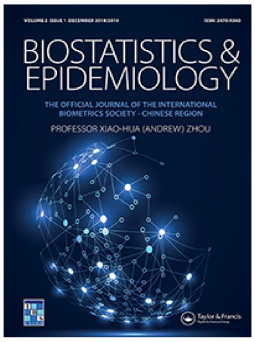
BIOSTATISTICS & EPIDEMIOLOGY
THE OFFICIAL JOURNAL OF THE INTERNATIONAL BIOMETRICS SOCIETY - CHINESE REGION
PROFESSOR XIAO-HUA (ANDREW) ZHOU
Taylor & Francis

Taylor & Francis
Taylor & Francis Group

Biostatistics & Epidemiology

ISSN: 2470-9360 (Print) 2470-9379 (Online) Journal homepage: https://www.tandfonline.com/loi/tbep20

Exact inference on meta-analysis with generalized fixed-effects and random-effects models

Sifan Liu, Lu Tian, Steve Lee & Min-ge Xie

To cite this article: Sifan Liu, Lu Tian, Steve Lee & Min-ge Xie (2018) Exact inference on meta-analysis with generalized fixed-effects and random-effects models, Biostatistics & Epidemiology, 2:1, 1-22, DOI: 10.1080/24709360.2017.1400714

To link to this article: https://doi.org/10.1080/24709360.2017.1400714

Published online: 22 Nov 2017.

BIOSTATISTICS & EPIDEMIOLOGY
2018, VOL. 2, NO. 1, 1–22
https://doi.org/10.1080/24709360.2017.1400714

Taylor & Francis
Taylor & Francis Group

Exact inference on meta-analysis with generalized fixed-effects and random-effects models

Sifan Liu [a], Lu Tian [b], Steve Lee [c] and Min-ge Xie [a]

[a]Department of Statistics and Biostatistics, Rutgers University, Piscataway, NJ, USA; [b]Department of Biomedical Data Science, Stanford University, Palo Alto, CA, USA; [c]Genentech Inc., South San Francisco, CA, USA

ABSTRACT
Meta-analysis with fixed-effects and random-effects models provides a general framework for quantitatively summarizing multiple comparative studies. However, a majority of the conventional methods rely on large-sample approximations to justify their inference, which may be invalid and lead to erroneous conclusions, especially when the number of studies is not large, or sample sizes of the individual studies are small. In this article, we propose a set of 'exact' confidence intervals for the overall effect, where the coverage probabilities of the intervals can always be achieved. We start with conventional parametric fixed-effects and random-effects models, and then extend the exact methods beyond the commonly postulated Gaussian assumptions. Efficient numerical algorithms for implementing the proposed methods are developed. We also conduct simulation studies to compare the performance of our proposal to existing methods, indicating our proposed procedures are better in terms of coverage level and robustness. The new proposals are then illustrated with the data from meta-analyses for estimating the efficacy of statins and BCG vaccination.

ARTICLE HISTORY
Received 9 December 2016
Accepted 19 August 2017

KEYWORDS
Meta-analysis; exact inference; fixed-effects model; random-effects model

1. Introduction

Meta-analysis has been widely used to combine information from multiple studies, especially in medical research. One important objective is to make inference about the overall effect often relative to a standard care of therapy. The fixed-effects and random-effects models, often coupled with the DerSimonian and Laird (D-L) approach [1], are two most commonly used statistical models in meta-analysis. However, the D-L method is suboptimal and may lead to too many statistically significant results when the number of studies is small and there is moderate or substantial heterogeneity [2]. Depending on specific settings, the coverage probability of the confidence interval (CI) by the D-L method may fail to achieve the target level even when the number of studies is as high as 20–35 [2,3]. The key reason is that the validity of the CI depends on large-sample approximation of the combined point estimator. The goal of this paper is to propose a family of test statistics for constructing exact CIs under fixed-effects and random-effects models, that are valid

CONTACT Sifan Liu sifan.liu@rutgers.edu

regardless of the number of studies. The validity of the proposal does not rely on large-sample approximations and the corresponding coverage probabilities can always achieve the specified nominal level.

This research is partially motivated by need of evaluating the efficacy of BCG vaccine in the prevention of tuberculosis. Though the use of BCG has a long history with billions of doses given, there has been an on-going debate on its efficacy [4,5]. Multiple clinical studies are identified for meta-analysis and the D-L approach was used to combine the information [6]. However, given the limited number of studies, the results based on the D-L method may not be reliable and hence more robust statistical inference is wanted.

Indeed, various CI procedures aiming to correct the under-coverage of the D-L method for random-effects model have been developed recently. Likelihood approaches, such as constructing CIs by profile likelihood [7] and by the restricted maximum likelihood method [8], are considered. Modifications are proposed to account for the between-study variability, e.g. [9], and the Student's t-distribution is also used [10]. Another approach is to centre the CI at the fixed-effects estimate with a robust variance estimator [11]. In addition, there are some higher order asymptotical inference procedures such as the Bartlett-type correction for the likelihood ratio statistics [12]. However, all these inferences are still asymptotic with respect to the number of studies. More recently, confidence distribution (CD) [13–15] is proved to be a powerful vehicle in developing new meta-analysis methods [16–18]. But again, these CD-based methods are asymptotic procedures.

Some exact methods, including Tian et al.'s method of combining CIs [19] and Liu et al.'s method of combining p value functions [20], have been developed. Mainly focusing on meta-analysis of rare events, both of these methods can be unified under the general framework of combining CDs [21]. Nevertheless, they are developed under the setting of fixed-effects models. We are interested in developing an exact inference procedure for both fixed-effects and random-effects models without relying on exact test for each study. The most related work is the permutation method proposed by Follmann and Proschan [22]. However, its implementation is slow except for very small number of studies.

The problem considered here is also closely related to the well-known Behrens-Fisher (B-F) problem, i.e. comparing the means of two Gaussian distributions with unknown variances. Specifically, the Gaussian fixed-effects model in Section 2 is directly related to the B-F problem, and the random-effects model is even more complicated. Compared to the exact solutions to the B-F problem, which need a second stage sampling [23–25], our proposed solution is much more direct and simpler. In addition, our method can also be extended beyond the fixed-effects model and even the parametric distribution assumptions.

In the rest of the article, we first propose procedures of constructing the exact CIs for conventional Gaussian fixed- and random-effects models. We show that the coverage probabilities of the resulting CIs can always achieve the target level, regardless of the number of studies. Some easy-to-implement computation algorithms are provided for constructing the corresponding exact confidence intervals. Then, we consider generalized fixed- and random-effects models, where the Gaussian assumptions can be relaxed. Lastly, we report results from the conducted simulation studies and two real life meta-analyses.

2. Exact inference with Gaussian fixed-effects and random-effects models

2.1. Model settings and review of the D-L method

Suppose that we have K independent studies and the ith study has a summary statistic (observed treatment effect) Y_i for the true study-specific treatment effect θ_i. The standard meta-analysis random-effects model assumes that, independently,

$$Y_i \mid \theta_i \sim N\left(\theta_i, \sigma_i^2\right), \theta_i \sim N\left(\mu_0, \tau^2\right), i = 1, \ldots, K; \tag{1}$$

which is equivalent to the parametric model

$$Y_i \sim N\left(\mu_0, \sigma_i^2 + \tau^2\right), i = 1, \ldots, K. \tag{2}$$

Here, μ_0 is the overall effect, σ_i^2 is the within-study variance and τ^2 is the between-study variance which is generally unknown.

The D-L method [1] is to estimate μ_0 by

$$\hat{\mu}_{\mathrm{DL}} = \frac{\sum_{i=1}^K \hat{w}_i Y_i}{\sum_{i=1}^K \hat{w}_i},$$

where $\hat{w}_i = (\sigma_i^2 + \hat{\tau}_{DL}^2)^{-1}$ is the inverse-variance weight, $\hat{\tau}_{DL}^2$ is a moment estimate of τ^2, given by

$$\hat{\tau}_{\mathrm{DL}} = \max\left\{ \frac{\left\{\sum_{i=1}^K \sigma_i^{-2}(Y_i - \hat{\mu}_F)^2\right\} - (K-1)}{\sum_{i=1}^K \sigma_i^{-2} - \sum_{i=1}^K \sigma_i^{-4} / \sum_{i=1}^K \sigma_i^{-2}}, 0 \right\},$$

and

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

is an initial estimator for μ_0. Given$\{(Y_i, \sigma_i) \mid i = 1, \ldots, K\}$, the normal approximation $(\hat{\mu}_{\mathrm{DL}} - \mu_0) \sim N\left(0, 1/\sum_{i=1}^K \hat{w}_i\right)$ leads to $100(1-\alpha)\%$ CI for μ_0,

$$\left[\hat{\mu}_{\mathrm{DL}} - z_{\alpha/2}\left(\sum_{i=1}^K \hat{w}_i\right)^{-1/2}, \hat{\mu}_{\mathrm{DL}} + z_{\alpha/2}\left(\sum_{i=1}^K \hat{w}_i\right)^{-1/2}\right]$$

where $z_{\alpha/2}$ is the $\alpha/2$ upper quantile of the standard normal. The validity of this CI relies on the large-sample approximations with the assumption that the number of studies, K, goes to infinity. When K is small, $\hat{\tau}_{\mathrm{DL}}^2$ can be inaccurate but the D-L method does not account for its randomness. See also [26] and [27] for insightful discussions on related challenges.

