approach. However, in practice we expect that meta-analysts may have access to this information or otherwise can calculate or create the appropriate degrees of freedom for S_i^2. For instance, for a mean difference the degrees of freedom would become $n_{i0} + n_{i1} - 2$ for equal variances or equal to the Satterthwaite degrees of freedom for unequal variances (see Section 3), with n_{i0} and n_{i1} the sample sizes of the control and exposed group in study i, respectively. Pearson's correlation coefficient is typically associated with $n_i - 2$ degrees of freedom (Fisher, 1915), while the degrees of freedom for a meta-analysis of regression parameters of linear regression analyses, is equal to the degrees of freedom $n_i - p_i$ for the residual variance, with n_i the number of participants in study i and $p_i - 1$ the number of confounders used in the linear regression analysis.

2.4 | Case studies from the literature

To illustrate the approaches, we applied them to two different meta-analyses with different types of effect sizes. One meta-analysis studies the effect of coronary artery disease (CAD) on the mean platelet volume (MPV) using mean differences (Sansanayudh et al., 2014), while the other meta-analysis studies Spearman's correlation coefficient for correlation between apparent diffusion coefficient (ADC) and tumor cellularity (TC) in patients (Chen et al., 2013).

2.4.1 | Meta analysis on CAD and MPV

One of the aims of Sansanayudh et al. (2014) was to conduct a systematic review and meta-analysis comparing mean differences in MPV between patients (CAD) and controls. Forty studies were included in this meta-analysis based on the authors' eligibility criteria, but only 31 studies compared the mean MPV between CAD patients and controls. They applied DL approach to the mean differences and reported an overall mean difference of 0.70 (0.55; 0.85).

We used the data of these 31 studies that were presented in fig. 2 of Sansanayudh et al. (2014). For these studies, we extracted the means, standard deviations, and sample sizes for patients and controls and calculated the mean difference Y_i, the standard error S_i, and the accompanying degrees of freedom df_i according to Section 3. These values are provided in Table 1.

The overall mean difference with their 95% confidence intervals and the estimate for the between-study variance τ^2 for our seven approaches are presented in Table 2. The estimates from HT, NB, and GS are all equal, since they are the maximum likelihood estimates for likelihood function (4). The estimates for DL and HKSJ are also equal since they both use the weighted average in (3). They only differ in the calculation of confidence intervals. DL has the smallest pooled estimate, the smallest estimate for τ^2, and the narrowest 95% confidence interval. The results match the results reported by Sansanayudh et al. (2014), but they reported a somewhat smaller confidence interval because they may have used the normal quantile. The HKSJ clearly enlarges the confidence interval due to a correction factor that is equal to 1.195. The

TABLE 1 Overview of the effect sizes, standard errors, and degrees of freedom for the association of MPV with CAD

Study	Y_i	S_i	df_i	Study	Y_i	S_i	df_i
Cameron (1983)	0.75	0.106	239	Lippi (2009)	0.57	0.028	644
Trowbridge (1984)	0.80	0.163	10.07	Senen (2010)	0.05	0.110	276
Glud (1986)	−0.08	0.390	38.6	Tavil (2010)	1.41	0.150	256
Erne (1988)	0.40	0.177	78.3	Pawlus (2010)	1.24	0.136	97.9
Hendra (1988)	0.70	0.131	276	Jurcut (2010)	0.55	0.042	137
Mcgill (1994)	0.52	0.161	92.2	Ulusoy (2011)	0.20	0.130	117
Halbmayer (1995)	0.10	0.094	222	Chu (2011)	1.00	0.116	105
Pizzulli (1998)	1.10	0.104	132	Cemin (2011)	−0.02	0.043	158
Senaran (2001)	1.35	0.180	33.6	Assiri (2012)	0.73	0.116	202
Kilichli-Camur (2005)	0.64	0.120	153	Kunicki (2012)	0.35	0.199	235
Khandekar (2006)	0.70	0.184	43.2	Khode (2012)	0.31	0.131	110
Ihara (2006)	−0.01	0.117	193	Lopez-Cuence (2012)	1.80	0.523	308
Boos (2008)	1.00	0.199	19.0	Ozkan (2012)	0.65	0.076	337
Ranjith (2009)	1.38	0.091	88.7	Mizaie (2012)	0.35	0.198	234
Varol (2009)	1.00	0.231	108	Ozlu (2013)	0.71	0.168	122
Sen (2009)	2.13	0.222	165				

TABLE 2 Combined estimate of mean difference of MPV, along with 95% confidence limits and the between-study variance estimate

Method	$\hat{\theta}$ with 95% Confidence limits	$\hat{\tau}^2$
DL	0.699 (0.544; 0.854)	0.1532
HKSJ	0.699 (0.513; 0.885)	0.1532
HT	0.703 (0.526; 0.883)	0.2137
NB	0.703 (0.520; 0.890)	0.2137
GS	0.703 (0.522; 0.890)	0.2137
REML	0.704 (0.520; 0.887)	0.2223
BD	0.713 (0.510; 0.917)	0.2339

HT pooled estimate is slightly larger than the DL estimate, but has an almost 40% higher estimate for the between-study variance. The 95% confidence intervals for NB and GS are slightly wider than the 95% confidence interval of HT, which is the intention of these two methods. The pooled estimate of REML is very close to the ML estimate, but has a confidence interval close to the two higher-order profile likelihoods (NB and GS). The pooled estimate of BD is the highest of all pooled estimates and the 95% confidence interval is the widest of all 95% confidence intervals. The estimate of the between-study variance of BD is also slightly larger than the other between-study estimates.

2.4.2 | Meta-analysis on ADC and TC

The ADC is a measure of the magnitude of diffusion of water molecules within tissues. It can be calculated from MRI's with diffusion-weighted imaging. TC affects the diffusion of water in tissue and in vitro and animal studies demonstrate that TC is inversely correlated with ADC. Chen et al. (2013) conducted a meta-analysis to explore the correlation between TC and ADC. The authors selected 30 studies from 189 papers and collected or calculated Spearman's correlation coefficient for each study (Table 1 in Chen et al., 2013). The authors applied the DL approach on the Fisher z transformed correlation coefficients without mentioning how the standard error was calculated. They reported a pooled correlation coefficient of −0.57 (−0.62; −0.52).

We extracted the sample sizes (n_i) and Spearman correlation coefficients (R_i) from their paper and calculated the Fisher z transformation as the effect size $Y_i = 0.5[\log(1 + R_i) - \log(1 - R_i)]$. We calculated two different standard errors for the Fisher z transformed Spearman correlation coefficient: $S_i^2(\text{FP}) = 1.06/[n_i - 3]$ suggested by Fieller and Pearson (1961) and $S_i^2(\text{BW}) = [1 + 0.5R_i^2]/[n_i - 3]$ suggested by Bonett and Wright (2000). Although alternative calculations for the standard

TABLE 3 Overview of the effect sizes, standard errors, and degrees of freedom for the correlation of TC with ADC

