Prediction intervals for random-effects meta-analysis: a confidence distribution approach

Kengo Nagashima[1]*, Hisashi Noma[2], and Toshi A. Furukawa[3]

Abstract

For the inference of random-effects models in meta-analysis, the prediction interval was proposed as a summary measure of the treatment effects that explains the heterogeneity in the target population. While the Higgins–Thompson–Spiegelhalter (HTS) plug-in-type prediction interval has been widely used, in which the heterogeneity parameter is replaced with its point estimate, its validity depends on a large sample approximation. Most meta-analyses, however, include less than 20 studies. It has been revealed that the validity of the HTS method is not assured under realistic situations, but no solution to this problem has been proposed in literature. Therefore, in this article, we describe our proposed prediction interval. Instead of using the plug-in scheme, we developed a bootstrap approach using an exact confidence distribution to account for the uncertainty in estimation of the heterogeneity parameter. Compared to the HTS method, the proposed method provides an accurate prediction interval that adequately explains the heterogeneity of treatment effects and the statistical error. Simulation studies demonstrated that the HTS method had poor coverage performance; by contrast, the coverage probabilities for the proposed method satisfactorily retained the nominal level. Applications to three published random-effects meta-analyses are presented.

Keywords

Confidence distributions; Coverage properties; Meta-analysis; Prediction intervals; Random-effects models

[1] *Department of Global Clinical Research, Chiba University Graduate School of Medicine, Japan.*

[2] *Department of Data Science, The Institute of Statistical Mathematics, Japan.*

[3] *Departments of Health Promotion and Human Behavior and of Clinical Epidemiology, Kyoto University Graduate School of Medicine / School of Public Health, Japan.*

*__Corresponding author__: nshi@chiba-u.jp

This is a preprent / submitted version (07-Apr-2017) of the following article: “Nagashima K, Noma H, Furukawa TA. Prediction intervals for random-effects meta-analysis: a confidence distribution approach. *Under review*.

1. Introduction

Meta-analysis is an important tool in scientific research for combining the results of multiple related studies. Two major approaches (i.e., fixed-effect models and random-effects models) have been widely applied. One frequently important objective of meta-analysis is to estimate the overall mean effect and its confidence interval [1].

Fixed-effect models assume that the true treatment effects are equal for all studies. The common treatment effect parameter estimate and its confidence interval provide valuable information for applying the results to other subpopulations. By contrast, random-effects models assume that the true treatment effects differ for each study. The average treatment effect across all studies and its confidence interval have been used together with heterogeneity measures that are also very important in terms of generalizability. For instance, the I^2-statistic [2, 3] has been widely used as a heterogeneity measure. However, researchers have interpreted summary results from random-effects models as an estimate of the average treatment effect rather than the common treatment effect [4, 5], which means that they tend to ignore the heterogeneity.

Subsequently, Higgins *et al.* [6] proposed a prediction interval for a treatment effect in a future study. It can be interpreted as the range of the predicted treatment effect in a new study, given the data. A prediction interval naturally takes into account the heterogeneity, and helps us apply the results to other subpopulations. Riley *et al.* [4] recommended that a prediction interval should be additionally reported along with a confidence interval and a heterogeneity measure.

The invalidity problem (i.e., the under-coverage property) of confidence intervals in random-effects meta-analysis has been well studied in the literature [6, 7, 8], since the number of synthesized studies is less than 20 in most meta-analyses in medical research [9] and a large sample approximation is not generally assured. By contrast, the small-sample problem of prediction intervals has not been very well examined so far, and there has been rising concern regarding this issue in meta-analysis. Recently, Partlett and Riley [10] revealed that the same problem occurs with prediction intervals. They showed that prediction intervals could have serious under-coverage properties under the general settings of medical meta-analyses, and it was considered that the ordinary methods of constructing prediction intervals, including the Higgins–Thompson–Spiegelhalter (HTS) prediction interval [6], are no longer valid. However, no explicit solution to this problem has been obtained thus far.

The HTS prediction interval has a fundamental problem. It can be regarded as a plug-in estimator that replaces the heterogeneity parameter τ^2 with its point estimate $\hat{\tau}^2$. The t distribution with $K-2$ degrees of freedom is used to approximately account for the uncertainty of $\hat{\tau}^2$, where K is the number of studies. The replacement with the t-approximation has a detrimental impact on the coverage probability, especially under a small number of studies. This is of particular concern since most meta-analyses include less than 20 studies. Thus, the HTS prediction interval can have severe under-coverage, as will be shown in Section 3. It is necessary to more precisely account for the uncertainty of $\hat{\tau}^2$.

In this article, to solve this very important problem, we develop a new prediction interval that is valid under more general and realistic settings of meta-analyses in medical research, including those whose K is especially small. To avoid using a plug-in estimator, we propose a parametric bootstrap approach using a confidence distribution to account for the uncertainty of $\hat{\tau}^2$ with an exact distribution estimator of τ^2 [11, 12, 13, 14, 15]. A confidence distribution, like a Bayesian posterior, is considered as a distribution function to estimate the parameter of interest in frequentist inference.

This article is organized as follows. In Section 2, we first briefly review the random-effects meta-analysis and the HTS prediction interval, then we provide the new method to construct an accurate prediction interval. In Section 3, we assess the performance of the HTS prediction interval and the proposed prediction interval via simulations. In Section 4, we apply the developed method to three meta-analysis data sets. Finally, we conclude the paper with a brief discussion.

2. Method

2.1 The random-effects model and the exact distribution of Cochran's Q statistic

We consider the random-effects model [6, 16, 17, 18, 19].

Condition 1. Let the random variable Y_k ($k = 1, 2, \ldots, K$) be an effect size estimate from the k-th study. The random-effects model can be defined as

$$\begin{aligned} Y_k &= \theta_k + \epsilon_k, \\ \theta_k &= \mu + u_k, \end{aligned} \tag{1}$$

where θ_k is the true effect size of the k-th study, μ is the grand mean parameter of the average treatment effect, ϵ_k is the random error within a study, and u_k is the random error across the studies. It is assumed

that ϵ_k and u_k are independent, with $\epsilon_k \sim N(0, \sigma_k^2)$ and $u_k \sim N(0, \tau^2)$, where the within-studies variances σ_k^2 are known and replaced by their valid estimates [20, 21], and the across-studies variance τ^2 is an unknown parameter that reflects the treatment effects heterogeneity.

Under Condition 1, the marginal distribution of Y_k is a normal distribution with the mean μ and the variance $\sigma_k^2 + \tau^2$.