In the remaining of this section, we construct the exact CIs for μ_0 by inverting appropriate exact tests. Note that when $\tau^2 = 0$, the random-effects model (1) degenerates into the simple fixed-effects model, independently,

$$Y_i \sim N\left(\mu_0, \sigma_i^2\right), i = 1, \ldots, K. \tag{3}$$

As a special case of random-effects models, the fixed-effects model requires that $\theta_1 = \cdots = \theta_K = \mu_0$ and the usual estimate of μ_0 is simply $\hat{\mu}_F$. It is not difficult to see that all proposed CIs for μ_0 in this article are valid under both fixed- and random-effects models. We shall avoid repeating this observation when each individual CI is discussed later. Throughout the developments, we let $V_i \sim \text{Bernoulli}\left(\frac{1}{2}\right), i = 1, \dots, K$.

2.2. Proposed test statistics and exact CIs for μ_0

Motivated by the 'exact' hypothesis testing procedure on median, we propose to consider the test statistics.

$$T_w(\mu) = \sum_{i=1}^{K} w_i \left\{ I(Y_i \le \mu) - \frac{1}{2} \right\}, \tag{4}$$

which is essentially a weighted sign test statistic. Here, $\{w_1, \dots, w_K\}$ is a set of positive weights given a priori, and $I(\cdot)$ is the indicator function. Model (2) implies that $\Pr(Y_i \le \mu_0) = 0.5$, and $T_w(\mu_0)$, at the true value μ_0, is a weighted sum of K independent Bernoulli random variables. Thus, we can define

$$T_w^* = \sum_{i=1}^{K} w_i \left\{ V_i - \frac{1}{2} \right\}. \tag{5}$$

Immediately, we have the key equivalence that $T_w(\mu_0)$ has the same distribution of T_w^*, i.e.

$$T_w(\mu_0) \sim T_w^* \tag{6}$$

which leads to the construction of the exact CI for μ_0 shown in the following theorem.

The rigorous justification is provided in the Appendix A.

Theorem 2.1. *For the random-effects model (1), consider the test statistic $T_w(\mu)$ (4) and the random variable T_w^* (5). Define*

$$p_w(t) = 2\min\left\{F_w^*(t), S_w^*(t)\right\},$$

where $F_w^*(t) = Pr\left(T_w^* \le t\right)$ and $S_w^*(t) = Pr\left(T_w^* \ge t\right)$. Then, the $100(1-\alpha)\%$ CI for μ_0 can be constructed as

$$C_{w\alpha} = [\mu : p_w\{T_w(\mu)\} > \alpha] \tag{7}$$

Here, $p_w\{T_w(\mu)\}$ can serve as the exact two-sided p value for testing $H_0 : \mu_0 = \mu$ versus $H_A : \mu_0 \neq \mu$. Operationally, $C_{w\alpha}$ can be constructed by assembling all μ's over a dense grid with the corresponding p value, $p_w\{T_w(\mu)\}$, greater than the significance level α. Note that, based on (6), $F_w^*\{T_w(\mu_0)\}$ is stochastically greater than or equal to the uniform distribution U(0, 1). Then, $F_w^*\{T_w(\mu)\}$ may serve as the exact p value in testing the null hypothesis $H_0 : \mu_0 \geq \mu$versus the alternative hypothesis $H_A : \mu_0 < \mu$ and the lower end of the one-sided $100(1-\alpha/2)\%$ CI for μ_0 can be found by inverting this exact test as $\mu_{wL} = \inf\left[\mu : F_w^*\{T_w(\mu)\} \geq \alpha/2\right]$. Similarly, $S_w^*\{T_w(\mu_0)\}$ is the exact p value in testing $H_0 : \mu_0 \leq \mu$ versus $H_A : \mu_0 > \mu$ and can be used to generate the upper end of the one-sided CI of μ_0 as $\mu_{wU} = \sup\left[\mu : S_w^*\{T_w(\mu)\} \geq \alpha/2\right]$. Therefore, based on these two one-sided exact CIs, we have an alternative expression of $C_{w\alpha}$ as (μ_{wL}, μ_{wU}).

Furthermore, since $T_w(\mu_0)$ has a discrete distribution, $p_w\{T_w(\mu)\}$ is a step function with respect to μ. In order to construct non-equal tailed CIs, which are potentially narrower than equal tailed CIs, the p value function $p_w(t)$ may be replaced by

$$p_w^\gamma(t) = \min\left\{\frac{F_w^*(t)}{\gamma}, \frac{S_w^*(t)}{1-\gamma}\right\},$$

for some $\gamma \in (0, 1)$.

Although the validity of the proposed CIs does not depend on the choice of $\{w_1, \ldots, w_K\}$in $T_w(\mu)$, the distribution of $T_w(\mu_0)$ does. It is reasonable to consider the (asymptotically) optimal weights, which tend to generate relatively narrow CIs. Specifically, for the fixed-effects model (3), we propose to use the inverse of standard deviation as the study-specific weight $w_i = \sigma_i^{-1}$, and define

$$T_1(\mu) = \sum_{i=1}^K \sigma_i^{-1}\left\{I(Y_i \leq \mu) - \frac{1}{2}\right\}$$

This is intuitive in that less informative studies are down-weighted. Compared to fixed-effects model, random-effects model (1) with τ^2 is widely used to deal with the study heterogeneity in meta-analysis. As to the cases with $\tau^2 = \tau_0^2$ known, we can use

$$w_i = \left\{\sigma_i^2 + \tau_0^2\right\}^{-\frac{1}{2}}, i = 1, \ldots, K. \tag{8}$$

When τ^2 is unknown, we may replace τ_0^2 in (8) by an estimator of τ^2, which leads to an extension of T_w with data-dependent weight components.

2.3. Extended test statistics and the corresponding exact confidence sets for μ_0

Similarly to previous developments, we first introduce the exact inference procedure in a general form, and then discuss some specific choices of the test statistics. Consider an extended version of T_w,

$$T_{\hat{w}}(\mu) = \sum_{i=1}^K \hat{w}_i(\mu)\left\{I(Y_i \leq \mu) - \frac{1}{2}\right\}, \tag{9}$$

where $\{\hat{w}_i(\mu) | i = 1, \ldots, K\}$ are positive and data-dependent. In order to construct exact confidence sets based on $T_{\hat{w}}(\mu)$, we impose the following condition:

- **Condition A:** Each component from the set of weights $\{\hat{w}_i(\mu_0) | i = 1, \ldots, K\}$ is independent of $\{I(Y_i \leq \mu_0) | i = 1, \ldots, K\}$.

Then, we let

$$T^*_{\hat{w}}(\mu) = \sum_{i=1}^{K} \hat{w}_i(\mu)\left\{V_i - \frac{1}{2}\right\}. \tag{10}$$

Under Condition A, we have the following key result of equivalence in distribution:

$$T_{\hat{w}}(\mu_0)|\{\hat{w}_1(\mu_0), \ldots, \hat{w}_K(\mu_0)\} \sim T_{\hat{w}}{}^*(\mu_0)|\{Y_1, \ldots, Y_K\}$$

which is followed from the fact that

$$I(Y_i \leq \mu_0) \sim \text{Bernoulli}\left(\frac{1}{2}\right), i = 1, \ldots, K.$$

The construction of exact confidence sets of μ_0 is then proposed and justified in the following theorem:

Theorem 2.2. *For the random-effects model (1), consider the test statistic $T_{\hat{w}}(\mu)$(9) and the random variable $T^*_{\hat{w}}(\mu)$ (10). Suppose that $\{\hat{w}_i(\mu)|i = 1, \ldots, K\}$ satisfy Condition A. Let $F^*_{\hat{w}}(t, \mu) = Pr(T_{\hat{w}}{}^*(\mu) \leq t \mid Y_1, \ldots, Y_K), S^*_{\hat{w}}(t, \mu) = Pr(T_{\hat{w}}{}^*(\mu) \geq t \mid Y_1, \ldots, Y_K)$; and define*

$$p_{\hat{w}}(t, \mu) = 2min\left\{F_{\hat{w}}{}^*(t, \mu), S^*_{\hat{w}}(t, \mu)\right\}.$$

Then, the $100(1-\alpha)\%$confidence set of μ_0 can be constructed as

$$C_{\hat{w}}\alpha = [\mu : p_{\hat{w}}\{T_{\hat{w}}(\mu), \mu\} > \alpha]. \tag{11}$$

The rigorous proof of Theorem 2.2 is given in Appendix B. Here, $p_{\hat{w}}\{T_{\hat{w}}(\mu), \mu\}$ may also serve as the exact two-sided p values for testing $H_0 : \mu_0 = \mu$ versus $H_A : \mu_0 \neq \mu$. Since $T_{\hat{w}}(\mu)$ is not guaranteed to be monotone in μ, the generated confidence set might be a union of disjointed intervals. In practice, it is common to report the conservative intervals $(\mu_{\hat{w}}L, \mu_{\hat{w}}U)$, where

$$\mu_{\hat{w}}L = \inf\left[\mu : F_{\hat{w}}{}^*\{T_{\hat{w}}(\mu), \mu\} \geq \frac{\alpha}{2}\right]; \mu_{\hat{w}}U = \sup\left[\mu : S_{\hat{w}}{}^*\{T_{\hat{w}}(\mu), \mu\} \geq \frac{\alpha}{2}\right].$$

Note that $(\mu_{\hat{w}}L, \mu_{\hat{w}}U)$ is the shortest interval containing $C_{\hat{w}}\alpha$.

More specifically, we may modify the fixed weights (8) by using the asymptotically optimal weights,

$$T_2(\mu) = \sum_{i=1}^{K} \left\{\sigma_i^2 + \hat{\tau}^2(\mu)\right\}^{-\frac{1}{2}}\left\{I(Y_i \leq \mu) - \frac{1}{2}\right\},$$

where $\hat{\tau}^2(\mu)$ is an estimator of τ^2. To meet Condition A, $\hat{\tau}^2(\mu_0)$ is required to be independent from $\{I(Y_i \leq \mu_0)|i = 1, \ldots, K\}$. Compared to $T_1(\mu)$, one natural advantage of T_2

(μ) is that it delivers better performance when K is adequately large. When K is small, our experience also suggests that the CIs based on $T_2(\mu)$ are no wider than those based on $T_1(\mu)$ (e.g. see Table 1).

