Biometrika (2017), **xx**, x, *pp*. 1–19

Printed in Great Britain

Exact Inference on the Random-Effects Model for Meta-Analyses with Few Studies

BY H. MICHAEL

Department of Statistics, Stanford University

haben.michael@gmail.com

S. THORNTON, M. XIE

Department of Statistics, Rutgers University

AND L. TIAN

Department of Biomedical Data Science, Stanford University

1. INTRODUCTION

The random effects model is often used to account for between-study heterogeneity when conducting a meta-analysis. When the distribution of the primary study treatment effect estimates is approximately normal, the simple normal-normal model is commonly used, and the DerSimonian-Laird ("DL") method and its variations are the most popular approach to estimating the model's parameters and performing statistical inference (DerSimonian & Laird, 1986). However, the DL method is based on an asymptotic approximation and its use is only justified when the number of studies is large. In many fields, the number of studies used in a meta-analysis or sub-meta-analysis rarely exceeds 20 and is typically fewer than 7 (Davey et al., 2011), leaving inferences based on the DL estimator questionable. Indeed, extensive simulation studies have

found that the coverage probability of the DL-based confidence interval (CI) can be substantially lower than the nominal level in various settings (Kontopantelis et al., 2010; IntHout et al., 2014), leading to false positives. One reason for the failure of the DL method is that the asymptotic approximation ignores the variability in estimating the heterogeneous variance, which can be substantial when the number of studies is small (Higgins et al., 2009).

Various remedies have been proposed to correct the under-coverage of DL-based confidence intervals. Hartung & Knapp (2001) proposed an unbiased estimator of the variance of the DL point estimator explicitly accounting for the variability in estimating the heterogenous variance. Sidik & Jonkman (2006) used the heavy-tailed t-distribution to approximate the distribution of a modified Wald-type test statistic based on the DL estimator. Using the more robust t- rather than normal distribution has also been proposed (Berkey et al., 1995; Raghunathan, 1993; Follmann & Proschan, 1999). Hardy & Thompson (1996); Vangel & Rukhin (1999); Viechtbauer (2005); and Raudenbush (2009) proposed procedures based on maximum-likelihood estimation. Noma (2011) further improved the performance of the likelihood-based inference procedure when the number of study is small by using a Bartlett-type correction. Bayesian approaches incorporating external information have been developed by many authors (Smith et al., 1995; Higgins & Whitehead, 1996; Bodnar et al., 2017). However, with few exceptions, most of these methods still depend on an asymptotic approximation and their performance with very few studies has only been examined by specific simulation studies. To overcome these difficulties, potentially conservative but "exact" inference procedures for the random effects model have been proposed (Follmann & Proschan, 1999; Wang et al., 2010; Liu et al., 2017) and Wang & Tian (2017). A permutation rather than the asymptotic limiting distribution is used to approximate the distribution of the relevant test statistics and thus the validity of the associated inference is guaranteed for any number of studies. However, due to the discreteness of the permutation distribution, the

highest significance level that may be achieved without randomization depends on the number of studies. For example, a 95% confidence interval can only be constructed with more than 5 studies.

The main contribution of this paper is to propose a set of new methods for constructing exact, unconditional, non-randomized CIs for the location parameter of the normal-normal model by inverting exact tests. The coverage level of the resulting CI is guaranteed to be above the nominal level, up to Monte Carlo error, as long as the meta-analysis contains more than 1 study. After employing several techniques to accelerate computation, the new CI can be easily constructed on a personal computer. Simulations suggest that the proposed CI typically is not overly conservative. In Section 2, we present our procedure for constructing exact CIs for the population mean; in Section 3, we report results from comprehensive simulation studies; in Section 4, we illustrate the proposed method with a real data example; and in Section 5 we conclude the paper with additional discussion.

2. METHOD

The observed data consist of $\mathcal{Y}_0 = \{Y_k, k = 1, \cdots, K\}$, where Y_k follows a random effects model,

$$Y_k \mid \theta_k \overset{ind.}{\sim} N(\theta_k, \sigma_k^2),\ \theta_k \overset{ind.}{\sim} N(\mu_0, \tau_0^2),\ k = 1, \cdots, K,$$

with the variances $\sigma_k > 0, k = 1, \ldots, K$, assumed known.The random effects model implies The simple parametric model

$$Y_k \overset{ind.}{\sim} N(\mu_0, \sigma_k^2 + \tau_0^2),\ k = 1, \cdots, K. \tag{1}$$

In the context of a meta-analysis, the pairs $(Y_k, \sigma_k^2), k = 1, \ldots, K$, are interpreted as observed effects and known within-study variances drawn from K studies, respectively. The unobserved

population effect and between-study variance are μ_0 and τ_0^2, respectively. The goal is inference on the location parameter μ_0, viewing τ_0^2 as a nuisance parameter. The typical number of studies depends on the area of research and can be small, e.g., $K \leq 10$.

With τ_0^2 known, the uniformly minimum variance unbiased estimator of μ_0 under (1) is given by

$$\frac{\sum_{k=1}^{K} Y_k(\tau_0^2 + \sigma_k^2)^{-1}}{\sum_{k=1}^{K}(\tau_0^2 + \sigma_k^2)^{-1}}.$$