Study	Y_i	S_i(FP)	S_i(BW)	df_i	Study	Y_i	S_i(FP)	S_i(BW)	df_i
Suguhara (1999)	−0.973	0.250	0.275	18	Yoshikawa (2008)	0.050	0.210	0.204	25
Gupta (2000)	−0.775	0.266	0.284	16	Woodhams (2009)	−0.950	0.297	0.326	13
Gauvai (2001)	−0.811	0.343	0.369	10	Wang (2009)	−0.741	0.179	0.191	34
Kono (2001a)	−0.973	0.275	0.303	15	Yamashita (2009)	−0.848	0.215	0.232	24
Kono (2001b)	−0.775	0.266	0.284	16	Gibbs (2009)	−0.829	0.250	0.269	18
Guo A (2002)	−0.497	0.206	0.210	26	Kikuchi (2009)	−0.793	0.389	0.417	8
Guo Y (2002)	−0.563	0.155	0.160	45	Jenkinson (2010)	0.040	0.275	0.267	15
Chen (2005)	−0.576	0.185	0.191	32	Ellingson (2010)	−1.376	0.275	0.315	15
Hayashida (2006)	−0.829	0.326	0.351	11	Barajas (2010)	−0.576	0.266	0.275	16
Plank (2007)	−0.758	0.460	0.491	6	Kyriazi (2010a)	−1.020	0.460	0.509	6
Matoba (2007)	−0.973	0.420	0.462	7	Kyriazi (2010b)	−0.908	0.515	0.561	5
Humphries (2007)	−0.908	0.257	0.281	17	Wang (2011)	−0.365	0.266	0.266	16
Zelhof (2008)	−0.523	0.174	0.179	36	Goyal (2011)	−0.321	0.179	0.178	34
Hatakenaka (2008)	−0.775	0.094	0.100	122	Doskaliyev (2012)	−0.662	0.225	0.236	22
Manenti (2008)	−0.887	0.210	0.228	25	Ginat (2012)	−0.662	0.266	0.279	16

TABLE 4 Pooled estimates of Spearman's correlation coefficients with their 95% confidence limits and between-study variance estimate

Method	Fieller and Pearson		Bonett and Wright	
	$\hat{\theta}$ with 95% Confidence limits	$\hat{\tau}^2$	$\hat{\theta}$ with 95% Confidence limits	$\hat{\tau}^2$
DL	−0.590 (-0.656; -0.515)	0.0240	−0.579 (-0.648; -0.501)	0.0222
HKSJ	−0.590 (-0.656; -0.515)	0.0240	−0.579 (-0.648; -0.501)	0.0222
HT	−0.590 (-0.659; -0.517)	0.0251	−0.580 (-0.653; -0.503)	0.0249
NB	−0.590 (-0.662; -0.513)	0.0251	−0.580 (-0.656; -0.499)	0.0249
GS	−0.590 (-0.664; -0.514)	0.0251	−0.580 (-0.658; -0.499)	0.0249
REML	−0.591 (-0.659; 0.513)	0.0285	−0.581 (-0.652; -0.499)	0.0283
BD	−0.592 (-0.658; -0.516)	0.0253	−0.589 (-0.656; -0.515)	0.0190

error exist (Caruso & Cliff, 1997), our choices for S_i are not or (expected to be) mildly correlated with the effect size Y_i. Since Spearman's coefficient is the rank-based Pearson's correlation coefficient, we choose $df_i = n_i - 2$, because the degrees of freedom for Pearson's correlation coefficient is $n_i - 2$. An overview of the studies, effect sizes, standard errors, and degrees of freedom are presented in Table 3.

All seven approaches have been applied to the data in Table 3 for both standard errors. The pooled estimator and its 95% confidence interval are then transformed back to the original scale of correlation. The results for all seven methods are reported in Table 4 for both standard errors. The between-study variance was obviously obtained in the Fisher z transformed scale.

When we use the Fieller and Pearson standard error we see that all methods are almost identical. This is to be expected since the standard errors are not random. Only the two finite-sample corrected profile likelihoods (NB and GS) have a slightly larger interval. When we use the Bonett and Wright standard error, we see that the pooled estimate of BD is the highest and the pooled estimate of DL is the smallest (although the differences are small). However, the 95% confidence interval for BD is now the smallest and it produced the smallest between-study variance (although differences are quite small). It is interesting to notice that the choice of calculation of the standard error has no impact on the pooled estimate of the BD approach, but it does affect the other approaches to some extent.

3 | SIMULATION MODEL

We use a heteroskedastic mixed effects model to generate individual participant data (IPD). The IPD is then used to calculate AD: an effect size Y_i, a standard error S_i, and its associated degrees of freedom df_i. The effect sizes are then pooled using the methods described in Section 2. Different settings for the IPD model parameters were selected, including simulations that allow for interaction between-study design parameters (i.e., the dependency between Y_i and S_i). A number of 1000 simulation runs were generated for each setting. For each simulation run, the parameter θ is estimated and accompanied with a 95% confidence interval using the methods described in Section 2. We present the bias, mean squared error (MSE), and the coverage probability for the main parameter θ. The Monte Carlo standard error for estimation of a coverage probability of 95% is equal to 0.69%.

3.1 | Individual participant data

We simulated IPD for m studies. The sample size n_i for study $i = 1, \dots, m$ varied from study to study. This sample size was drawn from an overdispersed Poisson distribution, that is, $n_i|\gamma_i \sim \text{Poi}(\lambda \exp\{0.5\gamma_i\})$, with $\gamma_i \sim \Gamma(a_0, b_0)$ drawn from a gamma distribution. Then within each study the participants are randomly allocated to two groups (e.g., treatments) with probabilities p and $1 - p$, resulting in n_{i0} participants in the control group (i.e., $n_{i0}|n_i \sim \text{Bin}(n_i, p)$) and $n_{i1} = n_i - n_{i0}$ participants in the exposed group. A continuous response Y_{ijk} for individual k $(= 1, \dots, n_{ij})$, in group j $(= 0, 1)$, of study i is then simulated according to a heteroskedastic linear mixed effects model (Davidian & Carroll, 1987; Quintero & Lesaffre, 2017):

$$Y_{ijk} = \mu_j + U_{ij} + \xi_j \exp(V_i)\epsilon_{ijk}, \tag{19}$$

with μ_j the mean of group j, U_{ij} a study-specific random effect for group j, ξ_j^2 a group-specific residual variance parameter, V_i a random effect for residual heteroskedasticity across studies, and $\epsilon_{ijk} \sim N(0, 1)$ standard normally distributed and independent of random effects U_{i0}, U_{i1}, and V_i. It is assumed that $(U_{i0}, U_{i1}, V_i)^T$ has a multivariate normal distribution with means 0 and a variance–covariance matrix Σ given by

$$\Sigma = \begin{pmatrix} \upsilon_0^2 & \rho_M \upsilon_0 \upsilon_1 & \rho_V \upsilon_0 \upsilon_2 \\ \rho_M \upsilon_0 \upsilon_1 & \upsilon_1^2 & \rho_V \upsilon_1 \upsilon_2 \\ \rho_V \upsilon_0 \upsilon_2 & \rho_V \upsilon_1 \upsilon_2 & \upsilon_2^2 \end{pmatrix}. \tag{20}$$

The value of ρ_M represents the correlation between the study-specific random effects U_{i0} and U_{i1} for the exposed and the control group, respectively. The value ρ_V represents the correlation between the study mean and the logarithm of the random heteroskedastic residual variance.

Note that there are two forms of residual heteroskedasticity in the IPD model (19). One is at the level of the participant and introduced via parameter ξ_j^2 and the other one is at the level of the study introduced via the random term $\exp(V_i)$. The variance ξ_j^2 indicates a fixed heteroskedasticity in variability between individuals for the two groups (i.e., the group affects both the level and the variability) that is consistent across studies, while $\exp(V_i)$ indicates random heteroskedasticity across studies that is consistent within studies (i.e., individuals are more or less alike within studies). This random heteroskedasticity will be referred to as the heteroskedasticity of within-study variances for AD meta-analyses (see also next section).