Since $\hat{\tau}^2_{DL}$ in the D-L method unfortunately does not satisfy Condition A, we propose another moment estimator as a simple and valid choice of $\hat{\tau}^2(\mu)$,

$$\hat{\tau}^2_{\text{mom}}(\mu) = \max\left[K^{-1}\left\{\sum_{i=1}^{K}(Y_i-\mu)^2-\sum_{i=1}^{K}\sigma_i^2\right\},0\right].$$

Based on the fact that, under the random-effects model (1), the sign of $Y_i-\mu_0$ is independent of its magnitude $\mid Y_i-\mu_0\mid$, for all $i=1,\ldots,K$, $\hat{\tau}_{\text{mom}}2(\mu)$ always satisfies Condition A.

Condition A is a mild requirement in practice. It is satisfied as long as $\hat{\tau}^2(\mu)$ is only a function of $\mid Y_i-\mu_0\mid$'s, which are independent of $I(Y_i\le\mu_0)$'s under model (1). For example, a robust alternative to $\hat{\tau}^2_{\text{mom}}(\mu)$ is the solution to the equation.

$$\text{median}\left\{\frac{(Y_i-\mu)^2}{\sigma_i^2+\tau^2}\mid i=1,\ldots,K\right\}=c_K,$$

where $c_K=E\left[\text{median}\left(Z_1^2,\ldots,Z_K^2\right)\right]$ and $Z_1,\ldots,Z_K$ are i.i.d standard normal random variables. However, if the distribution of θ_i is not symmetric at its centre, this condition may be violated.

Remark 1. Based on the Hodges–Lehmann estimator [28], we consider another test statistic

$$T_{\text{HL}}(\mu)=\sum_{1\le i\le j\le K}\left[I\left\{\frac{1}{2}\left(\hat{w}_i(\mu)(Y_i\le\mu)+\hat{w}_j(\mu)\left(Y_j\le\mu\right)\right)\le 0\right\}-\frac{1}{2}\right].$$

Let

$$T^*_{\text{HL}}(\mu)=\sum_{1\le i\le j\le K}\left[I\left\{\left[\left(V_i-\frac{1}{2}\right)\hat{w}_i(\mu)\mid Y_i\le\mu\mid+\left(V_i-\frac{1}{2}\right)\hat{w}_j(\mu)\mid Y_j\le\mu\mid\right]\le 0\right\}-\frac{1}{2}\right].$$

Then, under Condition A, we also have

$$T_{\text{HL}}(\mu_0)\big|\{\hat{w}_1(\mu_0)|Y_1\le\mu|,\ldots,\hat{w}_K(\mu_0)|Y_K\le\mu|\}\sim T^*_{HL}(\mu_0)\big|\{Y_1,\ldots,Y_K\}.$$

The exact confidence set for μ_0 can be constructed by the same procedure shown in Theorem 2.2. One typical example corresponding to the weight components proposed in $T_2(\mu)$ is that

$$T_3(\mu)=\sum_{1\le i\le j\le K}\left[I\left\{\frac{1}{2}\left(\frac{Y_i-\mu}{\left\{\sigma_i^2+\hat{\tau}^2(\mu)\right\}^{1/2}}+\frac{Y_j-\mu}{\left\{\sigma_j^2+\hat{\tau}^2(\mu)\right\}^{1/2}}\right)\le 0\right\}-\frac{1}{2}\right].$$

Remark 2. *Based on the previous developments, it is clear that any construction of* $\{\hat{w}_i(\mu)|i=1,\ldots,K\}$ *only via* $\{\,|\,Y_i-\mu\,|\,|i=1,\ldots,K\}$ *should meet Condition A. Our method can be conveniently extended to the D-L estimator by considering the test statistic*

$$T_4(\mu)=\sum_{i=1}^{K}\frac{1}{\sigma_i^2+\hat{\tau}^2}(\mu)\{Y_i-\mu\},$$

whose null distribution can be approximated by

$$T_4^*(\mu)=2\sum_{i=1}^{K}\frac{|Y_i-\mu|}{\sigma_i^2+\hat{\tau}^2(\mu)}\left\{V_i-\frac{1}{2}\right\}.$$

Follmann and Proschan (F-P) proposed a similar CI procedure [22]. However, in approximating the null distribution of the test statistics, the moment estimator $\hat{\tau}^2(\mu)$ *is updated based on permuted samples* $\{(2V_i-1)|Y_i-\mu|+\mu|i=1,\ldots,K\}$. *This unnecessary step introduces nontrivial computational burden compared with our method.*

2.4. Numerical computation

In the following, we present relevant numerical algorithms to calculate the proposed CIs. First, the computation of the exact confidence interval $C_{w\alpha}$ (7) is straightforward and may serve as the cornerstone for others. To this end, one may use the following simple algorithm for small Ks.

[***Algorithm A***]

1) Compute all 2^K values in the set $\left\{\sum_{\mathrm{i}=1}^{K}(v_i-1/2)w_i,\ (v_1,\ldots,v_K)'\in\{0,1\}^K\right\}$.
2) Find the $(\alpha/2)$th lower quantile of the set, denoted by $q_{wL}^*(\alpha/2)$. Let the $(\alpha/2)$th upper quantile $q_{wU}^*\ (\alpha/2)=\ -q_{wL}^*(\alpha/2)$.
3) Compute the CI (μ_{wL},μ_{wU}), where the lower and upper bounds are

$$\mu_{wL}=\inf\left[\mu:T_w(\mu)\geq q_{wL}^*(\alpha/2)\right];\mu_{wU}=\sup\left[\mu:T_w(\mu)\leq q_{wU}^*(\alpha/2)\right].$$

The overall computational complexity of the algorithm is in the order of $O(2^K)$. Furthermore, we rewrite the Hodges–Lehmann estimator-based test statistic as

$$T_{\mathrm{HL}}(\mu)=\sum_{i=1}^{K}I(Y_i\leq\mu)R_i(\mu)-\frac{K(K+1)}{4}.$$

Then, it is easy to see that $T_{\mathrm{HL}}^*(\mu)$ is uniformly distributed over the set

$$\left\{\sum_{i=1}^{K}iv_i-\frac{K(K+1)}{4},(v_1,\ldots,v_K)'\in\{0,1\}^K\right\},$$

which is independent of μ. Here $R_i(\mu)$ is the rank of the ith element of

$$\{\hat{w}_i(\mu)|Y_i-\mu|,i=1,\ldots,K\}.$$

Therefore, in order to construct the corresponding CI, denoted as C_α^{HL} we only need to modify Algorithm A slightly as following:

[***Algorithm A****]

1) Compute all 2^K values in the set $\left\{\sum_{\mathrm{i}=1}^{K} iv_i - \frac{K(K+1)}{4},\ (v_1,\ \ldots,\ v_K)' \in \{0,\ 1\}^K\right\}$.
2) Find the $\alpha/2$ lower quantile of 2^K values in the set, denoted by $q^*_{\mathrm{HL},L}(\alpha/2)$. Let $q^*_{\mathrm{HL},U}(\alpha/2) = -q^*_{\mathrm{HL},L}(\alpha/2)$.
3) Compute the CI $C_\alpha^{\mathrm{HL}} = \left(\mu_L^{\mathrm{HL}},\ \mu_U^{\mathrm{HL}}\right)$, where the lower and upper bounds are

$$\mu_L^{\mathrm{HL}} = \inf\left[\mu : T_{\mathrm{HL}}(\mu) \geq q^*_{\mathrm{HL},L}(\alpha/2)\right]; \mu_U^{\mathrm{HL}} = \sup\left[\mu : T_{\mathrm{HL}}(\mu) \leq q^*_{\mathrm{HL},U}(\alpha/2)\right].$$

Since the cut-off values $\{q^*_{\mathrm{HL},L}(\alpha/2),\ q^*_{\mathrm{HL},U}(\alpha/2)\}$ only depend on K, one may compute and store them in advance for different Ks to further accelerate the computation.

The construction of the confidence set $C_{\hat{w}}\alpha$ (11) can be more complicated due to the fact that the distribution of $T^*_{\hat{w}}(\mu)$ can depend on μ. Therefore, we propose the following algorithm for computing the shortest interval $(\mu_{\hat{w}L}, \mu_{\hat{w}U})$ containing $C_{\hat{w}\alpha}$:

[***Algorithm B***]

1) For each fixed μ, compute all 2^K values in the set

$$\Omega_K(\mu) = \left\{\sum_{i=1}^{K} v_i\hat{w}_i(\mu) - S_0(\mu), (v_1, \ldots, v_K)' \in \{0,1\}^K\right\},$$

where $S_0(\mu) = 0.5\sum_{i=1}^{K}\hat{w}_i(\mu)$.