As τ_0^2 is unknown, DerSimonian & Laird (1986) propose substituting a simplified method of moments estimator,

$$\hat{\tau}_{DL}^2 = \max\left\{0, \frac{\sum_{k=1}^{K}(Y_k - \hat{\mu}_F)^2/\sigma_k^2 - (K-1)}{\sum_{k=1}^{K}\sigma_k^{-2} - \frac{\sum_{k=1}^{K}\sigma_k^{-4}}{\sum_{k=1}^{K}\sigma_k^{-2}}}\right\},$$

where

$$\hat{\mu}_F = \frac{\sum_{i=1}^{K} Y_k \sigma_k^{-2}}{\sum_{i=1}^{K}\sigma_k^{-2}}$$

is the minimum variance unbiased estimator of μ_0 under a fixed effects model, i.e., when $\tau_0^2 = 0$. The resulting estimator is known as the "DerSimonian-Laird" estimator of μ_0:

$$\hat{\mu}_{DL} = \frac{\sum_{k=1}^{K} Y_k(\hat{\tau}_{DL}^2 + \sigma_k^2)^{-1}}{\sum_{k=1}^{K}(\hat{\tau}_{DL}^2 + \sigma_k^2)^{-1}}.$$

By an analogous substitution, a level $1-\alpha$ confidence interval for μ_0 is given by

$$\left\{\hat{\mu}_{DL} - z_{1-\alpha/2}\left(\sum_{k=1}^{K}(\hat{\tau}_{DL}^2 + \sigma_k^2)^{-1}\right)^{-1/2}, \hat{\mu}_{DL} + z_{1-\alpha/2}\left(\sum_{k=1}^{K}(\hat{\tau}_{DL}^2 + \sigma_k^2)^{-1}\right)^{-1/2}\right\}. \tag{2}$$

The justification of the CI given in (2) relies on the asymptotic approximation

$$T_0(\mu_0; \mathcal{Y}) = (\hat{\mu}_{DL} - \mu_0)^2 \sum_{k=1}^{K}(\hat{\tau}_{DL}^2 + \sigma_k^2)^{-1} \rightsquigarrow \chi_1^2 \tag{3}$$

as the number of studies, K, grows to infinity and $\max\{\sigma_k\}/\min\{\sigma_k\}$ is uniformly bounded. However, when K is moderate or small, the distribution of $T_0(\mu_0; \mathcal{Y})$ depends on τ_0^2 and may be very different from a χ_1^2 distribution. Consequently, the finite-sample performance of the CI given

by (2) is often unsatisfactory. We propose constructing an exact CI for μ_0 by first constructing an exact confidence region for (μ_0, τ_0^2). To this end, let $T\left\{(\mu, \tau^2); \mathcal{Y}_0\right\}$ denote a test statistic, which may depend on the null parameter (μ, τ^2), for the simple hypothesis $(\mu_0, \tau_0^2) = (\mu, \tau^2)$. The specific choice of $T\left\{(\mu, \tau^2); \mathcal{Y}_0\right\}$ will be discussed later and here we only assume that a high value of $T\left\{(\mu, \tau^2); \mathcal{Y}_0\right\}$ represents grounds for rejection. For a given choice of $T\left\{(\mu, \tau^2); \mathcal{Y}_0\right\}$, a $1 - \alpha$ level CI for μ_0 can be constructed as follows:

1. Obtain bounds $[\mu_{min}, \mu_{max}]$ and $[\tau^2_{min}, \tau^2_{max}]$ for μ_0 and τ_0^2.
2. For each pair of μ and τ^2 in an $R \times R$ grid of points on $[\mu_{min}, \mu_{max}] \times [\tau^2_{min}, \tau^2_{max}]$,
   a. Compute the null distribution of $T\left\{(\mu, \tau^2); \mathcal{Y}_0\right\}$, i.e., the distribution of $T\left\{(\mu, \tau^2); \mathcal{Y}(\mu, \tau^2)\right\}$, where
   $$\mathcal{Y}(\mu, \tau^2) = \left\{\widetilde{Y}_k, k = 1, \cdots, K\right\}$$
   with $\widetilde{Y}_k \overset{ind.}{\sim} N(\mu, \sigma_k^2 + \tau^2), k = 1, \cdots, K$.
   b. Compute the p-value $p_{\mu,\tau^2}(\mathcal{Y}_0) := P\left[T\left\{(\mu, \tau^2); \mathcal{Y}_0\right\} > T\left\{(\mu, \tau^2); \mathcal{Y}(\mu, \tau^2)\right\}\right]$.
3. Obtain a confidence region for (μ_0, τ_0^2) as $\Omega_{1-\alpha}(\mathcal{Y}_0) := \{(\mu, \tau^2) : p_{\mu,\tau^2}(\mathcal{Y}_0) > \alpha\}$.
4. Project $\Omega_{1-\alpha}(\mathcal{Y}_0)$ onto the μ axis to obtain a CI for $\mu_0 : \{\mu : (\mu, \tau^2) \in \Omega_{1-\alpha}(\mathcal{Y}_0)\}$.

This method generates the exact CI for μ_0 in the sense that

$$\text{pr}\left(\mu_0 \in \{\mu : (\mu, \tau^2) \in \Omega_{1-\alpha}(\mathcal{Y}_0)\}\right) \geq 1 - \alpha.$$

This is due to the fact that

$$\begin{aligned}
&\text{pr}\left(\mu_0 \in \{\mu : (\mu, \tau^2) \in \Omega_{1-\alpha}(\mathcal{Y}_0)\}\right) \\
\geq &\text{pr}\left\{(\mu_0, \tau_0^2) \in \Omega_{1-\alpha}(\mathcal{Y}_0)\right\} \\
= &\text{pr}\left\{p_{\mu_0,\tau_0^2}(\mathcal{Y}_0) \geq \alpha\right\} \\
= &\text{pr}(U \geq \alpha) = 1 - \alpha,
\end{aligned}$$

where the random variable U follows the unit uniform distribution. Here, we assume that $\tau_0 \in [\tau^2_{min}, \tau^2_{max}]$. If $\tau^2_{\min}$ and τ^2_{max} are chosen depending on the data in such a way that $\text{pr}(\tau^2_{min} < \tau^2 < \tau^2_{max}) \geq 1 - \beta$, then the guaranteed coverage probability of the proposed CI is $1 - \alpha - \beta \approx 1 - \alpha$ for very small β.