3.2 | Aggregated data

Based on the IPD of model (19), we can calculate the required study information for an AD meta-analysis. The observed effect size aggregated at the study level is given by the raw mean difference $Y_i = \bar{Y}_{i0.} - \bar{Y}_{i1.}$ for study i, where $\bar{Y}_{ij.} = \sum_{k=1}^{n_{ij}} Y_{ijk}/n_{ij}$ is the average value for group j in study i. The estimated standard error S_i for the effect size Y_i is given by $S_i^2 = S_{i0}^2/n_{i0} + S_{i1}^2/n_{i1}$, with $S_{ij}^2 = \sum_{k=1}^{n_{ij}}(Y_{ijk} - \bar{Y}_{ij.})^2/(n_{ij} - 1)$ the sample variance for group j in study i. The calculation of this standard error does not assume homoskedastic variances between the treatment and control group. The corresponding degrees of freedom df_i for S_i^2 can then be determined by Satterthwaite approach (Satterthwaite, 1946)

$$df_i = S_i^4/[S_{i0}^4/(n_{i0}^2(n_{i0} - 1)) + S_{i1}^4/(n_{i1}^2(n_{i1} - 1))]. \tag{21}$$

Using model (19), the observed effect size Y_i can be written in the form of the well-known random effects model for meta-analysis studies (Brockwell & Gordon, 2007) given in (1), with $\theta = \mu_0 - \mu_1$ the overall mean difference, $U_i \equiv U_{i0} - U_{i1}$ the effect size heterogeneity, $\varepsilon_i = \exp(V_i)(\xi_0\bar{\varepsilon}_{i0.} - \xi_1\bar{\varepsilon}_{i1.})$ the within-study residual, with $\bar{\varepsilon}_{ij.} = \sum_{k=1}^{n_{ij}} \epsilon_{ijk}/n_{ij}$. The usual distributional assumptions of normality of Y_i and independence of U_i and ε_i in model (1) are only met when the random heteroskedasticity V_i does not exist. Indeed, when $V_i = 0$, the terms U_i and ε_i are independent and normally distributed, that is, $Y_i \sim N(\theta, \upsilon_0^2 - 2\rho_M\upsilon_0\upsilon_1 + \upsilon_1^2 + \xi_0^2/n_{i0} + \xi_1^2/n_{i1})$. But the presence of V_i makes the distribution of the residuals (and thus the effect size Y_i) nonnormal and creates a dependence between U_i and ε_i when $\rho_V \neq 0$ (van den Heuvel et al., 2021). The distribution of the effect size Y_i becomes untraceable when V_i exists. Furthermore, heterogeneity of the effect sizes is present for all settings of $\rho_M < 1, \upsilon_0 > 0$, and $\upsilon_1 > 0$, but it vanishes when $\rho_M = 1$ and $\upsilon_0 = \upsilon_1$. Finally, without the presence of heteroskedastic within-study variances V_i, the residuals ε_i are still heteroskedastic across studies ($\mathsf{VAR}(\varepsilon_i) = \xi_0^2/n_{i0} + \xi_1^2/n_{i1}$) due to different sample sizes across studies, but the existence of V_i makes the standard errors of the effect size (i.e., the residuals ε_i in (1)) heteroskedastic even when all studies would have the exact same sample size (see also Cochran, 1954).

Again using model (19), the variance S_i^2 can be rewritten into

$$S_i^2 = \exp(2V_i)(\xi_0^2 s_{i0}^2/n_{i0} + \xi_1^2 s_{i1}^2/n_{i1}) \tag{22}$$

with $(n_{ij} - 1)s_{ij}^2 = \sum_{k=1}^{n_{ij}}(\epsilon_{ijk} - \bar{\epsilon}_{ij.})^2$ chi-square distributed with $n_{ij} - 1$ degrees of freedom. Thus the presence of term V_i in (22), shows that the *observed* standard errors in the meta-analysis are also heteroskedastic, and not just the residuals ε_i for the effect size Y_i in model (1). The conditional distribution of S_i^2 given V_i is approximately chi-square distributed using Satterthwaite approach (Satterthwaite, 1946), that is, $df_i S_i^2/[\exp\{V_i\}(\xi_0^2/n_{i0} + \xi_1^2/n_{i1})]$ is approximately chi-square distributed with df_i degrees of freedom when conditioned on V_i. The conditional distribution becomes exactly chi-square distributed with $n_{i0} + n_{i1} - 2$ degrees of freedom when both $\xi_0 = \xi_1$ and $n_{i0} = n_{i1}$ hold. However, the term V_i makes the marginal distribution of S_i^2 less traceable and possibly different from a chi-square distribution.

Finally, the random heteroskedasticity V_i causes a dependency between Y_i and S_i^2, even when there is no heterogeneity ($U_i = 0$), because V_i is present in both Y_i and S_i^2. However, the variance of Y_i given S_i^2 is unequal to S_i^2, since the conditional distributions of Y_i and S_i^2 given V_i are independent (van den Heuvel et al., 2021). Nevertheless, it can be demonstrated that the expectation of S_i^2 is equal to the variance of Y_i when the heterogeneity vanishes ($U_i = 0$), making S_i^2 an unbiased and appropriate estimator for the residual variance $\mathsf{VAR}(\varepsilon_i) = \exp\{\upsilon_2^2/2\}[\xi_0^2/n_{i0} + \xi_1^2/n_{i1}]$. The possible correlation between U_i and V_i will affect the dependency between the effect size Y_i and standard error S_i that is invoked by V_i alone. It can be demonstrated that the covariance of Y_i and S_i^2 is equal to $\mathsf{COV}(Y_i, S_i^2) = \upsilon_2\rho_V[\upsilon_0 - \upsilon_1]\exp\{\upsilon_2^2/2\}[\xi_0^2/n_{i0} + \xi_1^2/n_{i1}]$ (van den Heuvel et al., 2021). Thus $\rho_V = 0$ or $\upsilon_0 = \upsilon_1$ make the effect size Y_i and variance S_i^2 uncorrelated (but not independent), while $\rho_V \neq 0$ and $\upsilon_0 \neq \upsilon_1$ introduces a correlation between Y_i and S_i^2.

We believe that all proposed estimation methods for θ with their different confidence intervals in Section 2 are at best approximate methods for modeling the AD of our simulated meta-analysis. Moreover, none of the methods has an obvious direct advantage over any of the other methods a priori, since none of the methods incorporated the dependency between Y_i and S_i or used a nonnormal distribution for Y_i.