2) Find the $(\alpha/2)$th lower quantile of $\Omega_K(\mu)$, denoted by $q^*_{\hat{w}L}(\mu,\alpha/2)$. Let $q^*_{\hat{w}U}(\mu,\alpha/2) = -q^*_{\hat{w}L}(\mu,\alpha/2)$.
3) Repeat Steps 1) and 2) over a grid of values $\mu \in \{\mu_m \mathrm{m} = 1,\ \ldots,\ M\}$ ranging from $\min\{Y_1,\ \ldots,\ Y_K\}$ to $\max\{Y_1,\ \ldots,\ Y_K\}$. Compute $(\mu_{\hat{w}L}, \mu_{\hat{w}U})$ as

$$\mu_{\hat{w}L} = \min\left[\mu_m : T_{\hat{w}}(\mu_m) \geq q^*_{\hat{w}L}(\mu_m,\alpha/2)\right];$$
$$\mu_{\hat{w}U} = \max\left[\mu_m : T_{\hat{w}}(\mu_m) \leq q^*_{\hat{w}U}(\mu_m,\alpha/2)\right].$$

The complexity of the algorithm is in the order of $O(M2^K)$. The computation becomes time-consuming even for moderate K. However, when we are only interested in constructing the 95% CI, as commonly the case, the aforementioned algorithm can be greatly improved. Take $K = 9$ as an example, we only need to compare the observed test statistic $T_{\hat{w}}(\mu)$ with the smallest 12 values in $\Omega(\mu)$ to determine whether μ belongs to $C_{\hat{w}}(0.95)$.

Without the loss of generality, we assume that $\hat{w}_1(\mu) < \cdots < \hat{w}_9(\mu)$. It can be shown that the set $\tilde{\Omega}_{9,0.025}(\mu) = \tilde{\Omega}_0(\mu) \cup \tilde{\Omega}_1(\mu) \cup \tilde{\Omega}_2(\mu) \cup \tilde{\Omega}_3(\mu)$ contains the smallest 12 values of $\Omega_9(\mu)$, where

$$\tilde{\Omega}_0(\mu) = \{-S_0(\mu)\};$$
$$\tilde{\Omega}_1(\mu) = \{\hat{w}_i(\mu) - S_0(\mu) i = 1, \ldots, 9\};$$
$$\tilde{\Omega}_2(\mu) = \left\{\hat{w}_i(\mu) + \hat{w}_j(\mu) - S_0(\mu)(i,j) = (1,2), \ldots, (1,6), (2,3), (2,4), (3,4)\right\};$$
$$\tilde{\Omega}_3(\mu) = \{\hat{w}_1(\mu) + \hat{w}_2(\mu) + \hat{w}_i(\mu) - S_0(\mu) | i = 3, 4\}.$$

Therefore, instead of calculating $2^9 = 512$ different values in $\Omega_9(\mu)$, one only needs computing 20 values in $\tilde{\Omega}_{9,0.025}(\mu)$ for each given μ in constructing the 95% CI. The rational is that any number not belonging $\tilde{\Omega}_{9,0.025}(\mu)$ is greater than at least 12 members from $\Omega_9(\mu)$. For example, $\hat{w}_2(\mu) + \hat{w}_5(\mu) - S_0(\mu)$ is always greater than $-S_0(\mu), \hat{w}_i(\mu) - S_0(\mu), i = 1, \ldots, 5$ and $\hat{w}_j(\mu) + \hat{w}_k(\mu) - S_0(\mu), (j,k) = (1,2), \ldots, (1,5), (2,3), (2,4)$. Consequently, in computing the 95% exact CI based on $T_{\hat{w}}(\mu)$, one may replace the first two steps of Algorithm B by

1) For each fixed μ, compute all values in the set $\tilde{\Omega}_{K,0.025}(\mu)$.
2) Find the $[2^K \times 0.025]$th smallest value in $\tilde{\Omega}_{K,0.025}(\mu)$, denoted by $q^*_{\hat{w}L}(\mu, 0.025)$, where $[x]$ represents the largest integer no greater than x. Let $q^*_{\hat{w}U}(\mu, 0.025) = -q^*_{\hat{w}L}(\mu, 0.025)$.

Similar to the quantiles of $T^*_{\mathrm{HL}}(\mu)$, the membership of $\tilde{\Omega}_{K,0.025}(\mu)$ is independent of μ, and can be calculated for a sequence of Ks and stored in advance. Therefore, the computational complexity for a specified data-set can be quite low, although the implementation seems to be involved. Since the cardinality of $\tilde{\Omega}_{K,0.025}(\mu)$ is much smaller than that of $\Omega_K(\mu)$, the computation speed can be greatly improved.

To compare the computational efficiency of our proposed algorithms together with the permutation procedure (the F-P method [22]), we conduct a small experiment on the computation speed and the results are shown in Figure 1. The computation of C^{HL}_α based on $T_3(\mu)$ is substantially faster than others, while the computation of the F-P interval is the slowest as anticipated. For example, when $K = 18$, the speed of computing C^{HL}_α is 1959 times faster than the improved Algorithm B for computing $C_{\hat{w}}\alpha$ based on $T_2(\mu)$. This improved Algorithm B is eleven times faster than the original counterpart, which is still four times faster than computing the F-P interval. Note that the F-P interval can be obtained by using appropriately modified Algorithm B.

The faster computation speed of our methods comes from two sources: (i) for each given permutation, we do not need to update $\hat{\tau}^2(\mu)$ while F-P method recalculates $\hat{\tau}^2$ each time, which is a nontrivial computational burden; (ii) since $\hat{\tau}^2(\mu)$ is a constant, we only need to consider $2^K/20 \sim 2^K/10$ selected permutations instead of all 2^K permutations to identify the 2.5 percentile of the null distribution of the test statistics using the improved Algorithm B. The proposed methods are implemented in a newly developed R package 'RandMeta.'

Remark 3. When K is large, e.g. greater than 20, the Monte-Carlo simulation can be used for approximating the quantiles of the appropriate test statistics. For example, $q^*_{wL}(\alpha/2)$ in Algorithm A can be estimated by the $(\alpha/2)$th lower quantile of a large number of copies of T^*_w obtained by repeatedly simulating $V_i \sim Bernoulli\left(\frac{1}{2}\right), i = 1, \ldots, K$.

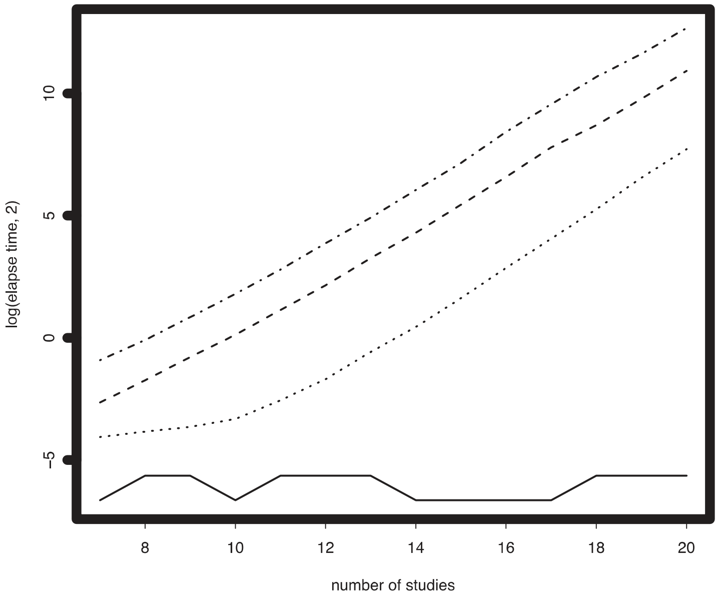

Figure 1. The computational speed of the proposed algorithms: Algorithm A* for C_α^{HL} based on $T_3(\mu)$ (solid line); improved Algorithm B for $C_{\hat{w}}\alpha$ based on $T_2(\mu)$ (dotted line); Algorithm B for $C_{\hat{w}}\alpha$ based on $T_2(\mu)$ (dashed line); and modified Algorithm B for the CI proposed by Follmann–Proschan (dash-dotted line).

3. Exact inference on generalized fixed-effects and random-effects models

In practice, the Gaussian assumptions are not always satisfied. Furthermore, μ_0 may not be limited to be the population mean. It can be other location measures for a distribution such as median and other quantiles. By carefully examining the developments in the previous sections, we note that, to guarantee the validity of $C_{\hat{w}}\alpha$, the key requirements are

- $\{\hat{w}_i(\mu)|\ \text{i}\ =\ 1,\ldots,\ K\}$ satisfy Condition A;
- $I(Y_i \le \mu_0) \sim \text{Bernoulli}\left(\frac{1}{2}\right),\ \text{i}\ =\ 1,\ldots,K.$

For the validity of C_α^{HL}, one additional requirement is

- $I(Y_i \le \mu_0)$ and $|Y_i - \mu_0|$ are independent, for i $=$ $1,\ldots,$ K.

Based on these observations, we generalize the proposed exact inference procedure to more general settings beyond Gaussian models. Specifically, we assume that, independently,

$$Y_i \sim \mathcal{F}_i, i = 1,\ldots,K; \tag{12}$$

where $\{\mathcal{F}_i, i = 1,\ldots,K\}$ is a set of distributions sharing a common location parameter of interest

$$\mu_0 = \mu_0(\mathcal{F}_i).$$

For instance, denoting by $F_i(t)$ the CDF of $\mathcal{F}_i$, μ_0 could be the population mean $\mu_0(\mathcal{F}_i) = \int tdF_i(t)$, the population median $F_i^{-1}(1/2)$ or any quantile $F_i^{-1}(p)$ for given $p \in (0,1)$.

To generalize the model assumptions for the random-effects model, we consider

- **Assumption I:** $\Pr(Y_i \le \mu_0) \ge p, \ i = 1, \dots, K$, for a known $p \in (0, 1)$.
- **Assumption II:** $\Pr(Y_i \ge \mu_0) \ge q, \ i = 1, \dots, K$, for a known $q \in (0, 1)$.

In order to construct exact confidence sets, we impose the following condition on the weight components:

- **Condition B:** There is a set of positive weights $\{\hat{w}_i(\mu) | \ i = 1, \dots, K\}$ such that $\{\hat{w}_i(\mu_0) | \ i = 1, \dots, K\}$ are independent of $\{I(Y_i \le \mu_0), \ I(Y_i \ge \mu_0) \ i = 1, \dots, K\}$.

