RESEARCH ARTICLE

Combined *P*-value Functions for Compatible Effect Estimation and Hypothesis Testing in Drug Regulation

Samuel Pawel | **Małgorzata Roos** | **Leonhard Held**

Epidemiology, Biostatistics and Prevention Institute (EBPI), Center for Reproducible Science (CRS), University of Zurich, Zurich, Switzerland

Correspondence
Samuel Pawel, Epidemiology, Biostatistics and Prevention Institute, Hirschengraben 84, 8001, Zurich, Switzerland.
Email: samuel.pawel@uzh.ch

Abstract

The two-trials rule in drug regulation requires statistically significant results from two pivotal trials to demonstrate efficacy. However, it is unclear how the effect estimates from both trials should be combined to quantify the drug effect. Fixed-effect meta-analysis is commonly used but may yield confidence intervals that exclude the value of no effect even when the two-trials rule is not fulfilled. We systematically address this by recasting the two-trials rule and meta-analysis in a unified framework of combined *p*-value functions, where they are variants of Wilkinson's and Stouffer's combination methods, respectively. This allows us to obtain compatible combined *p*-values, effect estimates, and confidence intervals, which we derive in closed-form. Additionally, we provide new results for Edgington's, Fisher's, Pearson's, and Tippett's *p*-value combination methods. When both trials have the same true effect, all methods can consistently estimate it, although some show bias. When true effects differ, the two-trials rule and Pearson's method are conservative (converging to the less extreme effect), Fisher's and Tippett's methods are anti-conservative (converging to the more extreme effect), and Edgington's method and meta-analysis are balanced (converging to a weighted average). Notably, Edgington's confidence intervals asymptotically always include individual trial effects, while meta-analytic confidence intervals shrink to a point at the weighted average effect. We conclude that all of these methods may be appropriate depending on the estimand of interest. We implement combined *p*-value function inference for two trials in the R package `twotrials`, allowing researchers to easily perform compatible hypothesis testing and effect estimation.

KEYWORDS
Confidence interval, estimand, median estimate, meta-analysis, two-trials rule

1 | INTRODUCTION

The "two-trials rule" in drug regulation requires "*at least two adequate and well-controlled studies, each convincing on its own*" for the demonstration of drug efficacy and subsequent market approval [1, p.3]. This criterion reflects the need for "substantiation" and "replication" of scientific results [2, p.8], and is typically implemented by requiring the *p*-values from the two trials to be statistically significant at the conventional (one-sided) $\alpha = 0.025$ level. However, this procedure alone does not provide a combined effect estimate nor a confidence interval (CI), and it has been suggested to pool the estimates with fixed-effect meta-analysis for this purpose [3,4,5]. Yet, the meta-analytic CI and point estimate are not always compatible with the two-trials rule. The meta-analytic CI may exclude the null value while the two-trials rule is not fulfilled, leading to discrepancies that are difficult to interpret and communicate.

The results from the two RESPIRE trials [6,7,8] in Table 1 illustrate this phenomenon. While the *p*-value for the null hypothesis of no effect from RESPIRE 1 is $p = 0.004 < 0.025$, the *p*-value from RESPIRE 2 is $p = 0.144 > 0.025$. Hence, the two-trials rule is not fulfilled at $\alpha = 0.025$. At the same time, the 95% CI for the log rate ratio based on combining the trials' log rate ratio effect estimates with fixed-effect meta-analysis ranges from –0.58 to –0.08 and thus excludes the value of 0.

A first attempt at resolving the apparent paradox could be to realize that the confidence level of the CI does not align with the level of the implicit test underlying the two-trials rule. Since the two-trials rule decision is based on two independent tests at level

TABLE 1 Results from the RESPIRE trials regarding the effect of ciprofloxacin after 14 days for the treatment of non-cystic fibrosis bronchiectasis[6,7,8].

	Log rate ratio	Confidence interval (95%)	*P*-value (one-sided)
RESPIRE 1	–0.49	–0.85 to –0.13	0.004
RESPIRE 2	–0.18	–0.53 to 0.16	0.144
Meta-analysis	–0.33	–0.58 to –0.08	0.004

$\alpha = 0.025$, the overall test is at level $\alpha^2 = 0.000625$, thus one could instead take a $(1 - 2\,\alpha^2) \times 100\% = 99.875\%$ meta-analytic CI[9,10]. For the RESPIRE trials, this would lead to a meta-analytic 99.875% CI from –0.71 to 0.05 which includes the value of 0 and hence aligns with the two-trials rule decision. However, the level $\alpha = 0.025$ is arbitrary and it would be desirable to have a CI that is compatible with the two-trials rule for any level, which is still not the case. For example, for $\alpha = 0.05$, the two-trials rule is still not fulfilled, while the $(1 - 2\,\alpha^2) \times 100\% = 99.5\%$ meta-analytic CI from –0.66 to –0.01 excludes zero.