The cumulative distribution function (CDF) of $T\left\{(\mu, \tau^2); \mathcal{Y}(\mu, \tau^2)\right\}$ may not be analytically tractable, but it is well defined for any given grid point (μ, τ^2) and can always be approximated by a Monte Carlo simulation. To be specific, given (μ, τ^2), we may approximate the distribution of $T\left\{(\mu, \tau^2); \mathcal{Y}(\mu, \tau^2)\right\}$ in 2a as follows:

2a For $b = 1, \cdots, B$,

a. Generate $e^*_{1b}, \cdots, e^*_{Kb} \overset{ind.}{\sim} N(0, 1)$.

b. Let $Y^*_{kb} = \mu + (\sigma_k^2 + \tau^2)^{1/2} e^*_{kb}$, $k = 1, \cdots, K$, and let $\mathcal{Y}^*_b = \left\{Y^*_{kb}, k = 1, \cdots, K\right\}$.

c. Let $T^*_b = T\left\{(\mu, \tau^2); \mathcal{Y}^*_b\right\}$ be the corresponding test statistic based on the generated data $\mathcal{Y}^*_b$. The empirical distribution of $\{T^*_1, \cdots, T^*_B\}$ can be used to approximate the distribution of $T\left\{(\mu, \tau^2); \mathcal{Y}(\mu, \tau^2)\right\}$.

Since the estimation of the null distribution in 2a does not depend on any asymptotic approximation, both the p-value, $p_{\mu,\tau^2}(\mathcal{Y}_0)$, and the confidence region, $\Omega_{1-\alpha}(\mathcal{Y}_0)$, are "exact" if we can

safely ignore the errors of the grid approximation and the Monte Carlo simulation above, which can be controlled by increasing the grid density and B in step 2a, respectively.

Because the data $Y_k, k = 1, \cdots, K$, are distributed as $\mathcal{N}(\mu, \sigma_k^2 + \tau_0^2), k = 1, \ldots, K$, whenever the shifted data $Y_k - \mu, k = 1, \ldots, K$, are distributed as $\mathcal{N}(0, \sigma_k + \tau_0^2), k = 1 \ldots, K$, we restrict our focus to equivariant statistics (Lehmann & Romano, 2006), that is, T satisfying $T\left\{(\mu, \tau^2); \mathcal{Y}_0\right\} = T\left\{(0, \tau^2), \mathcal{Y}_0 - \mu\right\}$, where $\mathcal{Y}_0 - \mu = \{Y_k - \mu, k = 1, \ldots, K\}$. In this situation, testing the null $H_0 : (\mu_0, \tau_0^2) = (\mu, \tau^2)$ based on the data $\mathcal{Y}_0$ is the same as testing the null $H_0 : (\mu_0, \tau_0^2) = (0, \tau^2)$ based on the shifted data $\mathcal{Y}_0 - \mu$. When the test statistic is equivariant, the computations in step 2a need only be performed once for each τ^2 in the grid rather than each pair (μ, τ^2). Thus, although a 2-dimensional grid is used in the algorithm, the computational complexity remains linear in the grid size, R. More specifically, steps 2–3 become:

2′. For each τ^2 of an R-sized grid on $[\tau^2_{min}, \tau^2_{max}]$,

   a. Compute the distribution of $T\left\{(0, \tau^2); \mathcal{Y}(0, \tau^2)\right\}$.

   b. Compute $q_{1-\alpha;\tau^2}$, the $1 - \alpha$ quantile of $T\left\{(0, \tau^2); \mathcal{Y}(0, \tau^2)\right\}$.

   c. Compute $\Omega_{1-\alpha}(\tau^2; \mathcal{Y}_0) = \{(\mu, \tau^2) \mid T\left\{(\mu, \tau^2); \mathcal{Y}_0\right\} = T\left\{(0, \tau^2); \mathcal{Y}_0 - \mu\right\} \geq q_{1-\alpha;\tau^2}\}$.

3′. Compute a $(1 - \alpha)$-level confidence region for (μ_0, τ_0^2) as

$$\bigcup_{\tau^2 \in [\tau^2_{min}, \tau^2_{max}]} \Omega_{1-\alpha}(\tau^2; \mathcal{Y}_0).$$

In this paper, we propose using the test statistics

$$T\left\{(\mu, \tau^2); \mathcal{Y}_0\right\} = T_0(\mu; \mathcal{Y}) + c_0 T_{lik}\left\{(\mu, \tau^2); \mathcal{Y}\right\}, \tag{4}$$

where $T_0(\mu; \mathcal{Y})$ is the same Wald-type test statistic used in the Dersimonian-Laird procedure,

$$T_{lik}\left\{(\mu, \tau^2); \mathcal{Y}\right\} = -\frac{1}{2}\sum_{k=1}^{K}\left[\frac{(Y_k - \hat{\mu}_{DL})^2}{\hat{\tau}^2_{DL} + \sigma^2_k} + \log\left\{2\pi(\hat{\tau}^2_{DL} + \sigma^2_k)\right\}\right] + \sum_{k=1}^{K}\frac{1}{2}\left[\frac{(Y_k - \mu)^2}{\tau^2 + \sigma^2_k} + \log\left\{2\pi(\tau^2 + \sigma^2_k)\right\}\right],$$

and c_0 is a tuning parameter controlling the relative contributions of these two statistics. While $T_0(\mu; \mathcal{Y})$ directly focuses on the location parameter μ_0, $T_{lik}\left\{(\mu, \tau^2); \mathcal{Y}\right\}$, similar to the likelihood ratio test statistic, targets the combination of μ_0 and τ^2_0 and helps to construct a narrower CI of μ_0 when the number of studies is small. The proposed test statistics satisfy the equivariance condition, ensuring speedy computation when carrying out the procedure on a typical personal computer.