3.3 | Simulation settings

The settings of the parameters are not based on any real case study, but they are chosen such that the simulation may potentially correspond with a meta-analysis of clinical trials on hypertension treatment for lowering systolic blood pressure. Parameter settings used to generate the IPD are $m \in \{5, 10, 20, 30\}$, $\lambda = 100$, $a_0 = b_0 = 1$, $p = 0.5$, $\mu = 160$, $\theta = -2$, $\xi_0^2 = \xi_1^2 = 100$. We will run several combinations of the remaining parameters $\upsilon_0^2, \upsilon_1^2, \upsilon_2^2, \rho_M$, and ρ_V of the IPD model:

1. **Setting 1**: Homogeneous effect sizes and no heteroskedastic within-study variances: $\upsilon_0^2 = 0$, $\upsilon_1^2 = 0$, $\upsilon_2^2 = 0$, $\rho_M = 0$, and $\rho_V = 0$,
2. **Setting 2**: Heterogeneous effect sizes and no heteroskedastic within-study variances: $\upsilon_0^2 = 2$, $\upsilon_1^2 = 3$, $\upsilon_2^2 = 0$, $\rho_M = 0.7$, and $\rho_V = 0$,
3. **Setting 3**: Heterogeneous effect sizes and heteroskedastic within-study variances without correlation: $\upsilon_0^2 = 2$, $\upsilon_1^2 = 3$, $\upsilon_2^2 = 1$, $\rho_M = 0.7$, and $\rho_V = 0$,
4. **Setting 4**: Heterogeneous effect sizes and heteroskedastic within-study variances with low correlation: $\upsilon_0^2 = 2$, $\upsilon_1^2 = 3$, $\upsilon_2^2 = 1$, $\rho_M = 0.7$, and $\rho_V = 0.3$,

TABLE 5 Bias of the pooled estimators for the overall effect size under different simulation settings and for $\theta = -2$

Setting	$m = 5$				$m = 10$			
	DL	**HT**	**REML**	**BD**	**DL**	**HT**	**REML**	**BD**
1	−0.004	−0.006	−0.003	−0.005	−0.024	−0.024	−0.022	−0.024
2	−0.003	−0.003	−0.003	−0.002	−0.006	−0.009	−0.007	−0.007
3	0.006	−0.000	0.004	0.013	0.007	0.010	0.009	−0.010
4	−0.020	−0.032	−0.022	0.006	−0.034	−0.035	−0.031	−0.007
5	−0.038	−0.052	−0.040	0.001	−0.062	−0.066	−0.060	−0.005
6	−0.057	−0.076	−0.059	−0.005	−0.091	−0.099	−0.090	−0.005
Setting	$m = 20$				$m = 30$			
	DL	**HT**	**REML**	**BD**	**DL**	**HT**	**REML**	**BD**
1	−0.021	−0.021	−0.020	−0.020	−0.010	−0.010	−0.010	−0.011
2	−0.013	−0.014	−0.014	−0.012	−0.004	−0.004	−0.004	−0.004
3	−0.001	0.000	−0.001	−0.019	−0.001	−0.001	−0.001	−0.006
4	−0.043	−0.044	−0.043	−0.018	−0.045	−0.046	−0.045	−0.007
5	−0.073	−0.075	−0.072	−0.017	−0.075	−0.077	−0.076	−0.007
6	−0.103	−0.107	−0.102	−0.015	−0.106	−0.109	−0.106	−0.006

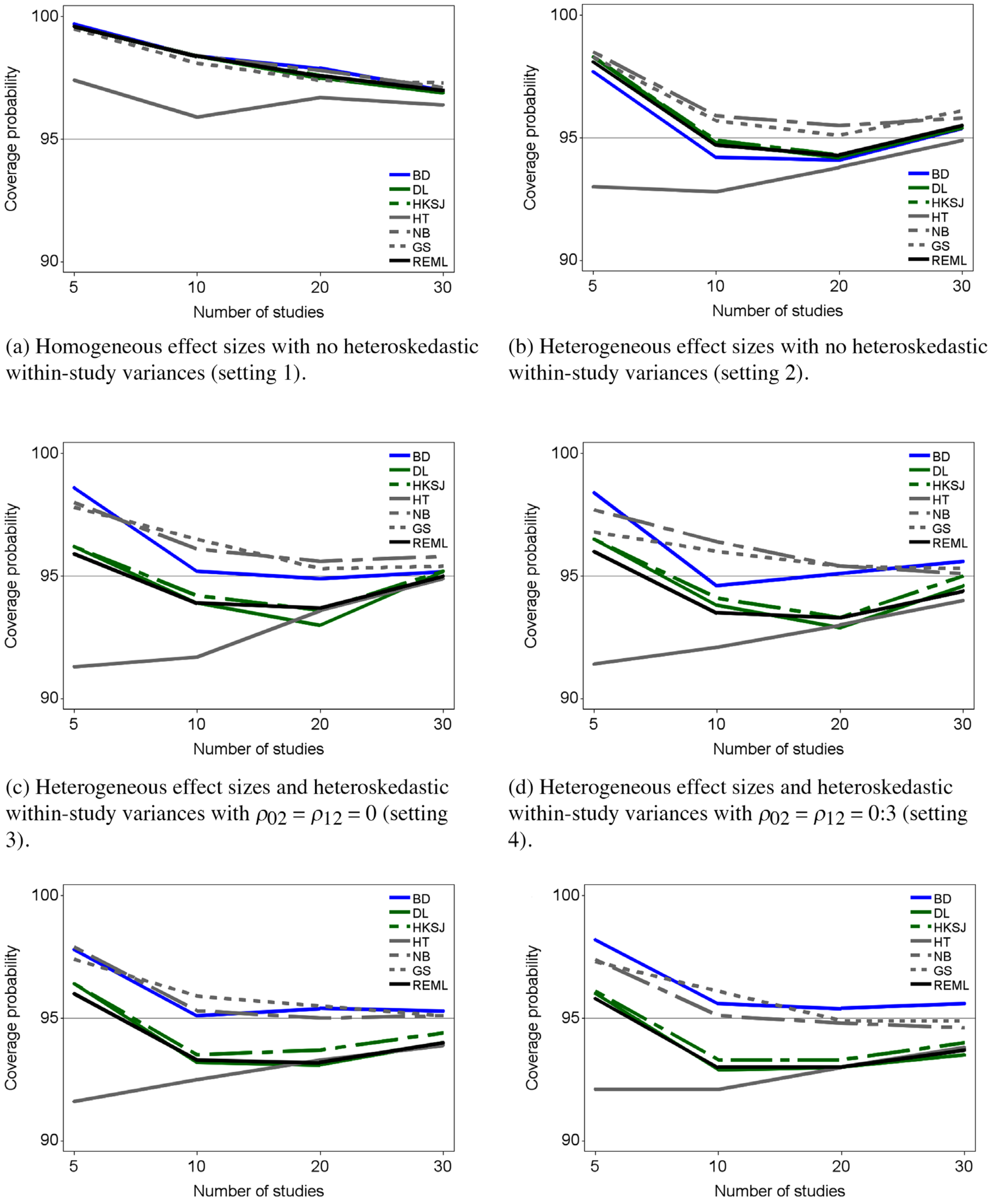
TABLE 6 MSE of the pooled estimators for the overall effect size under different simulation settings and for $\theta = -2$

Setting	$m = 5$				$m = 10$			
	DL	**HT**	**REML**	**BD**	**DL**	**HT**	**REML**	**BD**
1	0.505	0.501	0.504	0.496	0.239	0.238	0.240	0.233
2	0.878	0.876	0.878	0.872	0.435	0.436	0.434	0.429
3	0.930	0.936	0.932	1.213	0.434	0.439	0.437	0.589
4	0.918	0.927	0.926	1.204	0.431	0.435	0.432	0.602
5	0.909	0.919	0.919	1.207	0.432	0.436	0.433	0.610
6	0.904	0.912	0.915	1.209	0.436	0.441	0.436	0.616
Setting	$m = 20$				$m = 30$			
	DL	**HT**	**REML**	**BD**	**DL**	**HT**	**REML**	**BD**
1	0.105	0.104	0.105	0.102	0.069	0.069	0.069	0.068
2	0.197	0.197	0.197	0.197	0.128	0.128	0.127	0.127
3	0.202	0.203	0.202	0.282	0.127	0.127	0.126	0.181
4	0.202	0.204	0.202	0.277	0.126	0.127	0.126	0.182
5	0.204	0.206	0.204	0.273	0.128	0.129	0.128	0.182
6	0.208	0.210	0.208	0.266	0.132	0.133	0.132	0.183

5. **Setting 5**: Heterogeneous effect sizes and heteroskedastic within-study variances with medium correlation: $\upsilon_0^2 = 2$, $\upsilon_1^2 = 3$, $\upsilon_2^2 = 1$, $\rho_M = 0.7$, and $\rho_V = 0.5$,
6. **Setting 6**: Heterogeneous effect sizes and heteroskedastic within-study variances with high correlation: $\upsilon_0^2 = 2$, $\upsilon_1^2 = 3$, $\upsilon_2^2 = 1$, $\rho_M = 0.7$, and $\rho_V = 0.7$.