Despite the widespread use of the two-trials rule in regulatory decision-making[11], it remains unclear how point and interval estimation should be reconciled with it. This paper aims to resolve this issue with a new approach. The key idea is to look at both the two-trials rule and meta-analysis from the perspective of *p*-value functions[12,13,14,15,16] and *p*-value combination methods[17,18,19,20,21]. The two-trials rule can be understood as a combined *p*-value function based on the squared maximum of two *p*-values[22] which is a special case of Wilkinson's combination method[23], while meta-analysis corresponds to the combined *p*-value function based on Stouffer's *p*-value combination method[24] with suitable weights. Both can be used to obtain combined *p*-values for the null hypothesis of no effect, CIs, and point estimates. These quantities are compatible in the sense that the (two-sided) *p*-value for a null value is less than α if and only if the null value is excluded by the $(1 - \alpha) \times 100\%$ CI, and that the point estimate is included in the CI at any confidence level $(1 - \alpha) \in (0, 1)$. However, as we will show, the two methods implicitly target different estimands, which explains their different behaviors, and highlights the need to choose the method depending on the scientific question and corresponding estimand of interest. Moreover, the combined *p*-value function pespective suggests considering alternative *p*-value combination methods, for example, Edgington's method based on the sum of *p*-values[25] or Fisher's method based on the product of *p*-values[26]. All these *p*-value combination methods have been studied before in terms of hypothesis testing properties, such as admissibility or monotonicity[27,17]. In this paper, we take an alternative estimation perspective motivated by practical issues in drug regulation.

This paper is organized as follows: We begin by summarizing the general theory of combined *p*-value functions (Section 2), followed by investigating combined *p*-value functions based on the two-trials rule (Section 2.1), meta-analysis (Section 2.2), Tippett's method (Section 3.1), Fisher's and Pearson's methods (Section 3.2), and Edgington's method (Section 3.3) in more detail. For each, we derive corresponding point and interval estimates and investigate their properties. Results from two pairs of clinical trials are analyzed to illustrate the characteristics of the methods (Section 4). Extensions to more than two trials are discussed in Section 5. The paper ends with concluding discussions, limitations, and an outlook for future research (Section 6). Appendix A illustrates our R package `twotrials` for conducting *p*-value function inference, while Appendix B provides additional technical details.

2 | COMBINED *P*-VALUE FUNCTIONS

Suppose that two trials yield the effect estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ with corresponding standard errors σ_1 and σ_2, each estimate quantifying the effect of the treatment in the corresponding trial. Typically, it is reasonable to assume that the effect estimates (after suitable transformation) are approximately normally distributed around the trial-specific true effects θ_1 and θ_2 with variance equal to their squared standard error, i.e., $\hat{\theta}_i \mid \theta_i \sim \mathrm{N}(\theta_i, \sigma_i^2)$ for $i \in \{1, 2\}$. One-sided *p*-values can then be computed by

$$p_i(\mu) = \begin{cases} 1 - \Phi(Z_i) & \text{for } H_{1i}\colon\ \theta_i > \mu \text{ (alternative = "greater")} \\ \Phi(Z_i) & \text{for } H_{1i}\colon\ \theta_i < \mu \text{ (alternative = "less")} \end{cases} \tag{1}$$

with *z*-values

$$Z_i = \frac{\hat{\theta}_i - \mu}{\sigma_i},$$

cumulative distribution function of the standard normal distribution $\Phi(\cdot)$, null value μ, and alternative hypothesis H_{1i} chosen based on the orientation of the effect. For example, if a positive effect indicates treatment benefit, the alternative "greater" would

be chosen. We will not consider p-values with two-sided alternatives here, as the hypotheses tested in clinical trials usually have a well-defined direction. Moreover, combined p-value functions based on two-sided p-values can behave irregularly, e.g., they can be non-monotone so that the resulting confidence sets consist of empty or disjoint intervals, which is unintuitive and hard to communicate[28].

A combined p-value function $p(\mu)$ is then defined by the function g

$$p(\mu) = g\left(p_1(\mu), p_2(\mu)\right),$$

which combines the individual p-value functions $p_1(\mu)$ and $p_2(\mu)$ into a p-value function $p(\mu)$, which is a valid p-value function in the sense of having a uniform distribution for a particular μ if both $p_1(\mu)$ and $p_2(\mu)$ are also uniformly distributed for that μ[13,28]. A two-sided $(1-\alpha)\times 100\%$ CI can then be obtained by determining the null values μ for which the p-value function is equal to $\alpha/2$ and $1-\alpha/2$. The so-called median estimate is given by the null value μ for which the p-value function equals 1/2[29]. To obtain these quantities, it is useful to define a "combined estimation function"

$$\hat{\mu}(a) = \{\mu : p(\mu) = a\}$$

which is the inverse of the combined p-value function. It returns the median estimate when setting $a = 1/2$, while the limits of a $(1-\alpha)\times 100\%$ CI are obtained from $a = \alpha/2$ and $a = 1-\alpha/2$, respectively. As we will show, combined estimation functions (and hence the median estimate and any CI) are available in closed-form for several combined p-value functions, including the two-trials rule and meta-analysis.