Random-effects meta-analyses generally estimate μ to evaluate the average treatment effect and τ^2 to evaluate the treatment effects heterogeneity. The average treatment effect μ is estimated by $\sum_{k=1}^{K}(\sigma_k^2 + \hat{\tau}^2)^{-1}Y_k / \sum_{k=1}^{K}(\sigma_k^2 + \hat{\tau}^2)^{-1}$, where $\hat{\tau}^2$ is an estimator of the heterogeneity parameter τ^2. Estimators of τ^2, such as the DerSimonian and Laird estimator [18], have been proposed by a number of researchers [22]. In this paper, we shall discuss prediction intervals using the DerSimonian and Laird estimator that is defined as $\hat{\tau}^2_{DL} = \max[0, \{Q - (K-1)\}/(S_1 + S_2/S_1)]$, with its untruncated version defined as $\hat{\tau}^2_{UDL} = \{Q - (K-1)\}/(S_1 + S_2/S_1)$, where $Q = \sum_{k=1}^{K} v_k(Y_k - \bar{Y})^2$ is Cochran's Q statistic, $v_k = \sigma_k^{-2}$, $\bar{Y} = \sum_{k=1}^{K} v_k Y_k / \sum_{k=1}^{K} v_k$, and $S_r = \sum_{k=1}^{K} v_k^r$ for $r = 1, 2$. Under Condition 1, Biggerstaff and Jackson [21] derived the exact distribution function of Q, $F_Q(q; \tau^2)$, to obtain confidence intervals for τ^2. Cochran's Q is a quadratic form that can be written $\mathbf{Y}^{\mathrm{T}}\mathbf{A}\mathbf{Y}$, where $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_K)^{\mathrm{T}}$, $\mathbf{A} = \mathbf{V} - \mathbf{v}\mathbf{v}^{\mathrm{T}}/v_+$, $\mathbf{V} = \mathrm{diag}(v_1, v_2, \ldots, v_K)$, $\mathbf{v} = (v_1, v_2, \ldots, v_K)^{\mathrm{T}}$, $v_+ = \sum_{k=1}^{K} v_k$, and the superscript 'T' denotes matrix transposition. Here and subsequently, $\mathbf{Z} = \mathbf{\Sigma}^{-1/2}(\mathbf{Y} - \boldsymbol{\mu}) \sim N(\mathbf{0}, \mathbf{I})$, $\mathbf{S} = \mathbf{\Sigma}^{1/2}\mathbf{A}\mathbf{\Sigma}^{1/2}$, $\boldsymbol{\mu} = (\mu, \mu, \ldots, \mu)^{\mathrm{T}}$, $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1^2 + \tau^2, \sigma_2^2 + \tau^2, \ldots, \sigma_K^2 + \tau^2)$, $\mathbf{0} = (0, 0, \ldots, 0)^{\mathrm{T}}$, and $\mathbf{I} = \mathrm{diag}(1, 1, \ldots, 1)$.

Lemma 1. *Under Condition 1, Q can be expressed as $\mathbf{Z}^{\mathrm{T}}\mathbf{S}\mathbf{Z}$; then Q has the same distribution as the random variable $\sum_{k=1}^{K} \lambda_k \chi_k^2(1)$ where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{K-1} > \lambda_K = 0$ are the ordered eigenvalues of the matrix $\mathbf{S}$, and $\chi_1^2(1), \chi_2^2(1), \ldots, \chi_K^2(1)$ are K independent central chi-square random variables each with one degree of freedom.*

Lemma 1 was proven by Biggerstaff and Jackson [21] using the location invariantness of Q (e.g., Q can be decomposed as $\sum_{k=1}^{K} v_k(Y_k - \mu)^2 - v_+(\bar{Y} - \mu)^2$), and distribution theories of quadratic forms in normal variables that have been extensively studied in the literature [23, 24, 25].

2.2 The Higgins–Thompson–Spiegelhalter prediction interval

The HTS prediction interval was proposed by Higgins *et al.* [6]. Suppose τ^2 is known, $\hat{\mu} \sim N(\mu, \mathrm{SE}[\hat{\mu}])$ and the observation in a future study $\theta_{new} \sim N(\mu, \tau^2)$, where $\mathrm{SE}[\hat{\mu}] = \sqrt{1/\sum_{k=1}^{K} w_k}$ is a standard error of $\hat{\mu}$ given τ^2, and $w_k = (\sigma_k^2 + \tau^2)^{-1}$. Assuming independence of θ_{new} and $\hat{\mu}$ given μ, $\theta_{new} - \mu \sim N(0, \tau^2 + \mathrm{SE}[\hat{\mu}])$. Since τ^2 is unknown, it should be replaced by an estimator $\hat{\tau}^2_{DL}$. If $(K-2)(\hat{\tau}^2_{DL} + \hat{\mathrm{SE}}[\hat{\mu}])/(\tau^2 + \mathrm{SE}[\hat{\mu}])$ is approximately distributed as $\chi^2(K-2)$, then $(\theta_{new} - \hat{\mu})/\sqrt{\hat{\tau}^2_{DL} + \hat{\mathrm{SE}}[\hat{\mu}]} \sim t(K-2)$, where $\hat{\mathrm{SE}}[\hat{\mu}] = \sqrt{1/\sum_{k=1}^{K} \hat{w}_k}$ is the standard error estimator of $\hat{\mu}$, and $\hat{w}_k = (\sigma_k^2 + \hat{\tau}^2_{DL})^{-1}$. By this approximation, the HTS prediction interval is obtained by

$$\left[\hat{\mu} - t^{\alpha}_{K-2}\sqrt{\hat{\tau}^2_{DL} + \hat{\mathrm{SE}}[\hat{\mu}]^2},\ \hat{\mu} + t^{\alpha}_{K-2}\sqrt{\hat{\tau}^2_{DL} + \hat{\mathrm{SE}}[\hat{\mu}]^2}\right],$$

where t^{α}_{K-2} is the $100(1-\alpha/2)$ percentile of the t distribution with $K-2$ degrees of freedom. However, the t-approximation is clearly inappropriate, and has a detrimental impact on the coverage probability.

Several HTS-type prediction intervals following restricted maximum likelihood (REML) estimation of τ^2 have been proposed by Partlett and Reliy[10]. For example, they discussed a HTS-type prediction

interval following REML with the Hartung–Knapp variance estimator [26] (HTS-HK) that is defined as

$$\left[\hat{\mu}_R - t^{\alpha}_{K-2}\sqrt{\hat{\tau}^2_R + \hat{\mathrm{SE}}_{HK}[\hat{\mu}_R]^2},\ \hat{\mu}_R + t^{\alpha}_{K-2}\sqrt{\hat{\tau}^2_R + \hat{\mathrm{SE}}_{HK}[\hat{\mu}_R]^2}\right],$$

and a HTS-type prediction interval following REML with the Sidik–Jonkman bias-corrected variance estimator [27] (HTS-SJ) that is defined as

$$\left[\hat{\mu}_R - t^{\alpha}_{K-2}\sqrt{\hat{\tau}^2_R + \hat{\mathrm{SE}}_{SJ}[\hat{\mu}_R]^2},\ \hat{\mu}_R + t^{\alpha}_{K-2}\sqrt{\hat{\tau}^2_R + \hat{\mathrm{SE}}_{SJ}[\hat{\mu}_R]^2}\right],$$

where $\hat{\tau}^2_R$ is the REML estimator for the heterogeneity varinace [28, 29, 22] which is an iterative solution of the equation

$$\hat{\tau}^2_R = \frac{\sum_{k=1}^K \hat{w}^2_{R,k}\{(Y_k - \hat{\mu}_R)^2 + 1/\sum_{l=1}^K \hat{w}_{R,k} - \sigma^2_k\}}{\sum_{k=1}^K \hat{w}^2_{R,k}},$$