4 | RESULTS

First we discuss the bias and the MSE of the pooled estimators for the overall effect size for the four different estimation methods: DL, HT, REML and our BD, respectively. The results are presented in Tables 5 and 6. Recall that the NB and GS confidence intervals make use of the maximum likelihood estimators of Hardy and Thompson and that the HKSJ method makes use of the DL estimators. Then, we study the coverage probabilities of all seven approaches of confidence intervals on the overall effect size. The results are presented in Figure 1a–f. Finally, we discuss the estimation of the between-study variance (τ^2) for the four estimation methods. The results are listed in Table 7.



(a) Homogeneous effect sizes with no heteroskedastic within-study variances (setting 1).

(b) Heterogeneous effect sizes with no heteroskedastic within-study variances (setting 2).

(c) Heterogeneous effect sizes and heteroskedastic within-study variances with $\rho_{02} = \rho_{12} = 0$ (setting 3).

(d) Heterogeneous effect sizes and heteroskedastic within-study variances with $\rho_{02} = \rho_{12} = 0{:}3$ (setting 4).

(e) Heterogeneouseffect sizes and heteroskedastic within-study variances with $\rho_{02} = \rho_{12} = 0{:}5$ (setting 5).

(f) Heterogeneous effect sizes and heteroskedastic within-study variances with $\rho_{02} = \rho_{12} = 0{:}7$ (setting 6).

FIGURE 1 Coverage probabilities of 95% confidence intervals of seven methods for the overall effect size under different settings and study sizes

4.1 | Estimation of the overall effect size

For the settings without heteroskedastic within-study variances (settings 1 and 2) the biases of DL, HT, REML, and BD are all similar. Irrespective of the sample size, biases remain within 1.2% of the true effect size ($\theta = -2$) for the homogeneous effect sizes and are negligible for the heterogeneous effect sizes. In the presence of uncorrelated heterogeneous effect

TABLE 7 Between-study variances of the three estimation methods under different simulation settings

Setting	τ^2	$m = 5$				$m = 10$			
		DL	HT	REML	BD	DL	HT	REML	BD
1	0	0.664	0.346	0.680	0.329	0.436	0.281	0.433	0.253
2	1.5707	1.920	1.221	1.933	1.195	1.779	1.404	1.793	1.355
3	~1.5707	1.984	1.221	2.067	1.670	1.728	1.319	1.728	1.723
4	~1.5707	1.993	1.214	2.075	1.726	1.725	1.324	1.741	1.730
5	~1.5707	1.989	1.199	2.065	1.779	1.723	1.331	1.746	1.733
6	~1.5707	1.972	1.168	2.036	1.838	1.722	1.332	1.749	1.730

Setting	τ^2	$m = 20$				$m = 30$			
		DL	HT	REML	BD	DL	HT	REML	BD
1	0	0.296	0.207	0.277	0.181	0.240	0.177	0.222	0.154
2	1.5707	1.601	1.410	1.605	1.362	1.589	1.468	1.602	1.417
3	~1.5707	1.560	1.378	1.582	1.669	1.569	1.448	1.582	1.649
4	~1.5707	1.564	1.379	1.581	1.673	1.575	1.445	1.575	1.615
5	~1.5707	1.565	1.379	1.579	1.672	1.576	1.441	1.574	1.600
6	~1.5707	1.563	1.375	1.577	1.650	1.575	1.439	1.572	1.585

sizes and heteroskedastic within-study variances (setting 3), again all biases are very close to zero for all four sample sizes. However, in the case of correlated heterogeneous effect sizes and heteroskedastic within-study variances (settings 4–6), only BD seems to have small biases for all sample sizes and it is never larger than 1.0% of the true effect size. The biases of DL, HT, and REML are away from zero, in particular when the heterogeneous effect sizes are strongly correlated with the heteroskedastic within-study variances. The sample size does not seem to affect this, although the bias is somewhat smaller for meta-analyses with five studies. The bias can reach a level of more than 5% of the true effect size.

The performance of MSE for the four estimation methods is very consistent across all settings. For all methods, the MSE increases with settings, which is expected due to the increased variability. Setting 1 has no study heterogeneity and no heteroskedastic within-study variances, and thus the smallest variability across meta-analysis studies. Setting 2 has heterogeneous effect sizes but no heteroskedastic within-study variances. Then for settings 3–6, the residual variance increases due to the heteroskedastic within-study variances and an increased positive correlation ρ_V, while the heterogeneity in effect sizes remains constant (although the correlation seems to have little effect). When no heteroskedastic within-study variances are present, the MSE of the four estimation approaches DL, HT, REML, and BD are almost identical. However, when heteroskedastic within-study variances are present, the MSE of BD is larger than the MSE of DL, HT, and REML, due to a choice of study weights that is not maximizing the precision of the pooled effect size anymore. Indeed, the study weight w_i that would maximizes precision of the pooled effect size in case of heteroskedastic within-study variances is $[\tau^2 + \sigma_i^2]^{-1} / \sum_{i=1}^{m} [\tau^2 + \sigma_i^2]^{-1}$, while the BD method is applying the weights $[\tau^2 + \sigma^2/df_i]^{-1} / \sum_{i=1}^{m} [\tau^2 + \sigma^2/df_i]^{-1}$ with an estimator $\hat{\sigma}^2$ for σ^2. Thus the BD method may lose some precision with respect to a pooled effect size using the optimal weights, but the standard error of $\hat{\theta}_{BD}$ also includes the uncertainty of estimating σ^2, since the BD approach jointly estimates the three parameters θ, τ^2, and σ^2. The MSE of DL, HT, and REML seem to be identical across all settings and sample sizes. It seems that the random heteroskedasticity does hardly affect the MSE of DL, HT, and REML, since it is (almost) at the same level as setting 2 which had no heteroskedastic within-study variances. These methods make use of the optimal weights $[\tau^2 + S_i^2]^{-1} / \sum_{i=1}^{m} [\tau^2 + S_i^2]^{-1}$ for a pooled effect size with maximum precision, but they ignore the uncertainty of having to estimate the heteroskedastic variances σ_i^2 and may therefore be at risk of a too small estimated standard error for the pooled effect size.

4.2 | Coverage probabilities for the overall effect size

Considering the coverage probabilities in Figure 1a–f, we see conservative coverage probabilities for the (unrealistic) case of homogeneous effect sizes and no heteroskedastic within-study variances (setting 1), although the HT method seems closer to nominal than the others. The conservative coverage probabilities are explained by the incorrect use of the degrees of freedom in the t-quantile. The estimates for the between-study variance frequently vanishes, which would imply that the degrees of freedom for the estimated standard error of the pooled effect size is much closer to the sum of all the degrees of freedom in the meta-analysis study (Cochran, 1954; Mzolo et al., 2013). In that case, a normal quantile is warranted or one could update the degrees of freedom in the t-quantile. When the number of studies increases we see that the coverage probability for all methods converges to nominal. We see the same phenomena for the setting with heterogeneous effect sizes (setting 2) when the number of studies is small. The between-study variance cannot be estimated reliably (see tab. 1 of McNeish & Stapleton, 2016) and becomes zero too frequent. It is somewhat surprising that the profile likelihood confidence interval of HT provides liberal coverage probabilities in this setting. All methods seem to become closer to nominal for setting 2 when the number of studies is increasing.