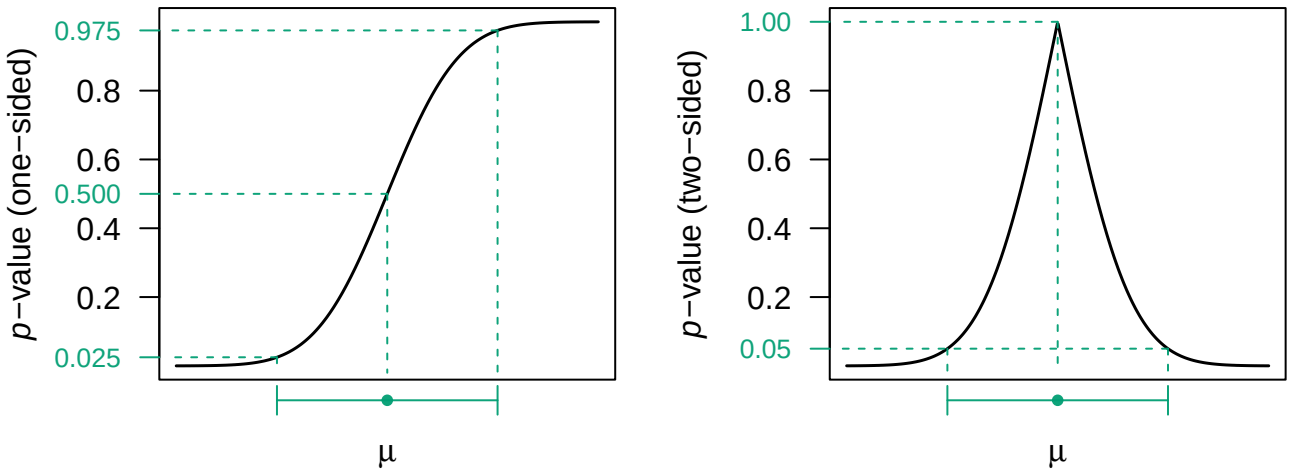


FIGURE 1 Illustration of a one-sided p-value function with alternative = "greater" (left plot), corresponding two-sided p-value function (right plot), and corresponding 95% CI and median estimate.

In practice, it is informative to plot the p-value function for a range of null values μ, see the left plot in Figure 1. For this purpose, it may also be converted to a two-sided p-value function using the transformation $2\min\{p(\mu), 1-p(\mu)\}$, known as "centrality function"[13]. Such a two-sided p-value function then peaks at the median estimate, and it can be thresholded at α to conveniently read off the $(1-\alpha)\times 100\%$ CI[28], see the right plot in Figure 1.

When both trials have the same underlying true effect ($\theta_1 = \theta_2 = \theta$), sometimes called "one population assumption" or "homogeneity"[30,31], a CI based on a combined p-value function has correct coverage and the median estimate is median unbiased for the true common effect θ, i.e., the probability of the median estimate being greater than θ is equal to the probability of it being smaller than θ see e.g.,[13]. However, it is unclear how other operating characteristics (e.g., mean bias or CI width) behave for different combined p-value functions g, and how they behave when the true effects are not the same ($\theta_1 \neq \theta_2$), known as "two populations assumption" or "heterogeneity"[30,31]. In the following, we will investigate this in detail for the two-trials rule,

TABLE 2 Summary of combined p-value functions and corresponding estimation functions. All are based on the alternative "greater". The median estimate is obtained from setting $a = 1/2$, while the limits of a $(1-\alpha) \times 100\%$ confidence interval (CI) are obtained from $a = \alpha/2$ and $a = 1-\alpha/2$, respectively.