$\hat{w}_{R,k} = (\sigma^2_k + \hat{\tau}^2_R)^{-1}$, $\hat{\mu}_R = \sum_{k=1}^K \hat{w}_{R,k}Y_k/\sum_{k=1}^K \hat{w}_{R,k}$, the Hartung–Knapp variance estimator is defined as

$$\hat{\mathrm{SE}}_{HK}[\hat{\mu}_R]^2 = \frac{1}{K-1}\sum_{k=1}^K \frac{\hat{w}_{R,k}(Y_k - \hat{\mu}_R)^2}{\sum_{l=1}^K \hat{w}_{R,l}},$$

the Sidik–Jonkman bias-corrected variance estimator

$$\hat{\mathrm{SE}}_{SJ}[\hat{\mu}_R]^2 = \frac{\sum_{k=1}^K \hat{w}^2_{R,k}(1-\hat{h}_k)^{-1}(Y_k - \hat{\mu}_R)^2}{(\sum_{k=1}^K \hat{w}_{R,k})^2},$$

and $\hat{h}_k = 2\hat{w}_{R,k}/\sum_{k=1}^K \hat{w}_{R,k} - \sum_{k=1}^K\{\hat{w}^2_{R,k}(\sigma^2_k + \hat{\tau}^2_R)\}/\{(\sigma^2_k + \hat{\tau}^2_R)\sum_{k=1}^K \hat{w}^2_{R,k}\}$. The HTS-HK and HTS-SJ prediction intervals can have a superior performance to other methods discussed in Partlett and Reliy[10] for a large heterogeneity variance and $K \geq 5$.

However, the HTS prediction intervals could have severe under-coverage under certain conditions (see Section 3 and Partlett and Reliy[10]). The uncertainty of the estimator of τ^2 should be more precisely taken into account. Therefore, we consider a new prediction interval that is valid under a small number of studies.

2.3 The proposed prediction interval

As an alternative approach to address the issue discussed in Section 2.2, we propose a new prediction interval. The proposed method starts with assumptions that differ from those of Higgins *et al.* [6] in order to address a small number of studies, and accounts for the uncertainty of $\hat{\tau}^2_{DL}$ via a parametric bootstrap with the exact distribution of $\hat{\tau}^2_{UDL}$ by using a confidence distribution (see Section 2.4).

From now on we make the following assumptions: Let the observation in a future study $\theta_{new} \sim N(\mu, \tau^2)$, $Y_k \sim N(\mu, \sigma^2_k + \tau^2)$ given σ^2_k and τ^2, and θ_{new} and $\bar{\mu} = \sum_{k=1}^K w_k Y_k/\sum_{k=1}^K w_k$ are independent.

In Hartung [30] and Hartung and Knapp [26], it was shown that assuming normality of Y_k, $(\mu - \bar{\mu})/\mathrm{SE}_H[\bar{\mu}]$ is t-distributed with $K-1$ degrees of freedom, and $\mathrm{SE}_H[\bar{\mu}]$ is stochastically independent of $\bar{\mu}$, where $\mathrm{SE}_H[\bar{\mu}]^2 = \frac{1}{K-1}\sum_{k=1}^K \frac{w_k}{w_+}(Y_k - \bar{\mu})^2$, and $w_+ = \sum_{k=1}^K w_k$. By replacing τ^2 in $(\bar{\mu} - \mu)/\mathrm{SE}_H[\bar{\mu}]$ with an appropriate estimate $\hat{\tau}^2$, $(\hat{\mu} - \mu)/\hat{\mathrm{SE}}_H[\hat{\mu}]$ is approximately t-distributed with $K-1$ degrees of freedom, where $\hat{\mathrm{SE}}_H[\hat{\mu}]^2 = \frac{1}{K-1}\sum_{k=1}^K \frac{\hat{w}_k}{\hat{w}_+}(Y_k - \hat{\mu})^2$, and $\hat{w}_+ = \sum_{k=1}^K \hat{w}_k$.

The above assumptions and results lead to a system of equations,

$$\begin{cases} \dfrac{\theta_{new} - \mu}{\tau} = Z \\ \dfrac{\bar{\mu} - \mu}{\mathrm{SE}_H[\bar{\mu}]} = t_{K-1} \end{cases}, \tag{2}$$

where $Z \sim N(0,1)$ and $t_{K-1} \sim t(K-1)$. Solving for θ_{new} in (2) yields

$$\bar{C} = \bar{\mu} + Z\tau - t_{K-1}\mathrm{SE}_H[\bar{\mu}], \tag{3}$$

and the prediction distribution has the same distribution as the statistic $\bar{C}$. By replacing τ^2 in (3) with an appropriate estimator (not an estimate), we have

$$\hat{C} = \hat{\mu} + Z\hat{\tau}_{UDL} - t_{K-1}\hat{\mathrm{SE}}_H[\hat{\mu}],$$

and an approximate prediction distribution can be given by the distribution of $\hat{C}$. We use the untruncated estimator $\hat{\tau}^2_{UDL}$ here, because we do not need the truncation to consider the distribution of an estimator of τ^2. Hence, $\Pr(c_l < \theta_{new} < c_u) = 1 - \alpha$ can be evaluated by the distribution of $\hat{C}$. Since $\hat{C}$ includes three random components, $\hat{\tau}^2_{UDL}$, Z, and t_{K-1}, this gives the following algorithm for the proposed prediction interval.

Algorithm 1. An algorithm for the proposed prediction interval.

1. Generate B bootstrap samples $\tilde{\tau}^2_b$ ($b = 1, \ldots, B$) that are drawn from the exact distribution of $\hat{\tau}^2_{UDL}$, z_b that are drawn from $N(0,1)$, and t_b that are drawn from $t(K-1)$.

2. Calculate $\tilde{\mu}_b = \sum_{k=1}^K \tilde{w}_{bk} y_k / \sum_{k=1}^K \tilde{w}_{bk}$, and $\tilde{C}_b = \tilde{\mu}_b + z_b\tilde{\tau}_b - t_b\tilde{\mathrm{SE}}_{H,b}[\tilde{\mu}_b]$, where $\tilde{w}_{bk} = (\sigma^2_k + \tilde{\tau}^2_b)^{-1}$, $\tilde{\mathrm{SE}}_{H,b}[\tilde{\mu}_b]^2 = \frac{1}{K-1}\sum_{k=1}^K \frac{\tilde{w}_{bk}}{\tilde{w}_{b+}}(y_k - \tilde{\mu}_b)^2$, and $\tilde{w}_{b+} = \sum_{k=1}^K \tilde{w}_{bk}$.

3. Calculate the prediction limits c_l and c_u that are $100 \times \alpha/2$ and $100 \times (1 - \alpha/2)$ percentage points of $\tilde{C}_b$, respectively.

However, an algorithm for sampling from the exact distribution of $\hat{\tau}^2_{UDL}$ has not been studied. We will discuss below the exact distribution of $\hat{\tau}^2_{UDL}$ and a sampling method from the exact distribution.

An R package implementing the new method with the three data sets (see Section 4) and a documentation may be available at the publisher's web-site, and will be published on the CRAN website (https://cran.r-project.org/) and GitHub (https://github.com/nshi-stat/pimeta/).