A further simplification afforded by this choice of test statistics is that step 2′c may be carried out by solving a simple quadratic inequality:

$$A(\tau)\mu_0^2 + B(\tau)\mu_0 + C(\tau) < 0,$$

where

$$\begin{aligned}
A(\tau) &= \sum_{k=1}^{K}\left\{\frac{1}{\hat{\tau}^2_{DL} + \sigma^2_k} + \frac{c_0}{2(\tau^2 + \sigma^2_k)}\right\} > 0,\\
B(\tau) &= -\sum_{k=1}^{K}\left\{\frac{2\hat{\mu}_{0DL}}{\hat{\tau}^2_{0DL} + \sigma^2_k} + \frac{c_0 Y_k}{\tau^2 + \sigma^2_k}\right\},\\
C(\tau) &= \sum_{k=1}^{K}\frac{c_0}{2}\left[\frac{Y_k^2}{\tau^2 + \sigma^2_k} + \log\frac{\tau^2 + \sigma^2_k}{\hat{\tau}^2_{DL} + \sigma^2_k} - \frac{(Y_k - \hat{\mu}_{DL})^2}{\hat{\tau}^2_{DL} + \sigma^2_k}\right] + \hat{\mu}^2_{DL}\sum_{k=1}^{K}\frac{1}{\hat{\tau}^2_{DL} + \sigma^2_k} - q_{1-\alpha;\tau^2}. \qquad (5)
\end{aligned}$$

As a result, the confidence interval of μ_0 when $\tau_0 = \tau$ $\Omega_{1-\alpha}(\tau^2; \mathcal{Y}_0)$, is simply the segment with endpoints

$$\left(\frac{-B(\tau) \pm \Delta^{1/2}}{2A(\tau)}, \tau^2\right),$$

when $\Delta(\tau) = B(\tau)^2 - 4A(\tau)C(\tau) \geq 0$, and an empty set, otherwise.

To choose τ^2_{min} and τ^2_{max} in step 1 of the algorithm, we may use the endpoints of a $100(1-\beta)\%$, e.g., 99.9%, confidence interval of τ^2_0. This CI can be constructed by inverting the pivotal statistic

$$T_3(\tau^2) = (\mathbf{WY})' \left\{\mathbf{W\Sigma}(\tau)\mathbf{W}'\right\}^{-1} (\mathbf{WY}),$$

where $\mathbf{Y} = (Y_1, \cdots, Y_K)'$, $\mathbf{\Sigma}(\tau) = \text{diag}\left\{\sigma_1^2 + \tau^2, \cdots, \sigma_K^2 + \tau^2\right\}$, and

$$\mathbf{W} = \begin{pmatrix} \sigma_1^{-2}/\sum_{i=1}^K \sigma_i^{-2} - 1 & \sigma_2^{-2}/\sum_{i=1}^K \sigma_i^{-2} & \cdots & \sigma_K^{-2}/\sum_{i=1}^K \sigma_i^{-2} \\ \sigma_1^{-2}/\sum_{i=1}^K \sigma_i^{-2} & \sigma_2^{-2}/\sum_{i=1}^K \sigma_i^{-2} - 1 & \cdots & \sigma_K^{-2}/\sum_{i=1}^K \sigma_i^{-2} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_1^{-2}/\sum_{i=1}^K \sigma_i^{-2} & \sigma_2^{-2}/\sum_{i=1}^K \sigma_i^{-2} & \cdots & \sigma_K^{-2}/\sum_{i=1}^K \sigma_i^{-2} - 1 \end{pmatrix}.$$

The pivot follows a χ^2_{K-1} distribution when $\tau^2 = \tau^2_0$.

Since our goal is a CI for μ_0, the shape of the confidence region is crucial to its performance: the projection of $\Omega_{1-\alpha}(\mathcal{Y}_0)$ onto the μ axis should be as small as possible, relative to the area of the confidence region. Figure 1 plots two confidence regions with the same confidence coefficient, but substantially different projected lengths. To avoid an overly conservative CI, we prefer a confidence region with boundaries parallel to the $\tau-$ axis, or nearly so. The shape of $\Omega_{1-\alpha}(\mathcal{X}_0)$ is determined by the way we combine $T_0(\mu; \mathcal{Y})$ and $T_{lik}\left\{(\mu, \tau^2); \mathcal{Y}\right\}$ or, more generally, by the choice of $T\left\{(\mu, \tau^2); \mathcal{Y}\right\}$. Because the proposed statistics (4) are quadratic in μ, the resulting confidence regions are a union of intervals with similar centers and tend not to produce overly conservative CIs when the tuning parameter c_0 is chosen appropriately.

The proposed test statistic was chosen to balance performance and computation costs. For example, the true likelihood ratio test statistic under model (1) may be more informative than $T_{lik}\left\{(\mu, \tau^2); \mathcal{Y}\right\}$, but its evaluation involves computing the maximum likelihood estimate and is substantially slower. The proposed algorithm is easily parallelized, so further gains in computing speed are available.