Method	Combined p-value function	Combined estimation function	Properties
Two-trials rule Maximum p-value, special case of Wilkinson's method, Section 2.1	$p_{2TR}(\mu) = \max\{p_1(\mu), p_2(\mu)\}^2$ R function `twotrials::p2TR`	$\hat{\mu}_{2TR}(a) = \min\{\hat{\theta}_1 + \sigma_1 z_{\sqrt{a}}, \hat{\theta}_2 + \sigma_2 z_{\sqrt{a}}\}$ R function `twotrials::mu2TR`	– Targets least extreme true effect (conservative) – Mean-biased when trials have the same true effects – CI shrinks to point with decreasing standard errors – Median estimate not equal to observed effect estimates when the same estimates in both trials – Median estimate standard error can be larger than trial standard errors
Fixed-effect meta-analysis Weighted Stouffer's method, inverse-normal method, Section 2.2	$p_{MA}(\mu) = 1 - \Phi(Z_{MA})$ with $Z_{MA} = \frac{\Phi^{-1}\{1-p_1(\mu)\}/\sigma_1 + \Phi^{-1}\{1-p_2(\mu)\}/\sigma_2}{\sqrt{1/\sigma_1^2 + 1/\sigma_2^2}}$ R function `twotrials::pMA`	$\hat{\mu}_{MA}(a) = \hat{\theta}_{MA} + \sigma_{MA} z_a$ with $\sigma^2_{MA} = 1/(1/\sigma_1^2 + 1/\sigma_2^2)$ $\hat{\theta}_{MA} = (\hat{\theta}_1/\sigma_1^2 + \hat{\theta}_2/\sigma_2^2)\,\sigma^2_{MA}$ R function `twotrials::muMA`	– Targets weighted average effect (inverse squared standard error weights) – Mean-unbiased when the same true effects – CI shrinks to point with decreasing standard errors – Median estimate equals observed effect estimates when the same estimates in both trials – Median estimate standard error cannot be larger than trial standard errors
Tippett's method Minimum p-value, special case of Wilkinson's method, Section 3.1	$p_T(\mu) = 1 - (1 - \min\{p_1(\mu), p_2(\mu)\})^2$ R function `twotrials::pTippett`	$\hat{\mu}_T(a) = \max\{\hat{\theta}_1 - \sigma_1 z_{\sqrt{1-a}}, \hat{\theta}_2 - \sigma_2 z_{\sqrt{1-a}}\}$ R function `twotrials::muTippett`	– Targets most extreme true effect (anti-conservative) – Mean-biased when the same true effects – CI shrinks to point with decreasing standard errors – Median estimate not equal to observed effect estimates when the same estimates in both trials – Median estimate standard error can be larger than trial standard errors
Fisher's method Product of p-values, Section 3.2	$p_F(\mu) = 1 - \Pr(\chi^2_4 \leq F)$ with $F = -2[\log\{p_1(\mu)\} + \log\{p_2(\mu)\}]$ R function `twotrials::pFisher`	$\hat{\mu}_F(a)$ not analytically available R function `twotrials::muFisher`	– Targets most extreme true effect (anti-conservative) – CI shrinks to point with decreasing standard errors – Median estimate not equal to observed effect estimates when the same estimates in both trials – Median estimate standard error can be larger than trial standard errors
Pearson's method Product of $1-p$-values, Section 3.2	$p_P(\mu) = \Pr(\chi^2_4 \leq K)$ with $K = -2[\log\{1 - p_1(\mu)\} + \log\{1 - p_2(\mu)\}]$ R function `twotrials::pPearson`	$\hat{\mu}_P(a)$ not analytically available R function `twotrials::muPearson`	– Targets least extreme true effect (conservative) – CI shrinks to point with decreasing standard errors – Median estimate not equal to observed effect estimates when the same estimates in both trials – Median estimate standard error can be larger than trial standard errors
Edgington's method Sum of p-values, Section 3.3	$p_E(\mu) = \begin{cases} E^2/2 & \text{if } 0 \leq E \leq 1 \\ 1 - (2-E)^2/2 & \text{if } 1 < E \leq 2 \end{cases}$ with $E = p_1(\mu) + p_2(\mu)$ R function `twotrials::pEdgington`	Median estimate analytically available $\hat{\mu}_E(a = 1/2) = \frac{\hat{\theta}_1/\sigma_1 + \hat{\theta}_2/\sigma_2}{1/\sigma_1 + 1/\sigma_2}$ $\hat{\mu}_E(a)$ not analytically available for $a \neq 1/2$ R function `twotrials::muEdgington`	– Targets weighted average effect (inverse standard error weights) – Mean-unbiased when the same true effects – CI asymptotically always includes both true effects (only shrinks to point when both are equal) – Median estimate equals observed effect estimates when the same estimates in both trials – Median estimate standard error can be larger than trial standard errors

meta-analysis, and four other types of combined p-value functions. As these investigations are somewhat technical, readers may choose to look only at the summary in Table 2 and then jump directly to the applications in Section 4.