2.4 Sampling from the exact distribution of the estimator of τ^2

In frequentist inference, sometimes a distribution function like a Bayesian posterior is needed to estimate a parameter of interest. Confidence distribution is an appropriate solution in this situation. Confidence distribution is a distribution estimator that can be defined and interpreted in a frequentist framework in which the parameter is a non-random quantity. A confidence distribution of the parameter of interest, as described below, can be easily defined by the cumulative distribution function of a statistic, which includes the parameter of interest. Confidence distribution has a theoretical relationship to the fiducial approach [31], and recent developments [11, 12, 13, 14, 15] provide useful statistical tools that are more widely applicable than the previous method. For example, Efron's bootstrap distribution [32] is a confidence

distribution and a distribution estimator of a parameter of interest. In meta-analysis, the Q-profile method for an approximate confidence interval for τ^2 [33] can be considered as an application of confidence distribution [12]. In this section, we propose the exact distribution of $\hat{\tau}^2_{UDL}$, which is a distribution function for estimating the parameter τ^2 using a confidence distribution, and then develop a method of sampling from the exact distribution. A useful theorem (Theorem 1) is proved that provides conditions in the case of a statistic with a continuous cumulative distribution function.

The following definition of a confidence distribution was presented in [15]. In the definition, Φ is the parameter space of the unknown parameter of interest ϕ, $\mathbf{Y}$ is a random vector, and $\mathcal{Y}$ is the sample space corresponding to sample data $\mathbf{y} = (y_1, y_2, \ldots, y_K)^{\mathrm{T}}$.

Definition 1. (R1) A function $H(\cdot) = H(\mathbf{y}, \phi)$ on $\mathcal{Y} \times \Phi \to [0, 1]$ is called a confidence distribution for a parameter ϕ; (R2) If for each given $\mathbf{y} \in \mathcal{Y}$, $H(\cdot)$ is a cumulative distribution function on ϕ; (R3) At the true parameter value $\phi = \phi_0$, $H(\phi_0) \equiv H(\mathbf{y}, \phi_0)$, as a function of the sample $\mathbf{y}$, follows the uniform distribution $U(0, 1)$.

Theorem 1. *If a cumulative distribution function of a statistic, $T(\mathbf{Y})$, is $F_T(T(\mathbf{y}); \phi) \equiv F_T(T(\mathbf{Y}) \leq T(\mathbf{y}); \phi)$, and F_T is a continuous and strictly monotonic (without loss of generality, assume it is decreasing) function in ϕ with the parameter space $\Phi = \{\phi : \phi_{\min} \leq \phi \leq \phi_{\max}\}$ for each sample $\mathbf{y}$, then $H(\phi) = 1 - F_T(T(\mathbf{y}); \phi)$ is a confidence distribution for ϕ that satisfies the requirements in Definition 1.*

Lemma 2. *Under Condition 1, $H(\tau^2) = 1 - F_Q(q; \tau^2)$ is a confidence distribution for τ^2.*

Technical proofs are collected in Appendix I. Lemma 2 can be easily proved by using Theorem 1.

It follows from the above that we propose an algorithm of sampling from the confidence distribution, $H(\tau^2) = 1 - F_Q(q_{obs}; \tau^2)$, where q_{obs} is the observed value of Q. By applying Lemma 2 and the inverse transformation method, if U is distributed as $U(0, 1)$ then $H^{-1}(U)$ follows the distribution $H(\tau^2)$. A sample $\tilde{\tau}^2 = H^{-1}(u)$ can be computed by a numerical inversion [34] of $H(\tilde{\tau}^2) = u$, where u is an observed value of the random variable U. If $H(0) > u$, then the sample is truncated to zero ($\tilde{\tau}^2 = 0$). It follows from Lemma 1 that $F_Q(q; \tau^2)$ is the distribution function of a positive linear combination of χ^2 random variables. It can be calculated with the Farebrother's algorithm [35].

3. Simulations

We assessed the empirical properties of the HTS and the proposed prediction intervals via simulation studies.

Simulation data was generated by the random-effects model (1), assuming independent normal errors $\epsilon_k \sim N(0, \sigma_k^2)$ and $u_k \sim N(0, \tau^2)$. We conducted two sets of simulations described below.

(i) By reference to Brockwell and Gordon [7, 36] and Jackson [37], parameter settings that mimic meta-analyses for estimating an overall mean log odds-ratio were determined in simulation (i). The average treatment effect μ was fixed at 0, because the coverage probability is not dependent on the value of μ. The across-studies variance was set to $\tau^2 = 0.01, 0.05, 0.1, 0.2, 0.3, 0.4$, or 0.5 [38, 39]. The within-studies variances σ_k^2 were generated from a scaled χ^2 distribution with one degree of freedom, multiplied by 0.25, then truncated to lie within $[0.009, 0.6]$. The number of studies was set to $K = 3, 5, 7, 10$, or 25.

(ii) In reference to Partlett and Reliy[10], parameter settings were determined to evaluate the empirical performance of prediction intervals under various relative degrees of heterogeneity scenarios in