Then, we can make the exact inference for μ_0. Specifically, denote

$$\xi_i \sim U(0,1), i = 1, \dots, K.$$

We may let

$$\tilde{T}_{p\hat{w}}(\mu) = \sum_{i=1}^{K} \hat{w}_i(\mu) I(Y_i \le \mu), \tilde{Z}_{q\hat{w}}(\mu) = \sum_{i=1}^{K} \hat{w}_i(\mu) I(Y_i \ge \mu);$$
$$T^*_{p\hat{w}}(\mu) = \sum_{i=1}^{K} \hat{w}_i(\mu) I(\xi_i \le p), Z^*_{q\hat{w}}(\mu) = \sum_{i=1}^{K} \hat{w}_i(\mu) I(\xi_i \ge 1 - q).$$

Then, Condition B and Assumptions I and II imply that

$$\tilde{T}_{p\hat{w}}(\mu_0) \big| \{\hat{w}_1(\mu_0), \dots, \hat{w}_K(\mu_0)\} \gtrsim T_{p\hat{w}}{}^*(\mu_0) \big| \{Y_1, \dots, Y_K\}$$

and

$$\tilde{Z}_{q\hat{w}}(\mu_0) \big| \{\hat{w}_1(\mu_0), \dots, \hat{w}_K(\mu_0)\} \gtrsim Z_{q\hat{w}}{}^*(\mu_0) \big| \{Y_1, \dots, Y_K\}.$$

Here, '$\gtrsim$' indicates stochastic ordering of two random variables; i.e. $U \gtrsim V$ means $\Pr(U \le t) \le Pr(V \le t)$ for all t. Based on the previous developments in Section 2, denote the CDFs of $T^*_{p\hat{w}}(\mu)$and $Z^*_{q\hat{w}}(\mu)$given $\{Y_1, \dots, Y_K\}$ by $F^*_{p\hat{w}}(\cdot, \mu)$ and $\overline{F}^*_{q\hat{w}}(\cdot, \mu)$, respectively. The exact CI for μ_0 may be constructed by

$$\left(\inf\left[\mu : F_{p\hat{w}}{}^*\{\tilde{T}_{p\hat{w}}(\mu), \mu\} \ge \alpha/2\right], \sup\left[\mu : \overline{F}_{q\hat{w}}{}^*\{\tilde{Z}_{q\hat{w}}(\mu), \mu\} \ge \alpha/2\right]\right).$$

Then, the exact CI for μ_0 can be constructed accordingly. We summarize the above discussions in the following general theorem, which is a generalization of Theorem 2.2 to more general cases:

Theorem 3.1. *Consider the generalized model* (12) *with Assumptions I and II. Under Condition B, let* $F^*_{p\hat{w}}(t,\mu) = Pr\big(T_{p\hat{w}}{}^*(\mu) \le t \mid Y_1,\ldots,Y_K\big), \overline{F}^*_{q\hat{w}}(z,\mu) = Pr\big(Z_{q\hat{w}}{}^*(\mu) \le z \mid Y_1,\ldots,Y_K\big)$*and*

$$\tilde{p}_{\hat{w}}(t,z;\mu) = 2\min\Big\{F_{p\hat{w}}{}^*(t,\mu), \overline{F}^*_{q\hat{w}}(z,\mu)\Big\}.$$

Then, a $100(1-\alpha)\%$ *confidence set for* μ_0 *can be constructed by*

$$\big[\mu : \tilde{p}_{\hat{w}}\big\{\tilde{T}_{p\hat{w}}(\mu), \tilde{Z}_{q\hat{w}}(\mu);\mu\big\} > \alpha\big]$$

Similarly, the Hodges–Lehmann estimator-based approach can be generalized under the following condition:

- **Condition C:** $\{\mid Y_i - \mu_0 \mid, \text{ i } = 1,\ldots,K\}$ are independent of $\{I(Y_i \le \mu_0),\ I(Y_i \ge \mu_0),\ \text{i} = 1,\ \ldots,\ K\}$.

Specifically, we let

$$\tilde{T}_{\text{HL}}(\mu) = \sum_{1\le i\le j\le K} I\left[\frac{1}{2}\big(\hat{w}_i(\mu)(Y_i - \mu) + \hat{w}_j(\mu)\big(Y_j - \mu\big)\big) \le 0\right]$$

$$\tilde{Z}_{\text{HL}}(\mu) = \sum_{1\le i\le j\le K} I\left[\frac{1}{2}\big(\hat{w}_i(\mu)(Y_i - \mu) + \hat{w}_j(\mu)\big(Y_j - \mu\big)\big) \ge 0\right]$$

$$T^*_{p,\text{HL}}(\mu) = \sum_{1\le i\le j\le K} I\left[\left\{I(\xi_i \le p) - \frac{1}{2}\right\}\hat{w}_i(\mu)|Y_i - \mu| + \left\{I(\xi_j \le p) - \frac{1}{2}\right\}\hat{w}_j(\mu) \mid Y_j - \mu \mid \ge 0\right]$$

and

$$Z^*_{q,\text{HL}}(\mu) = \sum_{1\le i\le j\le K} I\left[\left\{I(\xi_i \ge 1-q) - \frac{1}{2}\right\}\hat{w}_i(\mu)|Y_i - \mu| \right.$$
$$\left. + \left\{I\big(\xi_j \ge 1-q\big) - \frac{1}{2}\right\}\hat{w}_j(\mu)\big|Y_j - \mu\big| \ge 0\right].$$

The exact confidence set can then be constructed in a similar way of Theorem 3.1. The details are omitted.

It may be difficult to verify Condition C in some cases. However, one typical example satisfying it is the random-effects models with symmetrically and continuously distributed θ_i, which is a direct generalization of the Gaussian assumptions. Specifically, consider model (12) and assume that $\{\mathcal{F}_i, i = 1,\ldots,K\}$ are continuous and symmetric around μ_0, which is the parameter of interest. It is then easy to verify that, when $p = q = 1/2$, Assumptions I and II and Condition B are satisfied.

Remark 4. *Under Assumptions I and II,* $\tilde{p}_{\hat{w}}\big\{\tilde{T}_{p\hat{w}}(\mu), \tilde{Z}_{q\hat{w}}(\mu);\mu\big\}$ *may still be used as the two-sided exact p value for testing the null hypothesis* $H_0 : \mu_0 = \mu$ *versus* $H_A : \mu_0 \ne \mu$*, if Conditions B is only satisfied for* $\mu_0 = \mu$.

4. Numerical studies

4.1. Simulations

Simulation 1. We first consider the standard random-effects model (1). The number of studies K is chosen to be 8, 12 or 16. For $i = 1, \ldots, K$, let $\sigma_i = 1 + 4(i-1)/(K-1)$ and simulate the observations Y_i as follows:

$$\theta_i \sim N(0, 25/2);\ Y_i \mid (\theta_i, \sigma_i) \sim N\left(\theta_i, \sigma_i^2\right).$$

For comparisons, 95% CIs are constructed based on the proposed test statistics $T_i(\mu), i = 1, 2, 3, 4,$ together with several commonly used conventional methods, e.g. the D-L method and the Sidik–Jonkman method. Note that when K is small, the latter has been especially recommended (see [29] for some detailed discussions). In addition, the F-P (permutation) method is also implemented based on test statistics similar to $T_2(\mu)$ and $T_4(\mu)$, which are denoted as F-P(T_2) and F-P(T_4), respectively. For each K, the empirical coverage levels of the CIs, the average CI lengths, the standard deviation (std) of CI lengths and the average elapsed time (in secs) from 5000 simulated data-sets are reported in Table 1.

As K increases, the CIs become narrower and the stds of CI length become smaller as expected. But almost all conventional methods fail to achieve the desired level of coverage 95%, except the S-J method for K = 16. When K = 8, most of the empirical coverage probabilities are below 90%. Instead, our proposed methods can always achieve the desired coverage level for all the cases. Among our proposed methods, $T_4(\mu)$ performs the best with the shortest CIs and smallest stds of CI length. (The std for $T_4(\mu)$ is smaller than that for the D-L method when K = 16). $T_3(\mu)$ is conservative with the actual coverage levels being around 96%. Besides, $T_2(\mu)$ performs better than $T_1(\mu)$ when K = 12, 16, which

Table 1. The empirical coverage probabilities (Cov prob) of the 95% CIs, the average CI lengths (Length), the standard deviation of CI lengths (std) and the average elapsed time in seconds (Avg.t) by different methods in Simulation 1. (T_i: the proposed exact methods based on $T_i(\mu), i = 1, 2, 3, 4$; D-L: DerSimonian–Laird method; HE: Hedges method; H-S: Hunter–Schmidt method; S-J: Sidik–Jonkman method; ML: maximum-likelihood estimator; REML: restricted maximum-likelihood estimator; EB: empirical Bayes estimator; F-P(T_i): Follmann–Proschan permutation method based on $T_i(\mu)$, $i = 2$ or 4).