2.1 | The two-trials rule (maximum method)

The two-trials rule is fulfilled if $\max\{p_1, p_2\} \leq \alpha$, or equivalently if

$$p_{2\text{TR}}(\mu) = \max\{p_1(\mu), p_2(\mu)\}^2 \leq \alpha^2. \tag{2}$$

The formulation using the squared maximum (2) may be preferable because $p_{2\text{TR}}(\mu)$ is a valid p-value, i.e., it has a uniform distribution if both $p_1(\mu)$ and $p_2(\mu)$ are also uniformly distributed for a particular μ[22]. The combined p-value function (2) is also a special case of Wilkinson's p-value combination method based on the rth smallest out of k p-values with $r = k = 2$[23]. This relationship can be used to generalize the two-trials rule to different settings while preserving type I error control at level α^2, for example, settings with three rather than two trials[32]. We will discuss such extensions to more than two trials in Section 5 and focus first on effect estimation for two trials.

2.1.1 | Effect estimation

In order to obtain a CI and a point estimate based on the two-trials rule, we can equate the combined p-value function (2) to some value $a \in (0, 1)$ and solve for the null value μ. This leads to the combined estimation function

$$\hat{\mu}_{2\text{TR}}(a) = \begin{cases} \min\{\hat{\theta}_1 + \sigma_1 z_{\sqrt{a}}, \hat{\theta}_2 + \sigma_2 z_{\sqrt{a}}\} & \text{for alternative = "greater"} \\ \max\{\hat{\theta}_1 - \sigma_1 z_{\sqrt{a}}, \hat{\theta}_2 - \sigma_2 z_{\sqrt{a}}\} & \text{for alternative = "less"} \end{cases} \tag{3}$$

with z_q the $q \times 100\%$ quantile of the standard normal distribution. For $a = 1/2$ the median estimate is obtained, while the limits of an $(1 - \alpha) \times 100\%$ CI can be obtained from $a = \alpha/2$ and $a = 1 - \alpha/2$.

Now assume that the standard errors of both trials are the same ($\sigma_1 = \sigma_2 = \sigma$) and the alternative is "greater". The median estimate is then

$$\hat{\mu}_{2\text{TR}}(1/2) = \min\{\hat{\theta}_1, \hat{\theta}_2\} + \sigma \underbrace{z_{\sqrt{1/2}}}_{0.54} \tag{4}$$

and the 95% CI is given by

$$\left[\min\{\hat{\theta}_1, \hat{\theta}_2\} + \sigma \underbrace{z_{\sqrt{0.025}}}_{-1}, \min\{\hat{\theta}_1, \hat{\theta}_2\} + \sigma \underbrace{z_{\sqrt{0.975}}}_{2.24}\right]. \tag{5}$$

Both seem counterintuitive. For instance, if the trial effect estimates are the same ($\hat{\theta}_1 = \hat{\theta}_2 = \hat{\theta}$), the median estimate (4) is shifted away from the observed estimate by $\sigma \times z_{\sqrt{1/2}} \approx \sigma \times 0.54$, and also the CI (5) is not centered around it. This is illustrated in Figure 2 (panels A and C), where the hypothetical trial effect estimates are identical, but the median estimates based on the two-trials rule (black) are larger. Moreover, the CI (5) is skewed in the sense that the distance between the upper limit and the median estimate is larger than the distance between the lower limit and the median estimate although the estimates are the same.

While this CI has correct coverage and the median estimate is median unbiased we may look at other operating characteristics. The expectation of the median estimate (4) can be derived to be

$$\text{E}\left[\hat{\mu}_{2\text{TR}}(1/2)\right] = \theta_1 \, \Phi\left(\frac{\theta_2 - \theta_1}{\sqrt{2}\,\sigma}\right) + \theta_2 \, \Phi\left(\frac{\theta_1 - \theta_2}{\sqrt{2}\,\sigma}\right) + \sigma \left\{z_{\sqrt{1/2}} - \sqrt{2}\,\phi\left(\frac{\theta_2 - \theta_1}{\sqrt{2}\,\sigma}\right)\right\} \tag{6}$$

where $\phi(\cdot)$ denotes the density function of the standard normal distribution, see Appendix B.1 for details. If the true effects from the two trials coincide ($\theta_1 = \theta_2 = \theta$) the first two terms of the expectation (6) reduce to the common effect θ, whereas the last term reduces to $\sigma \times (z_{\sqrt{1/2}} - 1/\sqrt{\pi}) \approx \sigma \times -0.019$. Hence, the median estimate with (4) is negatively biased, yet the bias vanishes as the standard error decreases. In a similar way, one can show that the median estimate for the alternative "less" is positively biased, so the median estimate from the two-trials rule exhibits a conservative bias in both cases.

Another interesting operating characteristic is the standard error of the median estimate, elaborated in more detiail in Appendix B.2. Notably, the standard error of the two-trials rule depends not only on the standard errors of the trials, but also on the true trial effects.