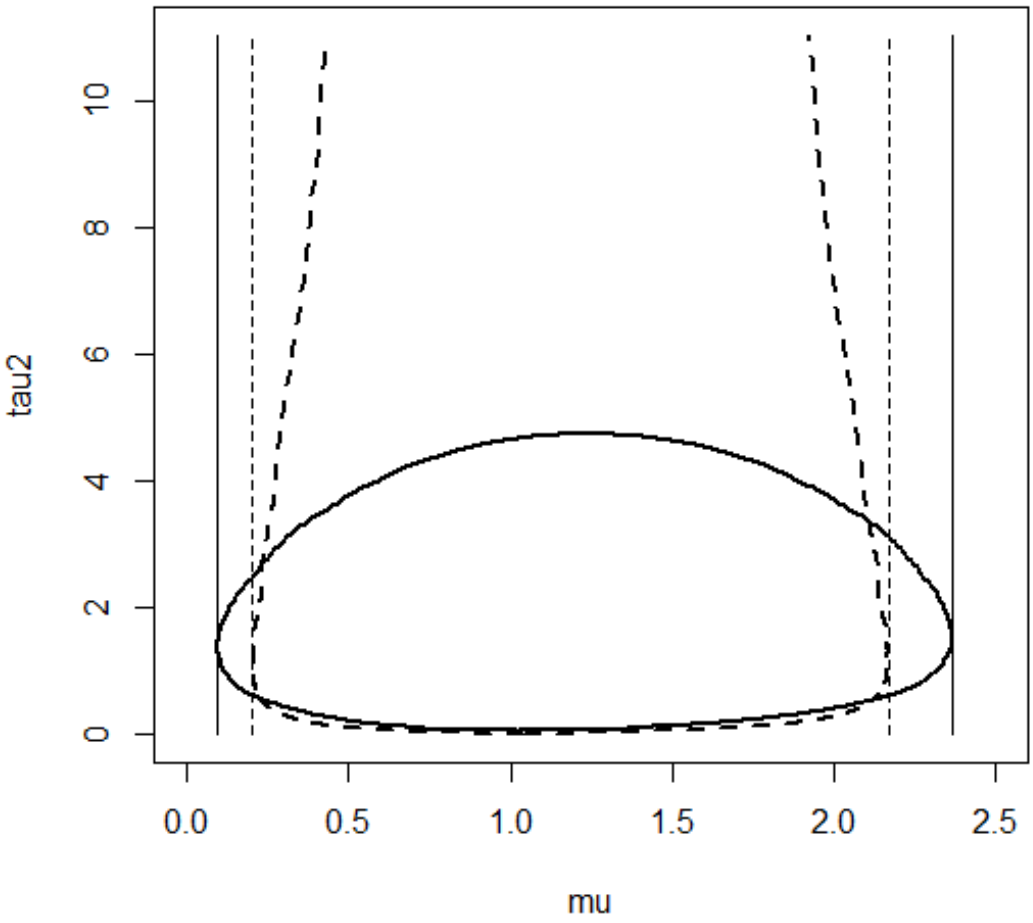

Fig. 1: The projection of the confidence region; the solid and dashed thick lines are boundaries of two confidence regions.

Remark 1. Projecting the confidence region parallel to the direction of the nuisance parameter is a geometric interpretation of a well-known approach to constructing non-randomized, unconditional, exact tests in the presence of nuisance parameters. In general, given a parameter of interest, θ, and nuisance parameter, η, let $p_{\theta,\eta}(\mathcal{Y}_0)$ denote the p-value for testing the null hypothesis $H_0 : (\theta_0, \eta_0) = (\theta, \eta)$ conditional on the observed data, $\mathcal{Y}_0$. An exact level α test for

the composite null hypothesis $H_0 : \theta_0 = \theta$ rejects the null if $\sup_\eta p_{\theta,\eta}(\mathcal{Y}_0) < \alpha$. This test is conservative by construction. A correspondingly conservative CI may be obtained by inversion as $\{\theta : \sup_\eta p_{\theta,\eta}(\mathcal{Y}_0) > \alpha\} = \{\theta : p_{\theta,\eta}(\mathcal{Y}_0) > \alpha \text{ for some } \eta\}$, i.e., the projection described in (2). See, e.g., Suissa & Shuster (1985) for an application to comparing proportions from two independent binomial distributions.

3. Numerical Study

In this section, we study the small-sample performance of the proposed method through a comprehensive simulation study. Observed data are simulated under the random effects model

$$Y_k \sim N(\mu_k, \tau_0^2 + \sigma_k^2), k = 1, \cdots, K,$$

where $\sigma_1, \cdots, \sigma_K$, are K equally spaced points in the interval $[1, 5]$, that is, $\sigma_k = 1 + 4(k - 1)/(K - 1), k = 1, \cdots, K$. The population variance τ_0^2 takes values 0, 12.5, and 25 to mimic settings with low, moderate, and high study heterogeneity, respectively. The corresponding I^2 measures of heterogeneity are approximately 0, 50%, and 70%, respectively.

In the first set of simulations, we examine the effect of the tuning parameter c_0 on the performance of the proposed method. For each set of simulated data, we construct a series of CIs using the proposed method with c_0 ranging from 0 to 2.5 in increments of 0.1, and the number of studies K ranges from 3 to 20. Based on results from 10000 simulated datasets under each combination of settings, we calculate the empirical coverage levels and average lengths of the resulting 95% CIs. In all settings, the empirical coverage levels of the proposed CIs are above the nominal level and therefore we optimize power by selecting the value of c_0 with the shortest CI lengths. When $K \geq 10$, the choice of c_0 does not have a pronounced effect on CI length. When K is between 3 and 6, the setting of primary interest, assigning more weight to the likelihood ratio-type statistic typically reduces the length of the CIs. We summarize the value of c_0 achieving

the minimum mean 95% CI length in Figure 2. Based on these results, we suggest for a tuning parameter $c_0 = 1.2$ for meta-analyses with fewer than 6 studies, $c_0 = 0.6$ for meta-analysess with 6–10 studies, $c_0 = 0.2$ for meta-analysis with 10-20 studies, and $c_0 = 0$ for analysis with more than 20 studies.