	$K = 8$			$K = 12$			$K = 16$		
Method	Cov prob	Length (std)	avg.t	Cov prob	Length (std)	Avg.t	Cov prob	Length (std)	Avg.t
D-L	0.887	5.90 (1.88)	0.006	0.916	4.96 (1.24)	0.006	0.923	4.32 (0.93)	0.005
HE	0.861	5.87 (2.11)	0.005	0.894	4.93 (1.42)	0.005	0.904	4.32 (1.06)	0.006
H-S	0.848	5.15 (1.56)	0.005	0.892	4.57 (1.11)	0.005	0.902	4.07 (0.86)	0.006
S-J	0.934	6.54 (1.59)	0.005	0.946	5.43 (1.04)	0.005	0.953	4.71 (0.77)	0.006
ML	0.848	5.40 (1.81)	0.011	0.894	4.72 (1.19)	0.011	0.907	4.16 (0.89)	0.011
REML	0.888	5.91 (1.89)	0.011	0.915	4.99 (1.22)	0.011	0.922	4.33 (0.90)	0.012
EB	0.890	6.00 (1.88)	0.010	0.916	5.03 (1.23)	0.010	0.923	4.36 (0.90)	0.011
T_1	0.954	9.16 (3.22)	0.010	0.950	7.48 (2.60)	0.042	0.950	6.22 (2.00)	0.763
T_2	0.954	9.16 (3.22)	0.033	0.950	7.01 (2.18)	0.381	0.950	5.90 (1.79)	0.901
T_3	0.962	8.35 (2.39)	0.026	0.959	6.18 (1.44)	0.026	0.962	5.23 (1.07)	0.018
T_4	0.955	7.79 (2.20)	0.034	0.950	5.77 (1.28)	0.381	0.951	4.80 (0.91)	0.892
F-P(T_2)	0.953	8.35 (2.70)	0.739	0.952	6.48 (1.88)	11.8	0.950	5.48 (1.52)	312.0
F-P(T_4)	0.951	7.75 (2.27)	0.745	0.951	5.91 (1.48)	12.1	0.953	4.90 (1.11)	319.3

indicates the advantage of using the asymptotically optimal weights even with a moderate number of studies.

In addition, although the F-P permutation CIs can also achieve the desired level, $T_4(\mu)$ has smaller sample std of the CI length suggesting more stable performance. When K = 12, 16, $T_4(\mu)$ has narrower CIs compared with the F-P method. More importantly, it is obvious that our proposed methods $(T_i(\mu), i = 1, 2, 3, 4)$ are much faster than the F-P method, which presents an important practical advantage.

Simulation 2. In the second set of simulations, K again is chosen to be 8, 12 or 16, and for $i = 1, \ldots, K$, $\sigma_i = 1 + 4(i-1)/(K-1)$. The observations Y_i are then simulated as follows:

$$\theta_i \sim t_2; Y_i \mid (\theta_i, \sigma_i) \sim N\left(\theta_i, \sigma_i^2\right),$$

where t_2 is the Student's t-distribution with degree of freedom 2. Since the treatment effects are generated from a t-distribution with much heavier tails than Gaussian distribution, the robustness of various methods against outlier study at the tail is of the primary interest. As in the first set of simulation, 5000 data-sets are simulated and 95% CIs are obtained for each K. Results including the empirical coverage levels and the average lengths the CIs are reported in Table 2.

Among the conventional methods, the S-J method is the only option that can achieve the desired coverage level. However, it is overly conservative and the actual coverage levels are above 97% even for K = 16, which results in unnecessarily wider CIs. Among our proposals, in terms of coverage level and the interval length, $T_4(\mu)$ and $T_3(\mu)$ perform the best when K = 8 and K = 12, 16, respectively. Especially, when K = 12 and 16, $T_3(\mu)$ can achieve 95% coverage level with CIs narrower than the S-J's, and its std's of CI length are smaller than all other methods'. Compared to the F-P's CIs, $T_3(\mu)$'s CIs have better coverage for all cases but also wider length except for K = 16.

Table 2. The empirical coverage probabilities (Cov prob) of the 95% CIs, the average CI lengths (Length), the standard deviation of CI lengths (std) and the average elapsed time in seconds (Avg.t) by different methods in Simulation 2. (T_i: the proposed exact methods based on $T_i(\mu)$, $i = 1, 2, 3, 4$; D-L: DerSimonian–Laird method; HE: Hedges method; H-S: Hunter–Schmidt method; S-J: Sidik–Jonkman method; ML: maximum-likelihood estimator; REML: restricted maximum-likelihood estimator; EB: empirical Bayes estimator; F-P(T_i): Follmann–Proschan permutation method based on $T_i(\mu)$, $i = 2$ or 4).

	$K = 8$			$K = 12$			$K = 16$		
Method	Cov prob	Length (std)	avg.t	Cov prob	Length (std)	Avg.t	Cov prob	Length (std)	Avg.t
D-L	0.917	4.62 (3.88)	0.006	0.923	3.87 (2.98)	0.006	0.929	3.43 (2.49)	0.006
HE	0.898	4.81 (4.14)	0.006	0.904	4.00 (2.87)	0.006	0.914	3.50 (2.37)	0.006
H-S	0.896	4.12 (3.19)	0.006	0.902	3.60 (2.66)	0.006	0.917	3.25 (2.30)	0.005
S-J	0.971	5.68 (3.88)	0.006	0.971	4.73 (2.66)	0.006	0.974	4.14 (2.18)	0.005
ML	0.893	4.24 (3.78)	0.012	0.906	3.63 (2.71)	0.012	0.917	3.27 (2.26)	0.010
REML	0.914	4.64 (4.05)	0.011	0.921	3.86 (2.83)	0.012	0.930	3.42 (2.33)	0.011
EB	0.918	4.79 (4.07)	0.010	0.923	3.98 (2.83)	0.011	0.932	3.50 (2.33)	0.009
T_1	0.952	7.09 (4.99)	0.010	0.941	5.14 (3.63)	0.048	0.946	4.02 (1.43)	0.816
T_2	0.952	7.09 (4.99)	0.038	0.942	4.89 (1.73)	0.418	0.948	4.02 (1.34)	1.35
T_3	0.962	6.86 (4.43)	0.031	0.952	4.71 (1.46)	0.025	0.953	3.75 (1.00)	0.023
T_4	0.953	6.29 (4.26)	0.044	0.946	4.54 (2.45)	0.416	0.,948	3.81 (1.98)	1.31
F-P(T_2)	0.951	6.45 (4.09)	0.857	0.940	4.52 (1.87)	13.7	0.944	3.74 (1.20)	362.6
F-P(T_4)	0.951	6.18 (4.26)	0.833	0.945	4.51 (3.02)	12.5	0.952	3.78 (2.21)	364.1

In conclusion, the simulation results demonstrate that when K is small, most conventional methods almost constantly fail to achieve the nominal coverage level except S-J method, whose actual coverage level can be erratic: sometimes substantially higher and sometimes lower than the nominal level. In contrast, the empirical performance of the proposed exact CIs is much more reliable in terms of both coverage level and the interval length.

4.2. Real data example: effect of statins in cholesterol reduction

Statins are the first line choices for reducing high blood cholesterol level, which increases the cardiovascular risk. There are ample evidences on the benefit of statins for patients with a history of cardiovascular disease. However, it is not entirely clear whether statins should be used as the primary prevention for people without history of cardiovascular disease. A systematic review is conducted to evaluate both the potential benefit and risk of statins for the primary prevention. Eighteen randomized controlled studies with 56,934 patients comparing statins with usual care are identified. The details of those studies are reported in [30]. Here, we focus on the meta-analysis for estimating the effect of statins in reducing the cholesterol level among participants without history of cardiovascular disease. Fourteen studies contributed data on measurements of total cholesterol level. The study specific results from these 14 studies are presented in Table 3. In the meta-analysis, we let Y_i be the observed group difference in changes of total cholesterol level during the follow-up. The within-study variance σ_i^2 can be calculated from the reported 95% CI reported in [30]. Because of the clear study heterogeneity, we adopt the random-effects model (1). The parameter of interest, μ_0, presents the mean difference in the total cholesterol level between statin and usual care groups. We first construct the 95% equal-tailed CIs of μ_0 based on D-L method. The resulting CI is $(-1.36, -0.74)$ with a point estimator of -1.05 indicating highly statistically significant treatment effect. Here, the point estimate is the μ value that yields the largest p value for testing $H_0 : \mu_0 = \mu$, i.e. the least significant testing result. When there are multiple μs with the same largest p value, the point estimator is set to be their average. We then construct the proposed exact CIs, which are

Table 3. The average change in total cholesterol levels by treatment arm and the observed treatment effects for 14 randomized clinical trails in the statins example.

Study	Statin group Mean (SD)	Control group Mean (SD)	Mean difference Mean (95% CI)
MEGA study	−0.72 (0.8)	−0.14 (0.8)	−0.58 (−0.62, −0.54)
ASPEN 2006	−0.51 (0.8)	−0.04 (0.8)	−0.47 (−0.54, −0.40)
CAIUS 1996	−1.01 (1.04)	0.18 (0.87)	−1.19 (−1.41, −0.97)
CARDS 2008	−1.24 (0.84)	−0.07 (0.87)	−1.17 (−1.23, −1.11)
CELL A 1996	−0.89 (0.86)	−0.18 (0.72)	−0.71 (−0.92, −0.50)
CELL B 1996	−0.86 (2.01)	−0.07 (0.64)	−0.93 (−1.31, −0.55)
CERDIA 2004	−1 (0.86)	0.14 (0.85)	−1.14 (−1.39, −0.89)
Derosa 2003	−1.63 (0.51)	−0.83 (0.58)	−0.80 (−1.11, −0.49)
HYRIM 2007	−0.56 (0.12)	0 (0.11)	−0.56 (−0.61, −0.51)
JUPITER 2008	−1.1 (0.8)	0.8 (0.8)	−1.90 (−1.92, −1.88)
KAPS 1995	−1.5 (0.66)	0 (0.66)	−1.50 (−1.63, −1.37)
METEOR 2010	−2.02 (0.77)	0 (0.7)	−2.02 (−2.13, −1.91)
PHYLLIS 2004	−1.1 (0.07)	−0.13 (0.06)	−0.97 (−0.98, −0.96)
PREVEND IT 2004	−1 (1)	−0.2 (1.05)	−0.80 (−0.94, −0.66)

Table 4. The point estimates (Est) and 95% CIs for the parameter of interest in real data examples under the Gaussian random-effects model.