An intuitively desirable property is that the standard error of a combined estimate should not be larger than either of the trials' standard errors. Assuming again that the true trial effects and standard errors coincide, the two-trials rule satisfies, as the standard error takes the simple form

$$\sigma_{2TR} = \sigma\sqrt{1 - 1/\pi} \approx \sigma \times 0.83,$$

so is approximately 17% smaller than the standard errors of each individual trial. However, this is no longer the case when the trial standard errors differ, see Appendix B.2.

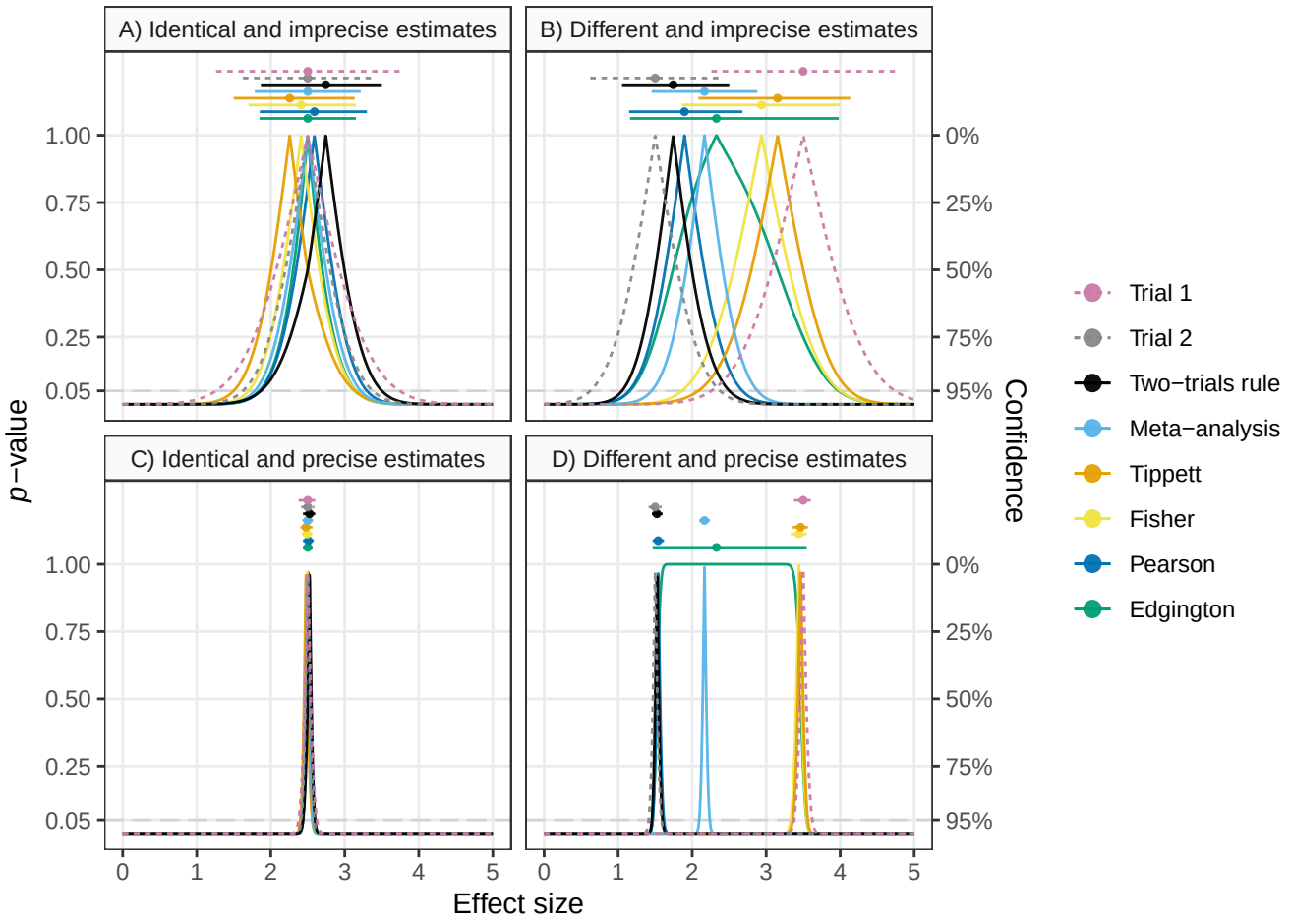


FIGURE 2 Four hypothetical pairs of effect estimates and standard errors from two trials. The standard errors are assumed to be of the form $\sigma = \sqrt{2/n}$. The sample size n in trial 1 is set to 5 (imprecise) or 500 (precise), and in trial 2 to twice the sample size of trial 1. The two-sided p-value functions of the individual trials (dashed lines), and the combined p-value functions (solid lines) based on the two-trials rule, fixed-effect meta-analysis, Tippett's, Fisher's, Pearson's, and Edgington's methods, and are shown along with the corresponding 95% CIs and median estimates (top). All p-values are based on the alternative "greater" and then converted to two-sided p-values via the centrality function $2\min\{p, 1-p\}$.