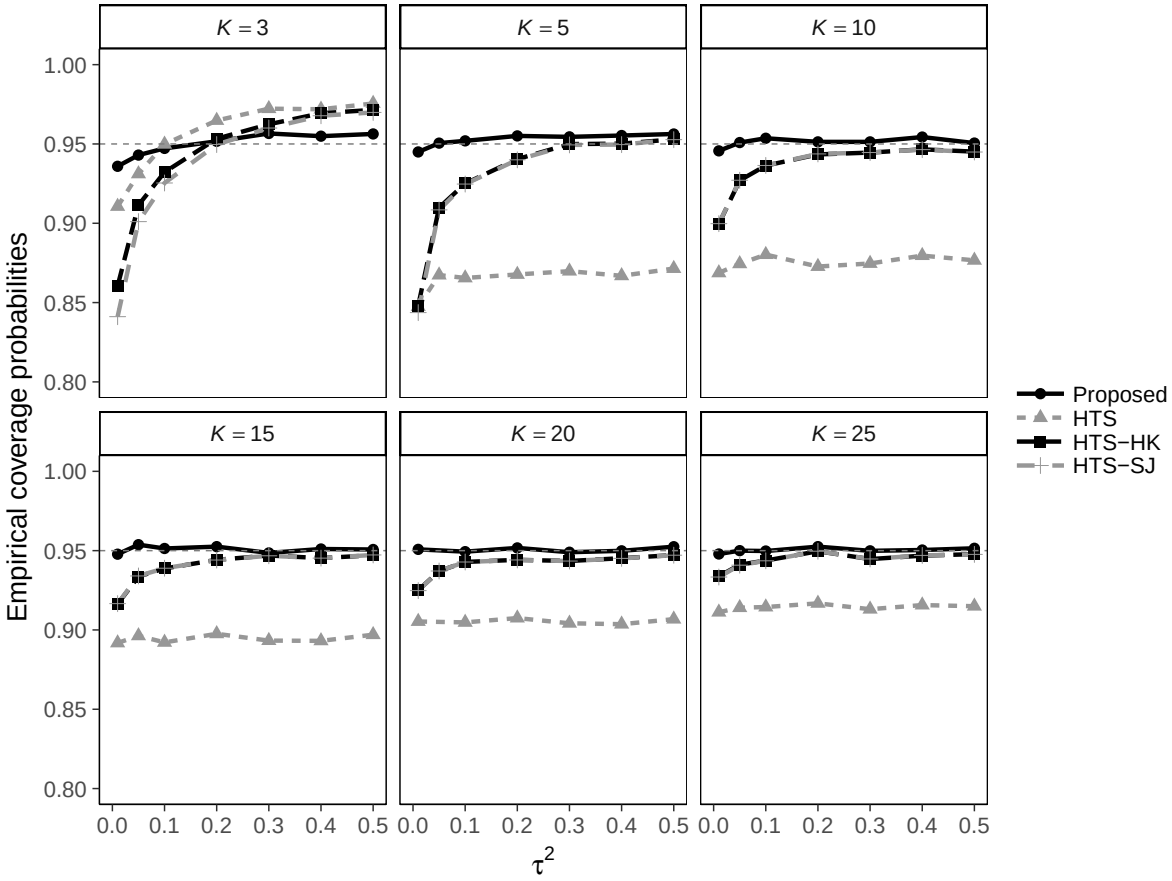

Figure 1. Simulation results (i): the performance of the HTS and proposed prediction intervals. The number of studies $K = 3, 5, 10, 15, 20$, or 25. The number of simulations was 25 000. Methods: Proposed, the proposed prediction interval; HTS, the HTS prediction interval; HTS-HK, the HTS-type prediction interval following REML with the Hartung–Knapp variance estimator; HTS-SJ, the HTS-type prediction interval following REML with the Sidik–Jonkman bias-corrected variance estimator.

simulation (ii). The within-studies variances σ_k^2 were generated from $\sigma^2\chi^2(n-1)/(n-1)$, an average within-study variance was set to $\sigma^2 = 0.1$, and the study sample size was set to $n = 30$, where $\chi^2(n-1)$ is a random number from a χ^2 distribution with $n-1$ degrees of freedom. The degree of heterogeneity is controlled using the ratio τ^2/σ^2. The heterogeneity parameter was set to $\tau^2 = 0.001, 0.01, 0.1$, or 1, which corresponds to $\tau^2/\sigma^2 = 0.01, 0.1, 1$, or 10. The average treatment effect μ was fixed at 1. The number of studies was set to $K = 3, 5, 10, 15, 20$, or 25.

For each setting, we simulated 25 000 replications. For each method, two-tailed 95% prediction intervals were calculated. The number of bootstrap samples B was set to 5 000. The coverage probability was estimated by the proportion of simulated prediction intervals containing the result of a future study θ_{new} that was generated from a normal distribution $N(\mu, \tau^2)$.

The results of simulation (i) are presented in Figure 1. The HTS prediction interval could not achieve the nominal level of 95%, and the coverage probabilities were around 90%. Since most meta-analyses include less than 20 studies [9], the coverage performance of the HTS prediction interval is therefore insufficient. The under-coverage of the HTS prediction interval reflects the rough t-approximation; in other words, the uncertainty of $\hat{\tau}^2_{DL}$ is ignored in the HTS prediction interval. The results show that the HTS-HK and HTS-SJ prediction intervals have similar performance. The coverage probabilities for the HTS-HK

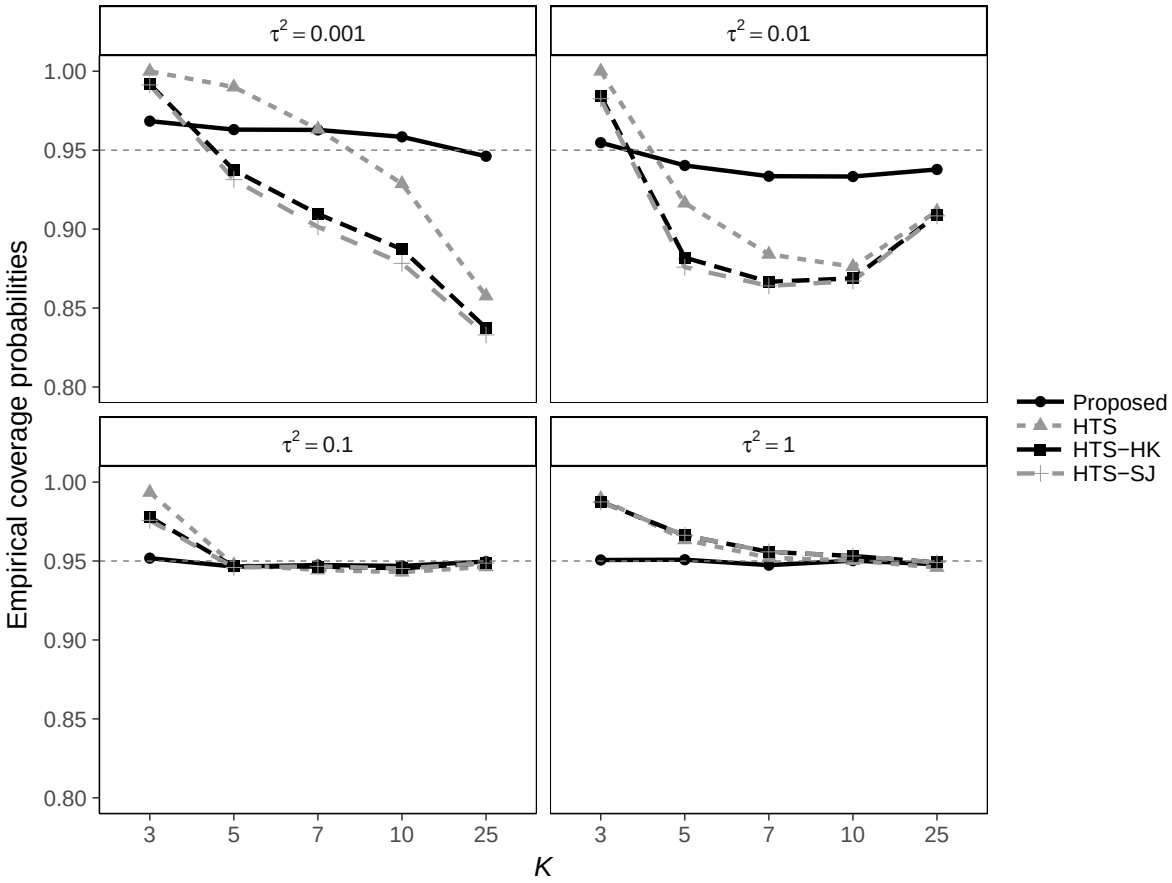

Figure 2. Simulation results (ii): the performance of the HTS and proposed prediction intervals. The heterogeneity parameters $\tau^2 = 0.001, 0.01, 0.1$, or 1. The number of simulations was 25 000. Methods: Proposed, the proposed prediction interval; HTS, the HTS prediction interval; HTS-HK, the HTS-type prediction interval following REML with the Hartung–Knapp variance estimator; HTS-SJ, the HTS-type prediction interval following REML with the Sidik–Jonkman bias-corrected variance estimator.