In the second set of simulations, we compare the performance of the proposed CIs with existing alternatives. For 10000 replicates at each data-generation setting described above, we construct CIs using the DerSimonian-Laird, Sidik-Jonkman, and restricted maximum likelihood asymptotic variance estimates, as well as the proposed CI with the recommended tuning parameter. In Figure 3 we summarize the average coverage and lengths of these CIs. In the presence of moderate heterogeneity, $I_2 = 0.5$, the empirical coverage level of the DL method is below 90% when $K \leq 10$, with the lowest coverage $\sim 75\%$ when the number of studies is 3. The CIs based on the Sidik-Jonkman estimator have better coverage, but still drop below 90% when $K \leq 5$. In contrast, the proposed exact CIs using the recommended tuning parameter settings do not fall below the nominal 95% coverage level. Morover, the coverage level is not overly conservative even for small Ks. The length of the 95% CI is comparable to the lengths of the asymptotic CIs, when these match the nominal coverage level, e.g., $K = 20$. When $I^2 = 0$, i.e., the random effects model degenerates to the fixed effects model, all methods, including the asymptotic estimators, control the Type 1 error. Sidik-Jonkman's CI is overly conservative even for moderate K values, while the proposed CIs, also overly conservative at lower values of K, improve steadily as K increases. When $I^2 = 0.70$, only the proposed CIs maintain the proper coverage level, while all methods fall below the nominal level for K as large as 10–20.

Several other estimators, including Hedges-Olkin, Hunter-Schmidt, and maximum likelihood, were also tested, with performance found to be generally intermediate between the performance of the DerSimonian-Laird and Sidik-Jonkman estimators.

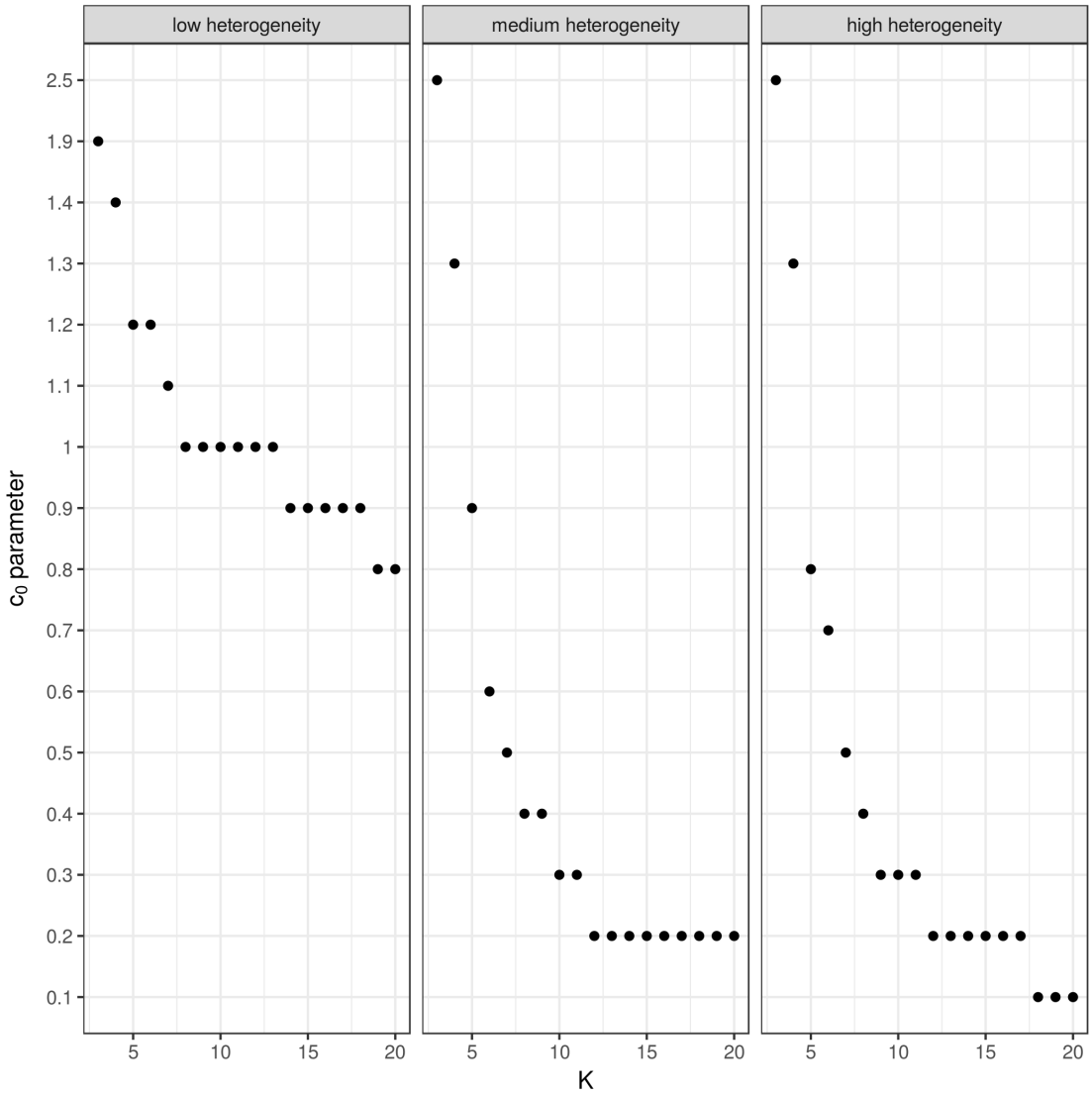

Fig. 2: The choice of c_0 achieving the minimum mean 95% CI length is plotted against the number K of studies, at 3 levels of between-study heterogeneity.