In case heteroskedastic within-study variances are introduced, the DL, HKSJ, HT, and REML methods seem to underperform and provide liberal coverage probabilities when the number of studies is equal to 10 and 20. The HKSJ is doing slightly better than the DL and REML method, but the difference is negligible. The HT approach is particularly bad for a small meta-analysis with a small number of studies. For the simulation setting of uncorrelated effect sizes and heteroskedasticity (setting 3), we believe that this lower coverage is caused by having to use S_i^2 for the true variance σ_i^2 in the study weights. The estimation uncertainty is not transferred to the standard error of the pooled effect size and the asymptotic standard errors do not work properly. The other methods GS, NB, and BD do take care of this issue of estimation uncertainty and show an improved coverage. When the number of studies then start to increase this issue of finite-sample uncertainty becomes less relevant and the methods DL, HKSJ, HT, and REML start to provide nominal coverages. Then these methods are clearly preferred over the BD method, since the BD method provides a lower precision that is induced by the nonoptimal study weights.

When the effect size and its standard error are becoming correlated (settings 4–6), a bias in the estimation of the overall effect size for the estimation methods DL, HT, and REML start to occur. This bias then worsens the coverage probability compared to the third setting (uncorrelated effect sizes and heteroskedasticity). This bias also occurs for meta-analyses with just five studies, but then the conservative coverage probability of DL, HKSJ, and REML from setting 2 is accidentally reduced to a coverage probability closer to nominal. This bias in estimating the pooled effect size is now also causing a somewhat lower coverage for higher numbers of study sizes, but the bias is not large enough to cause a concern in the coverage.

The dependency between the effect size and its standard error does not seem to affect the higher-order profile likelihood (NB and GS) and our bivariate approach (BD) for any of the four settings with heteroskedasticity. For 10 or more studies, the coverage probabilities are (very) close to nominal, but the GS method seems to be slightly and consistently conservative at $m = 10$ studies. These methods have a better bias-precision trade-off then the DL, HKSJ, HT, and REML methods, but BD is still a first-order approximation and it does as good (or better) as the higher-order profile likelihood approximations when the between-study variances can be reliably estimated. For meta-analyses with a very small number of studies, our approach can still be improved by implementing the appropriate degrees of freedom when the between-study variance is estimated zero. This would also be true for the other conservative methods (Mzolo et al., 2013).

4.3 | Estimation of the between-study variance

To complete the comparison, we also compared the estimates of the between-study variance τ^2 for the four estimation methods DL, HT, REML, and BD. For the first setting the variance $\mathsf{VAR}(U_i) = \mathsf{VAR}(U_{i0} - U_{i1})$ is $\tau^2 = 0$ and in the remaining settings this variance is $\tau^2 = \sigma_0^2 + \sigma_1^2 - 2\rho_M\sigma_0\sigma_1 = 2 + 3 - 2 \times 0.7 \times \sqrt{2} \times \sqrt{3} \approx 1.5707$. However, in the case of heteroskedastic within-study variances, the correlation between the heterogeneous effect sizes U_i and the random heteroskedasticity V_i may affect the estimation of the between-study variance, but we expect it to be still close to 1.5707. The results of the estimates are presented in Table 7.

Without heterogeneity and heteroskedastic within-study variances, all methods are biased for the estimation of the between-study variance, but the BD method is closest to the truth and the bias reduces with the number of studies. In the case of heterogeneity, but without heteroskedastic within-study variances, the DL and REML approach are closest to the true value when sample sizes m are 20 or larger. For smaller number of studies, HT and BD are slightly closer to the true variance. DL and REML seem to overestimate the between-study variance, while HT and BD underestimates the variance. This latter observation is a well-known characteristic of maximum likelihood estimation for variance components (McCulloch & Searle, 2001). In the case of heterogeneity and heteroskedastic within-study variances (settings 3–6), we see that DL and REML give unbiased estimates of the between-study variance when the number of studies is 20 or more, with REML the most accurate one. The BD approach slightly overestimates, but the exact between-study variance for these four settings is not exactly known due to a correlation of the residual (ε_i) and the heterogeneity (U_i). For meta-analyses with only five studies, the DL and REML estimator both overestimate the between-study variance by more than 25%, while the BD estimator has a bias of not more than 17.5%. But for 10 studies, the differences between DL, REML, and BD vanishes. The bias reduces to approximately 10%. The HT method seems to be (substantially) biased in all settings.

5 | DISCUSSION

The purpose of this paper was to introduce a joint analysis of the effect sizes and their estimated standard errors for AD meta-analyses. A combination of a normal and chi-square distribution was used to describe the distribution of the observed bivariate statistics, following and extending the work of Cochran (1937). The performance of this BD was compared to that of the DL method (with and without the HKSJ correction) and four likelihood-based methods. The likelihood-based methods assumed that the residual variance of the effect size is equal to the squared standard error. We studied the profile likelihood approach of HT, the Bartlett-corrected likelihood ratio, the Skovgaard corrected likelihood ratio, and the REML approach. They were all illustrated on two real case studies. Furthermore, a simulation study with different scenarios was carried out using various numbers of studies and correlation structures between the effect sizes and its standard error. The simulation settings explicitly studied heteroskedastic within-study variances, because we believe that heteroskedasticity is common in practice. None of the seven studied approaches are theoretically equipped to deal with this form of heteroskedasticity explicitly.

Differences between the methods for estimation of the overall effect size with its confidence intervals were relatively small, but some differences were observed. When there is no heteroskedastic within-study variances and the number of studies is small, all methods demonstrate conservative coverage probabilities, except for HT in the case of heterogeneous effect sizes. The degrees of freedom of $m - 1$ is too small and should be closer to the total sum of the degrees of freedom in the meta-analysis when the between-study variance is estimated at zero. Estimation of variance components with just a few studies is unreliable (McNeish & Stapleton, 2016).

When heteroskedastic within-study variances are introduced, the DL, HT, and REML approaches show a small bias in the overall effect size except when the effect size is uncorrelated with the heteroskedasticity (simulation setting 3). This bias certainly contributes to a liberal coverage probability (for meta-analyses with 10 or more studies), but not addressing the uncertainty of estimating the study weights in the overall standard error also lowers the coverage when the number of studies is small, a conclusion already established in the literature (Guolo, 2012; Noma, 2011). In case we apply the HKSJ standard error estimate for the DL method, the coverage improves slightly, but it remains almost similar to the DL results for all settings and is not considered a solution.

Since the corrected likelihood approaches use a finite-sample approximation of the distribution of the HT estimator, these corrected approaches provide the same bias as the HT method, but they do improve the coverage probability when the number of studies is 10 or more. The Bartlett-type and the Skovgaard corrected likelihood ratio methods have comparable results with the HT finite-sample approach. They are conservative when the number of studies is small, but for larger study sizes they provide nominal coverage probabilities. These conclusions have been established earlier too (Noma, 2011; Veroniki et al., 2019).