2.1.2 | Asymptotics

Suppose now that the sample size of the trials increases and in turn the standard errors of the effect estimates decrease toward zero. The combined estimation function (3) then converges to

$$\operatorname*{plim}_{\sigma_1,\sigma_2\downarrow 0} \hat{\mu}_{2\text{TR}}(a) = \begin{cases} \min\{\theta_1, \theta_2\} & \text{for alternative = "greater"} \\ \max\{\theta_1, \theta_2\} & \text{for alternative = "less",} \end{cases}$$

see Appendix B.3 for details. Hence, the median estimate ($a = 1/2$) and any CI limit ($a \neq 1/2$) approach $\min\{\hat{\theta}_1, \hat{\theta}_2\}$ or $\max\{\hat{\theta}_1, \hat{\theta}_2\}$, depending on the alternative hypothesis. This means that the CI shrinks to the more conservative of the two effects, while in case they coincide ($\theta_1 = \theta_2 = \theta$) it shrinks to the common effect θ. Both scenarios are illustrated in Figure 2: In case of very small standard errors and different effect estimates (panel D), the CI based on the two-trials rule (black) is tightly concentrated around the smaller effect estimate, while for identical effect estimates (panel C), it is tightly concentrated around the common effect estimate.

2.2 | Fixed-effect meta-analysis (Stouffer's method)

We will now compare the *p*-value function from the two-trials rule with its meta-analysis counterpart. The combined *p*-value based on fixed-effect meta-analysis is given by

$$p_{\text{MA}}(\mu) = \begin{cases} 1 - \Phi(Z_{\text{MA}}) & \text{for alternative = "greater"} \\ \Phi(Z_{\text{MA}}) & \text{for alternative = "less"} \end{cases} \tag{7}$$

with

$$Z_{\text{MA}} = \frac{Z_1/\sigma_1 + Z_2/\sigma_2}{\sqrt{1/\sigma_1^2 + 1/\sigma_2^2}} = \frac{\hat{\theta}_{\text{MA}} - \mu}{\sigma_{\text{MA}}} \tag{8}$$

where

$$\hat{\theta}_{\text{MA}} = \frac{\hat{\theta}_1/\sigma_1^2 + \hat{\theta}_2/\sigma_2^2}{1/\sigma_1^2 + 1/\sigma_2^2} \tag{9}$$

and

$$\sigma_{\text{MA}} = \frac{1}{\sqrt{1/\sigma_1^2 + 1/\sigma_2^2}}. \tag{10}$$

The first equation in (8) represents Stouffer's *p*-value combination method (after transforming *p*-values to *z*-values) using inverse standard errors as weights [19], whereas the second equation in (8) shows the corresponding representation via the meta-analytically pooled estimate (9) and standard error (10) [33]. While meta-analytic pooling could be extended to the random-effects model, this is typically not desired with only two studies for three reasons. First, the interest is in the true effects underlying the studies. Second, random-effects variance estimation is unreliable with only two studies. Finally, even if there is effect heterogeneity, fixed-effect meta-analysis is a valid procedure which estimates a well-defined average true effect [34].

2.2.1 | Effect estimation

To obtain meta-analytic CIs and point estimates we can also equate the *p*-value function (7) to a and solve for μ. This leads to the combined estimation function

$$\hat{\mu}_{\text{MA}}(a) = \begin{cases} \hat{\theta}_{\text{MA}} + \sigma_{\text{MA}}\, z_a & \text{for alternative = "greater"} \\ \hat{\theta}_{\text{MA}} - \sigma_{\text{MA}}\, z_a & \text{for alternative = "less".} \end{cases}$$

When $a = 1/2$ we obtain $\hat{\theta}_{\text{MA}}$ as the median estimate, while $a = \alpha/2$ and $a = 1 - \alpha/2$ give the limits of the $(1 - \alpha) \times 100\%$ CI corresponding to the usual fixed-effect meta-analytic Wald CI.

The standard error of the meta-analytic median estimate is given by σ_{MA} from (10), and it has two desirable properties: First, it it is never larger than either of trials' standard errors ($\sigma_{\text{MA}} \leq \min\{\sigma_1, \sigma_2\}$). Second, under effect homogeneity, the standard error is the smallest among all unbiased estimators of the common effect[17]. Both properties do not hold for the two-trials rule and the other p-value combination methods discussed below.