4. EXAMPLE

Tai et al. (2015) conduct a random effects meta-analysis of 59 randomized controlled trials to determine if increased calcium intake affects bone mineral density ("BMD"). Altogether, these trials measured the changes in BMD at five skeletal sites over three time points and measured

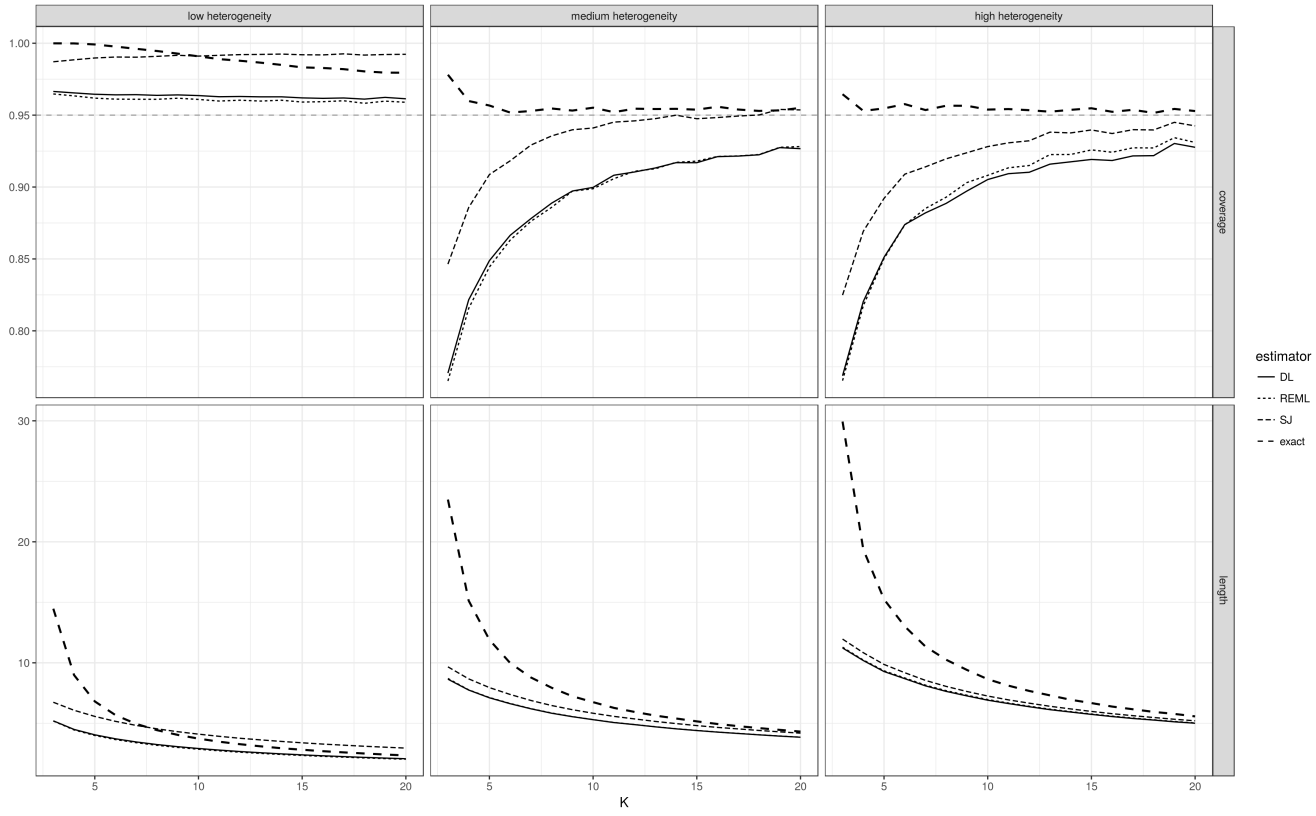

Fig. 3: Comparison by 95% CI coverage and length of the proposed estimator with 3 commonly used estimators based on asymptotic approximations. Data was generated according to model (1) with the number of studies K varying between 3 and 20 and the ratio of between- to average within-variance adjusted to give 3 levels of between-study heterogeneity. The proposed estimator achieves the nominal size at all configurations, with overcoverage evident where the heterogeneity is low or the studies is very few (3–4).

the effect of calcium intake on BMD from dietary sources and from calcium supplements. We illustrate the proposed method using four meta-analyses. The first meta-analysis investigates changes in BMD of the lumbar spine and is based on the findings of 27 trials that lasted fewer than 18 months. As shown in Table 1, the 95% CI produced by the proposed exact method does not differ very much from the 95% CI based on the DL method. The two intervals have a similar

length and are centered around a BMD difference of about 1.2. We also construct the exact CI by permuting a Hodge-Lehman type estimator (Liu et al., 2017). The resulting interval is very similar to the interval produced by the proposed method. These similarities are to be expected since the normality assumptions of the DL estimator may not be too unreasonable for a meta-analysis based on such a large number of primary studies.

Two of the other random effects meta-analyses investigate changes in BMD in the hip and forearm for trials of size six and five, respectively, that lasted for more than two years. The fourth analysis we consider here is the meta-analysis of three trials that lasted fewer than 18 months and measured changes in BMD for the total body of subjects. For these three meta-analyses, however, the number of studies is small, and the DL method may be expected to fall short of the nominal level. In the hip study, the proposed exact method and the DL method both yield the same conclusion, producing 95% confidence intervals rejecting the null of no change in BMD, although the exact method produces confidence intervals that are wider than their DL counterparts. In contrast, the DL 95% confidence intervals for the forearm and total body studies find a significant change in BMD whereas the exact method does not, suggesting that the DL method may be giving a false positive in these two cases. The intervals and their lengths are given in Table 1. Note that the exact 95% CI based on the permutation method is not available for the last two meta analyses, since the number of studies is fewer than 6.