More generally, all methods provide nominal coverages as the number of studies increases. Our bivariate approach provided similar and consistent results in all performance measures under heterogeneity and heteroskedastic within-study variances for meta-analyses with 10 or more studies, with coverage probabilities close to nominal. The coverage is very similar or better than the two finite-sample size corrected likelihood approaches and outperforms DL, HKSJ, HT approaches and REML for meta-analyses with 10–30 studies. However, for studies larger than and equal to 30, our approach is somewhat less precise than the other approaches when the effect sizes and heteroskedasticity are uncorrelated. For a smaller number of studies, our approach is comparable to these other methods.

A disadvantage of our approach is the need for a degrees of freedom. Therefore, our approach is currently limited to effect sizes from continuous clinical outcomes (e.g., functions of mean differences, regression parameters from linear regression analyses, and Fisher z-transformed correlation coefficients), since these measures typically come naturally with a degrees of freedom. Future research is needed to make our approach also suitable for effect sizes from contingency tables and survival analyses where no obvious degrees of freedom are present and where a dependency between the effect size and its standard error potentially needs a solution. One option is to approximate the distribution of the variance estimator S_i^2 for an effect size from a contingency table with a chi-square distribution, where the degrees of freedom is being estimated from the number of events and marginal totals, and model the dependency with Y_i with latent variables.

Our approach should then also be compared with alternative analysis approaches that would model the probability of an event directly.

The advantage of our approach is that the analysis is straightforward and based on first-order asymptotics that does not need a finite-sample correction. It also performed well when studies are heterogeneous in both the effect sizes and their standard errors and where independence between effect size and standard error was not guaranteed. The other analysis approaches studied in this paper, which condition on the standard error, led to a bias in the overall effect size (Böhning et al., 2002), while our method was not affected by the correlation between the effect size and its observed standard error. Our method is somewhat protected against such dependencies due to the proportionality assumption of the standard error and the study size implemented in the analysis phase, but could lose some precision by not using the study weights for optimal precision. In the presence of publication bias, where the effect size would depend on study size, and possibly for standardized mean differences (like Cohen's d), where a functional relationship between the effect size and standard error exists (Malzahn et al., 2000), our approach may fail and additional research is needed to investigate and address this issue. Generalizations of the joint likelihood of the effect sizes and their standard errors can be proposed for such issues (Jackson & White, 2018). Our approach is most beneficial for meta-analyses with small study sizes, since the uncertainty in the standard error is then most dominant.

ACKNOWLEDGMENTS

This research was funded by grant number 023.005.087 from the Netherlands Organization for Scientific Research. We would like to thank the four anonymous reviewers and associate editor for their very helpful suggestions in making our work better.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

ORCID

Osama Almalik https://orcid.org/0000-0001-6696-2286
Edwin R. van den Heuvel https://orcid.org/0000-0001-9157-7224

REFERENCES

Barndorff-Nielsen, O. E., & Hall, P. (1988). Bartlett adjustments to the likelihood ratio statistic and the distribution of the maximum likelihood estimator. *Biometrika*, *75*, 374–378.

Beath, K. (2016). metaplus: An R package for the analysis of robust meta-analysis and meta-regression. *R Journal*, *8*(1), 5–16.

Böhning, D., Malzahn, U., Dietz, E., Schlattmann, P., Viwatwongkasem, C., & Biggeri, A. (2002). Some general points in estimating heterogeneity variance with the DerSimonian–Laird estimator. *Biostatistics*, *3*(4), 445–457.

Bonett, D. G., & Wright, T. A. (2000). Sample size requirements for estimating Pearson, Kendall and Spearman correlations. *Psychometrika*, *65*(1), 23–28.

Brockwell, S. E., & Gordon, I. R. (2007). A simple method for inference on an overall effect in meta-analysis. *Statistics in Medicine*, *26*, 4531–4543.

Caron, M., St-Onge, P., Sontag, T., Wang, Y. C., Richer, C., Ragoussis, I., Sinnett, D., & Bourque, G. (2020). Single-cell analysis of childhood leukemia reveals a link between developmental states and ribosomal protein expression as a source of intra-individual heterogeneity. *Scientific Reports*, *10*(1), 1–12.

Caruso, J. C., & Cliff, N. (1997). Empirical size, coverage, and power of confidence intervals for Spearman's rho. *Educational and Psychological Measurement*, *57*(4), 637–654.

Chen, L., Liu, M., Bao, J., Xia, Y., Zhang, J., Zhang, L., Huan, X., & Wang, J. (2013). The correlation between apparent diffusion coefficient and tumor cellularity in patients: A meta-analysis. *PLoS One*, *8*(11), e79008.

Choi, S. C., & Wette, R. (1969). Maximum likelihood estimation of the parameters of the gamma distribution and their bias. *Technometrics*, *11*(4), 683–690.

Cochran, W. G. (1937). Problems arising in the analysis of a series of similar experiments. *Supplement to the Journal of the Royal Statistical Society*, *4*(1), 102–118.

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, *10*, 101–129.

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. Chapman and Hall.

Davidian, M., & Carroll, R. J. (1987). Variance function estimation. *Journal of the American Statistical Association*, *82*(400), 1079–1091.

DerSimonian, R., & Kacker, R. (2007). Random-effects model for meta-analysis of clinical trials: An update. *Contemporary Clinical Trials*, *28*(2), 105–114.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Control Clinical Trials*, *7*(3), 177–188.

Fieller, E. C., & Pearson, E. S. (1961). Tests for rank correlation coefficients: II. *Biometrika*, 29–40.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, *10*(4), 507–521.

Guolo, A. (2012). Higher-order likelihood inference in meta-analysis and meta-regression. *Statistics in Medicine*, *3*(1), 313–327.

Guolo, A., & Varin, C. (2012). The R package metaLik for likelihood inference in meta-analysis. *R Journal*, *50*(7), 1–4.

Gurka, M. J. (2006). Selecting the best linear mixed model under REML. *American Statistician*, *60*(1), 19–26.

Hardy, R. J., & Thompson, S. G. (1996). A likelihood approach to meta-analysis with random effects. *Statistics in Medicine*, *15*, 619–629.

Hartung, J., & Knapp, G. (2001). On tests of the overall treatment effect in the meta-analysis with normally distributed responses. *Statistics in Medicine*, *20*, 1771–1782.

In 't Hout, J., Ioannidis, J. P., & Borm, G. F. (2014). The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Medical Research Methodology*, *14*(25), 1–12.

Jackson, D., Bowden, J., & Baker, R. (2010). How does the DerSimonian and Laird procedure for random effects meta-analysis compare with its more efficient but harder to compute counterparts? *Journal of Statistical Planning and Inference*, *140*(4), 961–970.

Jackson, D., Law, M., Rücker, G., & Schwarzer, G. (2017). The Hartung-Knapp modification for random-effects meta-analysis: A useful refinement but are there any residual concerns? *Statistics in Medicine*, *36*(25), 3923–3934.

Jackson, D., & White, I. R. (2018). When should meta-analysis avoid making hidden normality assumptions? *Biometrical Journal*, *60*(6), 1040–1058.

Kontopantelis, E., & Reeves, D. (2012). Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: A comparison between DerSimonian– Laird and restricted maximum likelihood. *Statistical Methods in Medical Research*, *21*(6), 657–659.

Langan, D., Higgins, J. P., Jackson, D., Bowden, J., Veroniki, A. A., Kontopantelis, E., Viechtbauer, W., & Simmonds, M. (2019). A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Research Synthesis Methods*, *10*(1), 83–98.

Malzahn, U., Böhning, D., & Holling, H. (2000). Nonparametric estimation of heterogeneity variance for the standardised difference used in meta-analysis. *Biometrika*, *87*(3), 619–632.