and HTS-SJ prediction intervals almost retained the nominal level except in situations where the relative degree of heterogeneity is small or moderate. For example, the coverage probabilities of the HTS-HK prediction interval were 86.0%–93.3% for $\tau^2 = 0.01$, 91.0%–94.1% for $\tau^2 = 0.05$, and 92.5%–94.3% for $\tau^2 = 0.1$; the coverage probabilities of the HTS-SJ prediction interval were 84.1%–93.3% for $\tau^2 = 0.01$, 90.1%–94.1% for $\tau^2 = 0.05$, and 92.5%–94.4% for $\tau^2 = 0.1$. By contrast, the coverage probabilities for the proposed prediction interval almost always retained the nominal level. The only exception was when $K = 3$ and $\tau^2 = 0.01$, where the coverage probability for the proposed prediction interval was 93.6%, which was slightly below the nominal level. However, in this case, the coverage probability for the HTS, HTS-HJ, and HTS-SJ prediction intervals were even smaller, at 91.1%, 86.0%, and 84.1%, respectively. Analyses using a very small numbers of studies ($K < 5$) pose problems for random-effects models, as discussed by Higgins *et al.* [6]. Nevertheless, the proposed method performed well even when $K = 3$. It should be noted that, the nominal level was attained for almost all values of the heterogeneity parameter in the proposed prediction interval, and this parameter had very little effect on the interval's performance.

The results of simulation (ii) are presented in Figure 2. The results show that all HTS prediction intervals have similar performance except for $\tau^2 = 0.001$. The coverage probabilities for all HTS prediction intervals almost retained the nominal level for $\tau^2 \geq 0.1$ and $K \geq 5$. The coverage probabilities were too large for $K = 3$ and too small for $K \geq 5$ and $\tau^2 = 0.01$. In the case of $\tau^2 = 0.01$, the coverage probabilities of the HTS-HJ and HTS-SJ prediction intervals were too small for $K \geq 7$, and the coverage probability of the HTS prediction interval was too small for $K = 25$. By contrast, the coverage probabilities for the proposed prediction interval almost always retained the nominal level. The only exception was when $\tau^2 = 0.01$, where the coverage probabilities for the proposed prediction interval were 93.3%–95.5%, which was slightly below the nominal level.

In summary, the HTS prediction intervals had poor coverage performance except in situations where the relative degree of heterogeneity is large, and may show severe under-coverage under realistic meta-analysis settings involving medical research, possibly providing misleading results and interpretations. By contrast, since the proposed prediction interval could mostly achieve the nominal level, it can be recommended in practice.

4. Applications

We applied the methods to three published random-effects meta-analyses. The three data sets were

(A) Set-shifting data: Higgins *et al.* [6] re-analyzed data [40] that included 14 studies evaluating set-shifting ability in people with eating disorders by using a prediction interval.

(B) Pain data: Riley *et al.* [4] conducted a random-effects meta-analysis of published data [41]. The pain data included 22 studies comparing the treatment effect of antidepressants on reducing pain in patients with fibromyalgia syndrome.

(C) Systolic blood pressure (SBP) data: Riley *et al.* [4] analyzed a hypothetical meta-analysis with a random-effects model. They supposed that the data included 10 studies of the same antihypertensive drug.

These data sets are reproduced in Figure 3. The number of bootstrap samples B was set to 50 000.

Table 1 presents the estimated results for the average treatment effect and its confidence interval, heterogeneity measures, the P-value for the test of heterogeneity, the proposed prediction interval, and the HTS prediction intervals. None of the confidence intervals for the average treatment effect included 0 (Set-shifting data: $[0.19, 0.53]$; Pain data: $[-0.55, -0.30]$; SBP data: $[-0.48, -0.18]$). This means that on

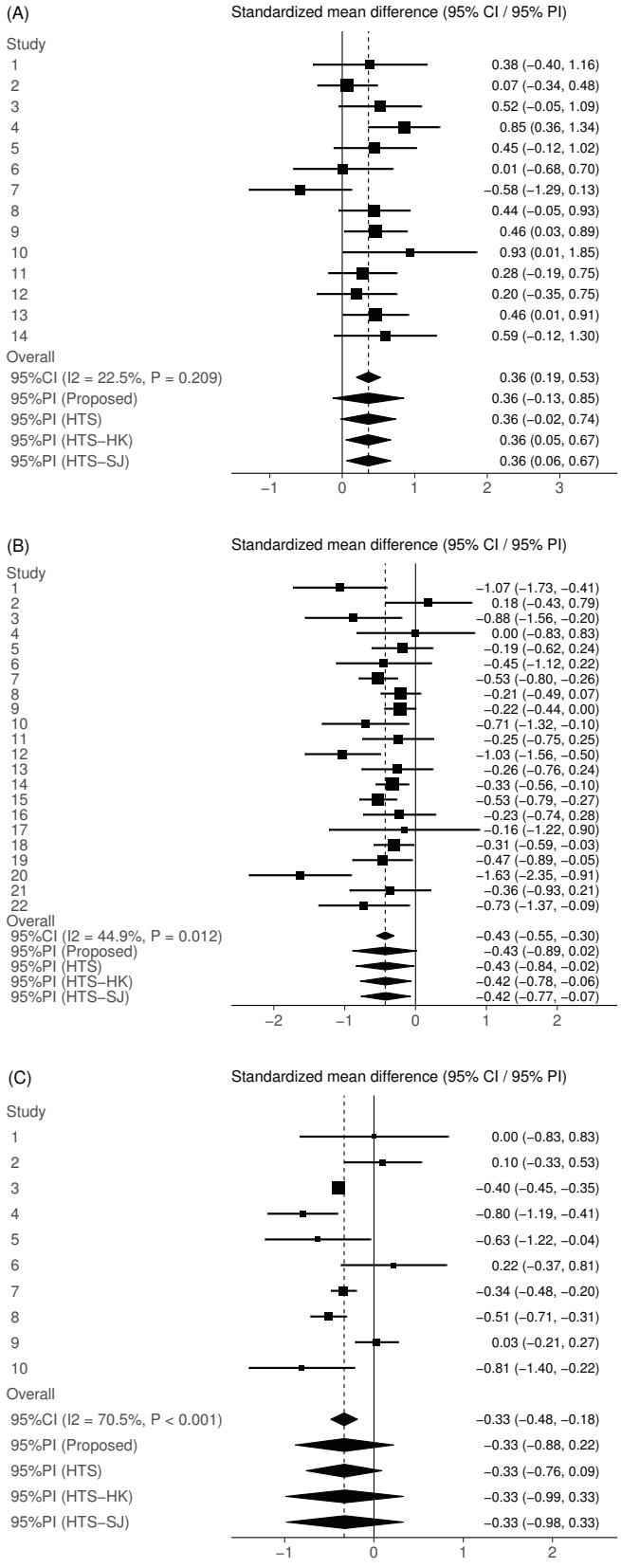

Figure 3. The three data sets and summary results: (A) Set-shifting data [40] ($K = 14$), (B) Pain data [41] ($K = 22$), and (C) SBP data [4] ($K = 10$). Abbreviations: CI, confidence interval; PI, prediction interval.