	Est (95% CI)	
Method	Statins	BCG vaccine
D-L	−1.05 (−1.36, −0.74)	0.49 (0.26, 0.91)
$T_1(\mu)$	−1.05 (−1.90, −0.56)	0.78 (0.23, 1.01)
$T_2(\mu)$	−0.96 (−1.19, −0.58)	0.50 (0.23, 1.01)
$T_3(\mu)$	−0.97 (−1.35, −0.76)	0.49 (0.23, 1.01)
$T_4(\mu)$	−1.05 (−1.34, −0.76)	0.49 (0.24, 1.00)

reported in Table 4. The exact CIs based on $T_3(\mu)$ and $T_4(\mu)$ are almost identical to that derived from the D-L method. However, the validity of our proposals does not rely on the large sample approximation with 14 studies.

4.3. Real data example: efficacy of BCG vaccine

BCG, or Bacille Calmette-Guerin, is a vaccine for tuberculosis (TB) prevention. In the United States, BCG vaccination is currently recommended for certain groups of people, such as children who have a negative tuberculin skin test and who are exposed to infection risk. On BCG's efficacy, more than one thousand articles or abstracts have been published. To combine the information from multiple sources, [6] conducted a meta-analysis using the random-effects model with D-L method. However, the number of studies involved is relatively small and therefore the asymptotical inference result may not be reliable. In this section, the proposed methods in Sections 2 and 3 are applied to make exact inferences.

Specifically, we consider eight independent 2 × 2 tables (see Table 5), formed by pairs of independent binomials $(X_{i1}, X_{i0}), i = 1, \ldots, 8$, representing the number of TB incidence in the treatment (BCG-vaccinated) and control (BCG-nonvaccinated) groups. Denote the corresponding sample sizes as (N_{i1}, N_{i0}) with the underlying true event rates (π_{i1}, π_{i0}). For ith study, the log-transformation of the study specific odds ratio (OR) is defined as

$$\theta_i = \log\left\{\frac{\pi_{i1}/(1-\pi_{i1})}{\pi_{i0}/(1-\pi_{i0})}\right\}$$

and the observed log-OR is

$$Y_i = \log\left\{\frac{\mathrm{X_{i1}/(N_{i1}-X_{i1})}}{\mathrm{X_{i0}/(N_{i0}-X_{i0})}}\right\}.$$

Table 5. The number of TB cases and group size by treatment arm in the BCG vaccine example. See detailed data sources in [6].

	BCG-vaccine group		Control group	
Study	# of TB	N_1	# of TB	N_0
Canada	6	306	29	313
N USA	4	123	11	139
Chicago	17	1716	65	1665
Georgia I	5	2498	3	2341
Georgia II	27	16,913	29	17,854
UK	62	13,598	248	12,867
South Africa	29	7429	45	7277
Madras	505	88,391	499	88,391

Based on large sample approximations, we first apply the random-effects model (1) assuming both $Y_i \mid \theta_i$ and θ_i are Gaussian random variables. For ith study, the within-study variance σ_i^2 is set as its consistent estimator.

$$\sigma_i^2 = \frac{1}{X_{i1}} + \frac{1}{N_{i1} - X_{i1}} + \frac{1}{X_{i0}} + \frac{1}{N_{i0} - X_{i0}}.$$

We first construct the 95% equal-tailed CIs of μ_0, which is the expectation of the log-transformed study-specific OR, using the conventional D-L and our proposed methods. Then, we obtain the 95% CIs of $\exp(\mu_0)$, which is the median of the distributions for study-specific OR's, together with its point estimate. If the upper end of the constructed CI is less than 1, we may conclude that the BCG vaccine significantly reduces the risk of TB. The results are also reported in Table 4. The upper bound of the CI based on D-L method is apparently lower than others, which implies that D-L method may overestimate the significance of its efficacy. Among the proposed test statistics, $T_4(\mu)$ yields the narrowest CI with an upper end of 1.00. Furthermore, the CIs are closely related to the two-sided test $H_0 : \mu_0 = \mu$ versus $H_A : \mu_0 \neq \mu$ as we discussed before. The test statistics and corresponding p value for different μ are shown in Figure 2.

In the BCG example, we are especially interested in testing the null hypothesis that there is no treatment effect, that is, the random vector (π_{i1}, π_{i0}) has the same distribution as that of (π_{i0}, π_{i1}). Note that all the trials in the current example are approximately balanced with $n_{i1} \approx n_{i0}$, implying that under the aforementioned null hypothesis,

$$\Pr(Y_i \leq 0) \geq 1/2 \text{ and } \Pr(Y_i \geq 0) \geq 1/2.$$

Thus, we can apply the development in Section 3 to perform the exact test for this hypothesis without Gaussian assumptions for either $Y_i|\theta_i$ or θ_i. We consider the test

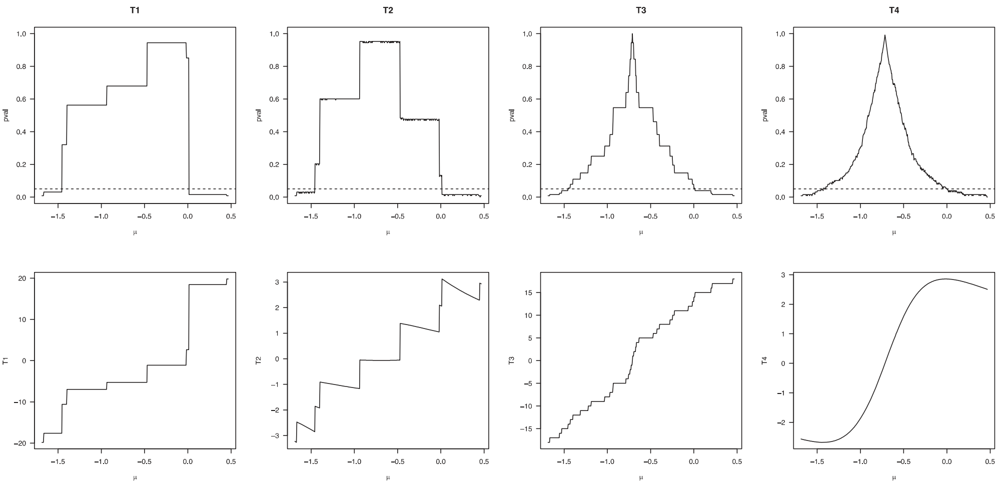

Figure 2. Based on $T_1(\mu)$, $T_2(\mu)$, $T_3(\mu)$ and $T_4(\mu)$, the p values (first row) and values of test statistics (second row) corresponding to different values of log(OR) for the BCG meta-analysis under Gaussian random-effects model.

statistics $\left\{\tilde{T}_{p\hat{w}}(0), \tilde{Z}_{q\hat{w}}(0)\right\}$ with two sets of weights.

$$\hat{w}_i = \sigma_i^{-1} or\ n_i^{1/2},\ \text{for } i = 1, \ldots, K.$$

The resulting p values are 0.078 and 0.106, respectively. We have also performed the tests based on D-L method and $T_3(\mu)$ assuming the commonly used Gaussian random-effects model for comparisons. The p value is 0.024 based on D-L method and 0.047 based on $T_3(\mu)$. Therefore, after relaxing the Gaussian assumption, the significance level for the efficacy of BCG vaccine reduces from approximately 5% to 10%.

5. Discussion

In this paper, we have proposed the exact inference procedures for fixed-effects and random-effects models in meta-analysis. The proposed CIs guarantee the coverage probabilities, while many other asymptotic methods including the commonly used D-L method may be invalid, especially when the number of studies is not large. Our proposals have a tight relationship with the nonparametric inference procedure by [31], whose validity does not require parametric distribution assumption for θ_i but is only justified asymptotically.

The normal assumption of θ_i may be restrictive and in general is difficult to verify especially when K is small. The proposed method slightly alleviates this concern by ensuring the validity of the inference procedure when the distribution of θ_i is symmetric at μ_0. On the other hand, when K is small and the information is limited, restrictive parametric assumptions are not avoidable for effectively summarizing the data via a statistical model. For example, even if all the $\sigma_i = 0$ and we observed θ_i's precisely, the statistical inference still requires a parametric model if K is small. Therefore, while we are making statistical inference under necessary model assumptions, we need to be cautious in interpreting the results when K is small.

For smaller K, say $K < 10$, the exact method can be conservative and generates wider CIs than the D-L method. In order to obtain narrower CIs, we propose specific choice of weights in $T_w(\mu)$ and $T_{\hat{w}}(\mu)$, together with $T_{\text{HL}}(\mu)$ (more specifically, $T_3(\mu)$). As demonstrated in the simulation studies, $T_3(\mu)$ and $T_4(\mu)$ can provide narrower CIs maintaining at least the same coverage levels of $T_1(\mu)$ and $T_2(\mu)$-based CIs. Besides, $T_3(\mu)$ and $T_4(\mu)$-based CIs are shown to be robust, although they can be a little more conservative in some cases.

The computation is a big obstacle for using many exact inference methods in practice. To construct the exact CIs, fast and easy-to-implement computational algorithms are proposed. Although the Follmann–Proschan's permutation CIs can perform similarly as $T_4(\mu)$-based CIs, our methods can significantly reduce the computational burden, especially when K is more than 10. We recommend to use our proposed methods for K between 6 and 20 (especially, when $K < 10$) or in the presence of substantial study heterogeneity, e.g. $I_2 \geq 50\%$. Note that the smallest two-sided p value of the proposed method is $1/2^{K-1}$ and we need $K \geq 6$ to generate significant result at the level of 0.05.