5. Discussion

We have proposed a method to construct an exact CI for the population mean under the normal-normal model commonly used in meta-analysis. Appropriate coverage is guaranteed, up to Monte Carlo error, even when the number of studies used in the meta-analysis is as small as 2. While convenient, the normal assumption for the study-specific treatment effect estimate may not be

Study	K	DerSimonian-Laird	Permutation	Proposal
lumbar spine	27	0.828–1.669 (0.841)	0.788–1.758 (0.970)	0.768–1.726 (0.958)
total hip	6	0.502–1.847 (1.345)	0.000–2.298 (2.298)	0.159–2.246 (2.087)
forearm	5	0.209–3.378 (3.169)		-0.459–4.124 (4.583)
total body	3	0.268–1.778 (1.511)		-0.74–2.796 (3.536)

Table 1: Random effects meta-analyses of the effect of calcium supplements on percentage change in bone mineral density (Tai et al. (2015), Figs. 1, 3, and 7). The meta-analyses were carried out using the DerSimonian-Laird variance estimator (as in Tai et al. (2015)), the permutation test of Wang & Tian (2017), applicable to meta-analyses with 6 or more studies, and the proposed exact method. On the two smaller meta-analyses ($K = 3, 5$) the proposed exact method fails to reject the null of no change, whereas the asymptotic DL method does reject.

valid in other settings. For example, the treatment effect estimate may be an odds ratio from a 2x2 contingency table. If the total sample sizes are small or if cell entries are close to 0, the normal assumption for the odds ratio may be inappropriate. More generally, Y_k may be a quantity relevant to a treatment effect with $Y_k|\theta_k$ following a non-normal, e.g., hypergeoemtric, distribution depending on the study-specific parameter θ_k. In such a case, the model for θ_k and the corresponding inference procedure warrant further research. More recently, there have been several new developments on confidence distribution and related generalized fiducial inference that have facilitated new inference procedures for meta-analysis (Xie & Singh, 2013; Claggett et al., 2014). These developments may also be promising directions for developing exact inference procedures for meta-analysis.

Routines in the R programming language for computing exact CIs for the population mean by the method proposed here are available at:

`https://github.com/haben-michael/rma-exact-pkg`.

REFERENCES

BERKEY, C. S., HOAGLIN, D. C., MOSTELLER, F. & COLDITZ, G. A. (1995). A random-effects regression model for meta-analysis. *Statistics in Medicine* **14**, 395–411.

BODNAR, O., LINK, A., ARENDACKÁ, B., POSSOLO, A. & ELSTER, C. (2017). Bayesian estimation in random effects meta-analysis using a non-informative prior. *Statistics in Medicine* **36**, 378–399.

CLAGGETT, B., XIE, M. & TIAN, L. (2014). Meta-analysis with fixed, unknown, study-specific parameters. *Journal of the American Statistical Association* **109**, 1660–1671.

DAVEY, J., TURNER, R. M., CLARKE, M. J. & HIGGINS, J. P. (2011). Characteristics of meta-analyses and their component studies in the cochrane database of systematic reviews: a cross-sectional, descriptive analysis. *BMC medical research methodology* **11**, 160.

DERSIMONIAN, R. & LAIRD, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials* **7**, 177–188.

FOLLMANN, D. A. & PROSCHAN, M. A. (1999). Valid inference in random effects meta-analysis. *Biometrics* **55**, 732–737.

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 meta-analysis with normally distributed responses. *Statistics in Medicine* **20**, 1771–1782.

HIGGINS, J., THOMPSON, S. G. & SPIEGELHALTER, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **172**, 137–159.

HIGGINS, J. & WHITEHEAD, A. (1996). Borrowing strength from external trials in a meta-analysis. *Statistics in Medicine* **15**, 2733–2749.

INTHOUT, 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.

KONTOPANTELIS, E., REEVES, D. et al. (2010). metaan: Random-effects meta-analysis. *Stata Journal* **10**, 395.

LEHMANN, E. L. & ROMANO, J. P. (2006). *Testing statistical hypotheses*. Springer Science & Business Media.

LIU, S., LEE, S. & XIE, M. (2017). Exact inference on meta-analysis with generalized fixed-effects and random-effects models. *Biostatistics & Epidemiology* , under review.

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

RAGHUNATHAN, T. (1993). Analysis of binary data from a multicentre clinical trial. *Biometrika* **80**, 127–139.

RAUDENBUSH, S. W. (2009). Analyzing effect sizes: Random-effects models. *The handbook of research synthesis and meta-analysis* **2**, 295–316.

SIDIK, K. & JONKMAN, J. N. (2006). Robust variance estimation for random effects meta-analysis. *Computational Statistics & Data Analysis* **50**, 3681–3701.

SMITH, T. C., SPIEGELHALTER, D. J. & THOMAS, A. (1995). Bayesian approaches to random-effects meta-analysis: a comparative study. *Statistics in Medicine* **14**, 2685–2699.

SUISSA, S. & SHUSTER, J. J. (1985). Exact unconditional sample sizes for the 2× 2 binomial trial. *Journal of the Royal Statistical Society. Series A (General)* , 317–327.

TAI, V., LEUNG, W., GREY, A., REID, I. R. & BOLLAND, M. J. (2015). Calcium intake and bone mineral density: systematic review and meta-analysis. *BMJ* **351**, h4183.

VANGEL, M. G. & RUKHIN, A. L. (1999). Maximum likelihood analysis for heteroscedastic one-way random effects ANOVA in interlaboratory studies. *Biometrics* **55**, 129–136.

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

WANG, R., TIAN, L., CAI, T. & WEI, L. (2010). Nonparametric inference procedure for percentiles of the random effects distribution in meta-analysis. *The Annals of Applied Statistics* **4**, 520.

WANG, Y. & TIAN, L. (2017). An efficient numerical algorithm for exact inference in meta analysis. *Journal of Statistical Computation and Simulation* , under review.

XIE, M.-G. & SINGH, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: A review. *International Statistical Review* **81**, 3–39.