Table 1. Results from the three data sets: the average treatment effect ($\hat{\mu}$) and its 95% confidence interval, heterogeneity measures ($\hat{\tau}^2_{DL}$, $\hat{\tau}^2_{REML}$ and I^2), the P-value for the test of heterogeneity, the proposed prediction interval, and the HTS prediction intervals.

Data		Set-shifting ($K = 14$)	Pain ($K = 22$)	SBP ($K = 10$)
$\hat{\mu}$ (DL)		0.36	-0.43	-0.33
95%CI (DL)		[0.19, 0.53]	[-0.55, -0.30]	[-0.48, -0.18]
$\hat{\tau}^2_{DL}$		0.023	0.034	0.023
$\hat{\tau}^2_{R}$		0.013	0.025	0.070
I^2 (DL)		22.5%	44.9%	70.5%
P-value for heterogeneity		0.209	0.012	<0.001
95%PI	Proposed	[-0.13, 0.85]	[-0.89, 0.02]	[-0.88, 0.23]
	HTS	[-0.02, 0.74]	[-0.84, -0.02]	[-0.76, 0.09]
	HTS-HK	[0.05, 0.67]	[-0.78, -0.06]	[-0.99, 0.33]
	HTS-SJ	[0.06, 0.67]	[-0.77, -0.07]	[-0.98, 0.33]

average the interventions are significantly effective. However, substantial between-studies heterogeneities were observed in the three data sets (Set-shifting data: $\hat{\tau}^2_{DL} = 0.023$, $I^2 = 22.5\%$; Pain data: $\hat{\tau}^2_{DL} = 0.185$, $I^2 = 44.9\%$; SBP data: $\hat{\tau}^2_{DL} = 0.023$, $I^2 = 70.5\%$). Accounting for the heterogeneities, prediction intervals would provide additional relevant statistical information.

As shown in Figure 3 and summarized in Table 1, the proposed prediction intervals (Set-shifting data: $[-0.13, 0.85]$, length = 0.98; Pain data: $[-0.89, 0.02]$, length =0.91; SBP data: $[-0.88, 0.22]$, length = 1.10) were consistently wider than the HTS prediction intervals (Set-shifting data: $[-0.02, 0.74]$, length = 0.76; Pain data: $[-0.84, -0.02]$, length = 0.82; SBP data: $[-0.76, 0.09]$, length = 0.85). The lengths of the proposed prediction intervals were 29%, 11%, and 31% wider than the HTS prediction intervals for the Set-shifting data, Pain data, and SBP data, respectively. The HTS-HK (Set-shifting data: $[0.05, 0.67]$, length = 0.62; Pain data: $[-0.78, -0.06]$, length =0.72; SBP data: $[-0.99, 0.33]$, length = 1.32) and HTS-SJ (Set-shifting data: $[0.06, 0.67]$, length = 0.61; Pain data: $[-0.77, -0.07]$, length =0.72; SBP data: $[-0.98, 0.33]$, length = 1.31) prediction intervals give similar results.

The prediction intervals may lead to different interpretations of the results. Especially in the Pain data, the HTS prediction intervals did not include 0, meaning that the intervention may be beneficial in most subpopulations. On the other hand, the proposed prediction interval included 0, which indicates that the intervention may not be beneficial in some subpopulations. The simulation results in Section 3 suggest that the HTS prediction intervals could have under-coverage in situations where the relative degree of heterogeneity is small or moderate. Since $\hat{\tau}^2_{DL}$ of three data sets and $\hat{\tau}^2_{R}$ of Set-shifting and Pain data were small (≈ 0.02), it may be too narrow under realistic situations and may provide misleading results. By contrast, our proposed method enables adequate evaluations of the statistical error in the predictive inference.

5. Discussion and conclusion

For the random-effects model in meta-analysis, the average treatment effect and its confidence interval have been used with heterogeneity measures such as the I^2-statistic and τ^2. However, results from random-effects models have sometimes been misinterpreted. Thus, the new concept "prediction interval" was proposed, which is useful in applying the results to other subpopulations and in decision making. The

HTS prediction intervals have a theoretical problem, namely that its rough t-approximation could have a detrimental impact on the coverage probability. Therefore, we have presented an appropriate prediction interval to account for the uncertainty of τ^2 by using a confidence distribution. We also proved a useful theorem for applying confidence distribution.

Simulation studies showed that the HTS prediction intervals could have severe under-coverage for realistic meta-analysis settings and might lead to misleading results and interpretations. The simulation results suggested that the HTS prediction interval may be too narrow when considering a small number of studies. This interval would be valid if $K \gg 25$, but such a large number of studies can rarely be expected in common meta-analysis settings. The HTS-HK and HTS-SJ prediction intervals may be too narrow when the relative degree of heterogeneity is small. By contrast, the coverage probabilities for the proposed prediction interval satisfactorily retained the nominal level. Although Higgins *et al.* [6] cautioned that the random-effects model may not work well under very small numbers of studies ($K < 5$), the proposed method performed well even when $K = 3$. Since the heterogeneity parameter had very little effect on the performance of the proposed prediction interval, the method would be valid regardless of the value of the heterogeneity parameter.

Applications to the three published random-effects meta-analyses concluded that substantially different results and interpretations might be obtained from the prediction intervals. Since the HTS prediction interval is always narrower and the HTS-HK and HTS-SJ prediction intervals are narrower when the heterogeneity parameter is small or moderate, we should be cautious in using and interpreting these approaches.

In conclusion, we showed that the proposed prediction interval works well and can be recommended for random-effects meta-analysis in practice. As shown in the three illustrative examples, quite different results and interpretations might be obtained with our new method. Extensions of these results to other complicated models such as network meta-analysis are now warranted.

Appendix I. Proofs of Theorem 1 and Lemma 2

Proof of Theorem 1. (R1) Since F_T is a continuous distribution function, $H(\phi) = 1 - F_T(T(\mathbf{y}); \phi)$ is continuous on $\mathbf{Y} \times \Phi \to [0, 1]$. (R2) By the continuity of F_T, a derivative, $g(\phi) = \mathrm{d}F_T(T(\mathbf{y}); \phi)/\mathrm{d}\phi$, exists, and $G(\phi) = \int g(\phi)\,\mathrm{d}\phi = F_T(T(\mathbf{y}); \phi)$. By (R1) and the monotone decreasingness of F_T, $G(\phi_{\min}) = 1$ and $G(\phi_{\max}) = 0$. Therefore, $H(\phi)$ can be written as $1 - \int_{\phi}^{\phi_{\max}} -g(s)\,\mathrm{d}s = 1 - G(\phi)$. Writing $h(\phi) = -g(\phi)$, we find $1 - \int_{\phi}^{\phi_{\max}} h(s)\,\mathrm{d}s = \int_{\phi_{\min}}^{\phi} h(s)\,\mathrm{d}s$. Thus, $H(\phi)$ is clearly a cumulative distribution function on ϕ. (R3) At the true parameter value $\phi = \phi_0$, it follows that $1 - F_T(T(\mathbf{y}); \phi_0) \sim U(0, 1)$. Thus, by Definition 1, $H(\phi)$ is a confidence distribution for the parameter ϕ, and $h(\phi)$ is a confidence density function for ϕ. □

Proof of Lemma 2. From Lemma 1, the cumulative distribution function of Q is $F_Q(q; \tau^2)$. $F_Q(q; \tau^2)$ is a continuous and strictly decreasing function in τ^2 [37]. Note that we have considered the untruncated version of an estimator of τ^2 with the parameter space $\Phi = [\tau^2_{\min}, \infty]$, and $\tau^2_{\min}$ can be negative. Applying Theorem 1, we easily show that $H(\tau^2) = 1 - F_Q(q; \tau^2)$ is a confidence distribution for τ^2. □

Acknowledgements

This study was supported by CREST from the Japan Science and Technology Agency (Grant number: JPMJCR1412).