For the Behrens–Fisher problem, it has been shown that there is no exact test that is also the most powerful for all values of the variances [32]. We anticipate the same statement will still hold for the inference problems under the more complicated settings we considered. Although it is out of the scope of the current article, it will be of interest to investigate whether the statement is correct in future research. Regardless, our proposals here provide a simple and valid inference procedure for general meta-analysis problems that are often encountered in medical research.

Acknowledgments

The authors thank Professor L. J. Wei for constructive comments on the paper and kindly providing the data for the statins example. The first and fourth authors gratefully acknowledge support from NSF grant: DMS1513483. The second author acknowledges support from NIH grant: R01 HL089778-05.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

NSF [grant number DMS1513483];NIH [grant number R01 HL089778-05].

Notes on contributors

Sifan Liu is postdoctoral associate, Department of Statistics and Biostatistics, Rutgers University, Piscataway, NJ 08854.

Lu Tian is associate professor, Department of Biomedical Data Science, Stanford University, Palo Alto, CA 94305.

Steve Lee is statistical consultant, South San Francisco, CA 94080.

Min-ge Xie is distinguished professor, Department of Statistics and Biostatistics, Rutgers University.

ORCID

Sifan Liu http://orcid.org/0000-0003-4473-1262

References

[1] DerSimoninan R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986;7: 177–188.

[2] IntHout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard Dersimonian-Laird method. BMC Med Res Methodol. 2014;14(25):1–12.

[3] Brockwell S, Gordon I. A comparison of statistical methods for meta-analysis. Stat Med. 2001;20:825–840.

[4] Macgregor G. BCG: bad news from India. Lancet. 1980;315(8159):73–74.

[5] BCG vaccination after the Madras study. Lancet. 1981;317(8215):309–310.

[6] Colditz G, Brewer T, Berkey C, et al. Efficacy of BCG vaccine in the prevention of tuberculosis. Meta-analysis of the published literature. JAMA. 1994;271(9):698–702.

[7] Hardy R, Thompson SG. A likelihood approach to meta-analysis with random effects. Stat Med. 1996;15:619–629.

[8] Normand S-LT. Tutorial in biostatistics meta-analysis: formulating, evaluating, combining and reporting. Stat Med. 1999;18(3):321–359.

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

[10] Sidik K, Jonkman JN. A simple confidence interval for meta-analysis. Stat Med. 2002;21: 3153–3159.

[11] Henmi M, Copas J. Confidence intervals for random effects meta-analysis and robustness to publication bias. Stat Med. ec 2010;29:2969–2983.

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

[13] Cox DR. Some problems connected with statistical inference. Ann Math Stat. 1958;29: 357–372.

[14] Efron B. Bayes and likelihood calculations from confidence intervals. Biometrika. 1993; 80:3–26.

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

[16] Xie M, Liu RY, Damaraju CV, et al. Incorporating external information in analyses of clinical trials with binary outcomes. Ann Appl Stat. 2013;7:342–368.

[17] Yang G, Liu D, Liu R, et al. Efficient network meta-analysis: a confidence distribution approach. Stat Methodol. 2014;20:105–125.

[18] Claggett B, Xie M, Tian L. Meta-analysis with fixed, unknown, study-specific parameters. J Am Stat Assoc. 2014;109:1660–1671.

[19] Tian L, Cai T, Pfeffer MA, et al. Exact and efficient inference procedure for meta-analysis and its application to the analysis of independent 2×2 tables with all available data but without artificial continuity correction. Biostatistics. 2009;10(2):275–281.

[20] Liu D, Liu R, Xie M. Exact meta-analysis approach for discrete data and its application to 2×2 tables with rare events. J Am Stat Assoc. 2014;109(508):1450–1465.

[21] Yang G, Liu D, Wang J, et al. Meta-analysis framework for exact inferences with application to the analysis of rare events. Biometrics. 2016;72(4):1378–1386.

[22] Follmann DA, Proschan MA. Valid inference in random effects meta-analysis. Biometrics. 1999;55(3):732–737.

[23] Dudewicz EJ, Ahmed SU. New exact and asymptotically optimal solution to the Behrens-Fisher problem, with tables. Am J Math Manage Sci. 1998;18:359–426.

[24] Chapman DG. Some two sample tests. Ann Math Stat. 1950;21:601–606.

[25] Prokof'yev VN, Shishkin AD. Successive classification of normal sets with unknown variances. Radio Eng Electron Phys. 1974;19(2):141–143.

[26] Lin DY, Zeng D. On the relative efficiency of using summary statistics versus individual level data in meta-analysis. Biometrika. 2010;97:321–332.

[27] Zeng D, Lin D. On random-effects meta-analysis. Biometrika. 2015;102(2):281–294.

[28] Hodges JL, Lehmann EL. Estimation of location based on ranks. Ann Math Stat. 1963;34 (2):598–611.

[29] DerSimonian R, Kacker R. Random-effects model for meta-analysis of clinical trials: an update. Contemp Clin Trials. 2007;28:105–114.

[30] Taylor F, Huffman MD, Macedo AF, et al. Statins for the primary prevention of cardiovascular disease. Cochrane Database Syst Rev. 2013;1(CD004816):1–20.

[31] Wang R, Tian L, Cai T, et al. Nonparametric inference procedure for percentiles of the random effects distribution in meta-analysis. Ann Appl Stat. 2010;4(1):520–532.

[32] Linnik YV. Latest investigations on Behren-Fisher problem. Sankyha Ser A. 1966;28(1):15–24.

Appendices

Appendix A. Proof of Theorem 2.1

Proof. One the one hand, $\Pr\{T_w(\mu_0) \le t\} = F_w^*(t)$. Define the inverse function,

$$F_w^{*-1}(\nu) = \sup\{t : F_w^*(t) \le \nu\}, \nu \in [0, 1];$$

which implies $F_w^*\{F_w^{*-1}(\nu)\} \le \nu$. Thus, we have $Pr\left[F_w^*\{T_w(\mu_0)\} \le \frac{\alpha}{2}\right] = Pr\left\{T_w(\mu_0) \le F_w^{*-1}\left(\frac{\alpha}{2}\right)\right\} = F_w^*\left\{F_w^{*-1}\left(\frac{\alpha}{2}\right)\right\} \le \frac{\alpha}{2}$. On the other hand, for $Pr\{T_w(\mu_0) \ge t\} = S_w^*(t)$, we may define

$$S_w^{*-1}(\nu) = \inf\{t : S_w^*(t) \le \nu\}, \nu \in [0, 1],$$

and $\Pr\left[S_w^*\{T_w(\mu_0)\} \le \alpha/2\right] = \Pr\left[T_w(\mu_0) \ge S_w^{*-1}(\alpha/2)\right] = S_w^*\{S_w^{*-1}\}(\alpha/2)\} \le \alpha/2$.

As a result, $\Pr(\mu_0 \in C_{w\alpha}) = 1 - \Pr(\mu_0 \notin C_{w\alpha}) = 1 - \Pr\left[F_w^*\{T_w(\mu_0)\} \le \alpha/2\right] - \Pr\left[S_w^*\{T_w(\mu_0)\} \le \alpha/2\right] \ge 1 - \alpha/2 - \alpha/2 = 1 - \alpha$.

Appendix B. Proof of Theorem 2.2

Proof. Based on the proof of Theorem 2.1, on the one hand, let $F_{\hat{w}}^*(t, \mu_0) = \Pr\{T_{\hat{w}}{}^*(\mu_0) \le tY_1, \dots, Y_K\} = Pr\{T_{\hat{w}}{}^*(\mu_0) \le t|\hat{w}_1(\mu_0), \dots, \hat{w}_K(\mu_0)\}$ and $F_{\hat{w}}^{*-1}(\nu, \mu_0) = \sup\{t : F_{\hat{w}}{}^*(t, \mu_0) \le \nu\}$, for $\nu \in [0, 1]$. Since $\{I(Y_1 < \mu_0), \dots, I(Y_K < \mu_0)\}$ and $\{\hat{w}_1(\mu_0), \dots, \hat{w}_K(\mu_0)\}$ are independent,

$$\begin{aligned}
&\Pr\left[\Pr\{T_{\hat{w}}{}^*(\mu_0) \le T_{\hat{w}}(\mu_0)\} \le \frac{\alpha}{2}|\hat{w}_1(\mu_0), \dots, \hat{w}_K(\mu_0)\right] \\
&= \Pr\left[F_{\hat{w}}{}^*\{T_{\hat{w}}(\mu_0), \mu_0\} \le \frac{\alpha}{2}\hat{w}_1(\mu_0), \dots, \hat{w}_K(\mu_0)\right] \\
&= \Pr\left[T_{\hat{w}}(\mu_0) \le F_{\hat{w}}^{*-1}\left(\frac{\alpha}{2}, \mu_0\right)\hat{w}_1(\mu_0), \dots, \hat{w}_K(\mu_0)\right] \\
&= F_{\hat{w}}^*\left[F_{\hat{w}}^{*-1}\left(\frac{\alpha}{2}, \mu_0\right), \mu_0\right] \le \frac{\alpha}{2}\bigg].
\end{aligned}$$

On the other hand, $\Pr\left[\Pr\{T_{\hat{w}}{}^*(\mu_0) \ge T_{\hat{w}}(\mu_0)\} \le \frac{\alpha}{2}\hat{w}_1(\mu_0), \dots, \hat{w}_K(\mu_0)\right] \le \frac{\alpha}{2}$. Following the same arguments for Theorem 2.1, $\Pr[\mu_0 \in C_{\hat{w}}\alpha\hat{w}_1(\mu_0), \dots, \hat{w}_K(\mu_0)] \ge 1 - \alpha$, which implies that $\Pr\{\mu_0 \in C_{\hat{w}}\alpha\} \ge 1 - \alpha$.