McCulloch, C. E., & Searle, S. R. (2001). *Generalized, linear, and mixed models*. Wiley.

McNeish, D. M., & Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review*, *28*(2), 295–314.

Moschopoulos, P. G. (1985). The distribution of the sum of independent gamma random variables. *Annals of the Institute of Statistical Mathematics*, *37*(1), 541–544.

Mzolo, T., Hendriks, M., & van den Heuvel, E. (2013). A comparison of statistical methods for combining relative bioactivities from parallel line bioassays. *Pharmaceutical Statistics*, *12*(6), 375–384.

Nagashima, K., Noma, H., & Furukawa, T. A. (2019). Prediction interval for random-effects meta-analysis: a confidence distribution approach. *Statistical Methods in Medical Research*, *28*(6), 1689–1702.

Nilsson, A., Bonander, C., Strömberg, U., & Björk, J. (2019). Assessing heterogeneous effects and their determinants via estimation of potential outcomes. *European Journal of Epidemiology*, *34*(9), 823–835.

Noma, H. (2011). Confidence intervals for a random-effects meta-analysis based on Bartlett-type corrections. *Statistics in Medicine*, *30*(28), 3304–3312.

Partlett, C., & Riley, R. D. (2017). Random effects meta-analysis: Coverage performance of 95% confidence and prediction intervals following REML estimation. *Statistics in Medicine*, *36*(2), 301–317.

Petropoulou, M., & Mavridis, D. (2017). A comparison of 20 heterogeneity variance estimators in statistical synthesis of results from studies: A simulation study. *Statistics in Medicine*, *36*(27), 4266–4280.

Quintero, A., & Lesaffre, E. (2017). Multilevel covariance regression with correlated random effects in the mean and variance structure. *Biometrical Journal*, *59*(5), 1047–1066.

Sansanayudh, N., Anothaisintawee, T., Muntham, D., McEvoy, M., Attua, J., & Thakkinstian, A. (2014). Mean platelet volume and coronary artery disease: A systematic review and meta-analysis. *International Journal of Cardiology*, *175*(3), 433–440.

SAS Institute. (1996) SAS OnlineDoc. *The NLMIXED procedure*. SAS Institute Inc.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, *2*, 110–114.

Schmidt, D., Germano, A. M., & Milani, T. L. (2019). Subjective sensitivity data: Considerations to treat heteroscedasticity. *Cogent Medicine*, *6*(1), 1673086.

Schwarzer, G. (2007). meta: An R package for meta-analysis. *R News*, *7*, 40–45.

Sidik, K., & Jonkman, J. N. (2005). A note on variance estimation in random effects meta-regression. *Journal of Biopharmaceutical Statistics*, *15*(5), 823–838.

Skovgaard, I. M. (2001). Likelihood asymptotics. *Scandinavian Journal of Statistics*, *28*, 3–32.

Stevens, N. T., Steiner, S. H., & MacKay, R. J. (2018). Comparing heteroscedastic measurement systems with the probability of agreement. *Statistical Methods in Medical Research*, *27*(11), 3420–3435.

Tanizaki, H. (2004). Power comparison of empirical likelihood ratio tests: Small sample properties through Monte Carlo studies. *Kobe University Economic Review*, *50*(13), 13–25.

Thorlund, K., Wetterslev, J., Awad, T., Thabane, L., & Gluud, C. (2011). Comparison of statistical inferences from the DerSimonian–Laird and alternative random-effects model meta-analyses—An empirical assessment of 920 Cochrane primary outcome meta-analyses. *Research Synthesis Methods*, *2*(4), 238–253.

van den Heuvel, E. R., Almalik, O., & Zhan, Z. (2021). Simulation models for aggregated data meta-analysis: Evaluation of pooling effect sizes and publication bias. Submitted, https://doi.org/10.48550/arXiv.2009.06305

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J. P. T., Langan, D., & Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods*, *7*(1), 55–79.

Veroniki, A. A., Jackson, D., Bender, R., Kuss, O., Langan, D., Higgins, J. P. T., Knapp, G., & Salanti, G. (2019). Methods to calculate uncertainty in the estimated overall effect size from a random-effects meta-analysis. *Research Synthesis Methods*, *10*(1), 23–43.

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational Behavioral Statistics*, *30*(3), 261–293.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48.

SUPPORTING INFORMATION

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

How to cite this article: Almalik, O., Zhan, Z., & van den Heuvel, E. R. (2022). Jointly pooling aggregated effect sizes and their standard errors from studies with continuous clinical outcomes. *Biometrical Journal*, *64,* 1340–1360. https://doi.org/10.1002/bimj.202100108

APPENDIX A: PROGRAMMING CODES

Here, we list the SAS codes that were used for the bivariate approach and for the (restricted) maximum likelihood. For the analysis of maximum likelihood (ML) and the restricted maximum likelihood (REML), the information (Y_i, S_i^{-2}) of study *i* should be in the same row, where Y_i and S_i^{-2} are in separate columns. For the bivariate approach, the information Y_i and S_i^2 should be put in different rows in the same column.

SAS codes for REML and Hardy–Thompson (HT)

The data set “`Effect_Sizes`” should have three columns: one for `study` *i*, one for the `effect` Y_i, and one for the `weight` S_i^{-2}. Then the SAS codes are given by

```
PROC MIXED DATA = Effect_Sizes METHOD = REML;
  CLASS study;
  MODEL effect = / SOLUTION CL;
  RANDOM study;
  PARMS (1) (1) / HOLD = 2;
  WEIGHT weight;
RUN;
```

The maximum likelihood estimator of Hardy and Thompson (1996) can be obtained by applying METHOD = ML. The profile likelihood confidence intervals are not available through the procedure `MIXED`.

SAS codes for bivariate approach

The programming codes in proc `NLMIXED` assume that there exists a data set “`Effect_Sizes`” with different columns and rows. The rows represent studies which are listed in column “`Study`.” For each study, we have two separate rows: one

TABLE A.1 Schematic overview of how the data of a meta-analysis should be organized to execute our bivariate distribution approach

Study	Response	Outcome	Degrees
1	Effect size	Y_1	df_1
1	Variance	S_1^2	df_1
2	Effect size	Y_2	df_2
2	Variance	S_2^2	df_2
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮
m	Effect size	Y_m	df_m
m	Variance	S_m^2	df_m

row for the effect size Y_i and a second row for the variance S_i^2. The effect size Y_i and variance S_i^2 are below each other in the same column called "`Outcome`" and to identify these different responses we have a column "`Response`" with levels "`effect size`" and "`variance`." Finally, there is a column with the degrees of freedom for each study. Table A.1 shows schematically how the data are organized.

The SAS programming codes for the bivariate analysis is given below. Note that we must specify the degrees of freedom (`DOF =m-1`) for the variance of the pooled effect size.

```
PROC NLMIXED DATA = Effect_Sizes DOF =m-1;
  PARMS THETA = 0 LNSTAU = 0 SD = 10;
  TAU2 = EXP(2*LNSTAU);
  VAR_I = (SD**2)/Degrees;
  VAR_T = TAU2 + (SD**2)/Degrees;
  IF Response = ‘‘ effect size ’’
    THEN DENS = -0.5*LOG(VAR_T)-0.5*(( Outcome - THETA)**2)/VAR_T;
  ELSE IF Response = ‘‘ variance ’’
    THEN DENS = ( Degrees /2)*LOG( Degrees * Outcome /VAR_I)-0.5* Degrees * Outcome /VAR_I;
  MODEL Outcome ~ GENERAL (DENS);
RUN; QUIT;
```