References

[1] Borenstein M, Hedges LV, Higgins JPT, et al. *Introduction to Meta-Analysis*. Chichester: Wiley, 2009.

[2] Higgins JPT and Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med* 2002; **21**: 1539–1558.

[3] Higgins JPT, Thompson SG, Deeks JJ, et al. Measuring inconsistency in meta-analyses. *BMJ* 2003; **327**: 557–560.

[4] Riley RD, Higgins JPT and Deeks JJ. Interpretation of random effects meta-analyses. *BMJ* 2011; **342**: d549.

[5] Riley RD, Gates SG, Neilson J, et al. Statistical methods can be improved within Cochrane pregnancy and childbirth reviews. *J Clin Epidemiol* 2011; **64**: 608–618.

[6] Higgins JPT, Thompson SG and Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *J R Stat Soc Ser A Stat Soc* 2009; **172**: 137–159.

[7] Brockwell SE and Gordon IR. A comparison of statistical methods for meta-analysis. *Stat Med* 2001; **20**: 825–840.

[8] Noma H. Confidence intervals for a random-effects meta-analysis based on Bartlett-type corrections. *Stat Med* 2011; **30**: 3304–3312.

[9] Kontopantelis E, Springate DA and Reeves D. A re-analysis of the Cochrane library data: the dangers of unobserved heterogeneity in meta-analyses. *PLoS One* 2013; **8**: e69930.

[10] Partlett C and Riley RD. Random effects meta-analysis: Coverage performance of 95% confidence and prediction intervals following REML estimation. *Stat Med* 2017; **6**: 301–317.

[11] Schweder T and Hjort NL. Confidence and likelihood. *Scand J Stat* 2002; **29**: 309–332.

[12] Schweder T and Hjort NL. *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*. New York: Cambridge University Press, 2016.

[13] Singh K, Xie M and Strawderman WE. Combining information from independent sources through confidence distributions. *Ann Stat* 2005; **33**: 159–183.

[14] Singh K, Xie M and Strawderman WE. Confidence distribution (CD)-distribution estimator of a parameter. In *Complex Datasets and Inverse Problems: Tomography, Networks, and Beyond*. Liu R, Strawderman WE and Zhang CH (eds). Beachwood: Institute of Mathematical Statistics, 2007; **54**: 132–150.

[15] Xie M and Singh K. Confidence distribution, the frequentist distribution estimator of a parameter: a review. *Int Stat Rev* 2013; **81**: 3–39.

[16] Cochran WG. Problems arising in the analysis of a series of similar experiments. *J R Stat Soc* 1937; (Supplment) **4**:102–118.

[17] Cochran WG. The combination of estimates from different experiments. *Biometrics* 1954; **10**: 101–129.

[18] DerSimonian R and Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986; **7**: 177–188.

[19] Whitehead A and Whitehead J. A general parametric approach to the meta-analysis of randomized clinical trials. *Stat Med* 1991; **10**: 1665–1677.

[20] Biggerstaff BJ and Tweedie RL. Incorporating variability of estimates of heterogeneity in the random effects model in meta-analysis. *Stat Med* 1997; **16**: 753–768.

[21] Biggerstaff BJ and Jackson D. The exact distribution of Cochran's heterogeneity statistic in one-way random effects meta-analysis. *Stat Med* 2008; **27**: 6093–6110.

[22] Sidik K and Jonkman JN. A comparison of heterogeneity variance estimators in combining results of studies. *Stat Med* 2007; **26**: 1964–1981.

[23] Scheffé H. *The Analysis of Variance*. New York: Wiley, 1959.

[24] Graybill FA. *Theory and Application of the Linear Model*. North Scituate: Duxbury Press, 1976.

[25] Mathai AM and Provost SB. *Quadratic Forms in Random Variables: Theory and Applications*. New York: Marcel Dekker, 1992.

[26] Hartung J and Knapp G. On tests of the overall treatment effect in meta-analysis with normally distributed responses. *Stat Med* 2001; **20**: 1771–1782.

[27] Sidik K and Jonkman JN. Robust variance estimation for random effects meta-analysis. *Comput Stat Data Anal* 2006; **50**: 3681–3701.

[28] Harville DA. Maximum likelihood approaches to variance component estimation and to related problems. *J Am Stat Assoc* 1977; **72**: 320–339.

[29] Raudenbush SW and Bryk AS. Empirical Bayes meta-analysis. *J Educ Stat* 1985; **10**: 75–98.

[30] Hartung J. An alternative method for meta-analysis. *Biom J* 1999; **41**: 901–916.

[31] Fisher RA. The fiducial argument in statistical inference. *Ann Eugen* 1935; **6**: 391–398.

[32] Efron B. R. A. Fisher in the 21st century. *Stat Sci* 1998; **13**: 95–122.

[33] Viechtbauer W. Confidence intervals for the amount of heterogeneity in meta-analysis. *Stat Med* 2007; **26**: 37–52.

[34] Forsythe GE, Malcolm MA and Moler CB. *Computer Methods for Mathematical Computations*. Englewood Cliffs: Prentice-Hall, 1977.

[35] Farebrother RW. Algorithm AS 204: the distribution of a positive linear combination of χ^2 random variables. *J R Stat Soc Ser C Appl Stat* 1984; **33**: 332–339.

[36] Brockwell SE and Gordon IR. A simple method for inference on an overall effect in meta-analysis. *Stat Med* 2007; **26**: 4531–4543.

[37] Jackson D. Confidence intervals for the between study variance in random effects meta-analysis using generalized Cochran heterogeneity statistics. *Res Synth Methods* 2013; **4**: 220–229.

[38] Turner RM, Davey J, Clarke MJ, et al. Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane Database of Systematic Reviews. *Int J Epidemiol* 2012; **41**: 818–827.

[39] Rhodes KM, Turner RM and Higgins JPT. Predictive distributions were developed for the extent of heterogeneity in meta-analyses of continuous outcome data. *J Clin Epidemiol* 2015; **68**: 52–60.

[40] Roberts ME, Tchanturia K, Stahl D, et al. A systematic review and meta-analysis of set-shifting ability in eating disorders. *Psychol Med* 2007; **37**: 1075–1084.

[41] Häuser W, Bernardy K, Üçeyler N, et al. Treatment of fibromyalgia syndrome with antidepressants: a meta-analysis. *JAMA* 2009; **301**: 198–209.