2.2.2 | Asymptotics

Since the meta-analytic combined estimation function is a linear combination of normally distributed effect estimates, its distribution is also normal and given by

$$\hat{\mu}_{\text{MA}}(a) \sim \text{N}\left(\frac{\theta_1}{1+c} + \frac{\theta_2}{1+1/c} - \frac{z_a}{\sqrt{1/\sigma_1^2 + 1/\sigma_2^2}}, \frac{1}{1/\sigma_1^2 + 1/\sigma_2^2}\right)$$

for the alternative "greater" and with variance ratio $c = \sigma_1^2/\sigma_2^2$. For the alternative "less", the minus in the mean has to be replaced with a plus. The median estimate ($a = 1/2$) hence targets the weighted average of the true effects

$$\frac{\theta_1}{1+c} + \frac{\theta_2}{1+1/c}$$

while the meta-analytic CIs becomes increasingly concentrated around the weighted average with decreasing standard errors, provided the relative variance c stays constant.

Meta-analysis thus shows a less conservative asymptotic behavior than the two-trials rule in the sense that a more extreme effect can compensate for a less extreme one, whereas the two-trials rule would converge to the less extreme of the two effects. Figure 2 illustrates this asymptotic behavior: In case both estimates are identical and the standard errors very small (panel C), the meta-analytic CI is concentrated around the trials estimate, while in case of different estimates the CI concentrates somewhere in between (panel D). Since in this example the relative variance is $c = 2$, the weighted average is slightly closer to the estimate from trial 2.

3 | OTHER *P*-VALUE COMBINATION METHODS

While the two-trials rule and meta-analysis are the most commonly used p-value combination methods in practice, several other combination methods exist[17]. In this section, we examine Tippett's, Fisher's, Pearson's, and Edgington's methods, which can also be used to obtain combined effect estimates, CIs, and p-values. Although these methods are not standard in drug regulation, they may have useful properties in certain settings, as we will demonstrate in the following.

3.1 | Tippett's (minimum) method

The combined p-value from Tippett's method[35] is based on the minimum of the two p-values and given by

$$p_{\text{T}}(\mu) = 1 - (1 - \min\{p_1(\mu), p_2(\mu)\})^2.$$

It is closely related to the two-trials rule in the sense that the combined p-value based on the alternative "greater" from Tippett's method is the same as one minus the combined p-value based on the alternative "less" from the two-trials rule, and vice versa[28]. Similarly, Tippett's method is a special case of Wilkinson's method based on the $r = 1$ smallest out of $k = 2$ p-values.

3.1.1 | Effect estimation

Following a similar approach as with the two-trials rule, CIs and point estimates based on Tippett's method can be obtained in closed-form with the combined estimation function

$$\hat{\mu}_{\mathrm{T}}(a) = \begin{cases} \max\{\hat{\theta}_1 - \sigma_1 z_{\sqrt{1-a}}, \hat{\theta}_2 - \sigma_2 z_{\sqrt{1-a}}\} & \text{for alternative = "greater"} \\ \min\{\hat{\theta}_1 + \sigma_1 z_{\sqrt{1-a}}, \hat{\theta}_2 + \sigma_2 z_{\sqrt{1-a}}\} & \text{for alternative = "less".} \end{cases} \tag{11}$$

The similarity to the two-trials rule is again visible as (11) looks similar to the estimation function from the two-trials rule (3) with the minimum and maximum flipped and using different normal quantiles. The same median estimates are obtained (i.e., $\hat{\mu}_{\mathrm{2TR}}(1/2) = \hat{\mu}_{\mathrm{T}}(1/2)$) if opposite alternatives are specified.

We can see that when the observed effect estimates are the same ($\hat{\theta}_1 = \hat{\theta}_2 = \hat{\theta}$), the median estimate ($a = 1/2$) based on Tippett's method is not equal to $\hat{\theta}$ but shifted from it, as the two-trials rule (see panels A and C in Figure 2 for an illustration). Similarly, CIs obtained from Tippett's method are typically skewed in the sense that the distances between the point estimate and the upper and lower limits are not the same.

3.1.2 | Asymptotics

It can be shown that as the standard errors σ_1 and σ_2 decrease, the combined estimation function (11) converges to

$$\operatorname*{plim}_{\sigma_1,\sigma_2 \downarrow 0} \hat{\mu}_{\mathrm{T}}(a) = \begin{cases} \max\{\theta_1, \theta_2\} & \text{for alternative = "greater"} \\ \min\{\theta_1, \theta_2\} & \text{for alternative = "less",} \end{cases}$$

that is, the more extreme of the two effects, see Appendix B.3 for details. In contrast to the two-trials rule, Tippett's method is hence anti-conservative. This is illustrated in panel D of Figure 2 where Tippett's CI is tightly concentrated around the larger effect estimate.

3.2 | Fisher's and Pearson's (product) methods

Pearson's and Fisher's combination method are two closely related *p*-value combination methods, that are based on the product of *p*-values, or equivalently, the sum of the log *p*-values. Fisher's method has been proposed for combining *p*-values from clinical trials[9,36,30], however, using the associated *p*-value function for effect estimation in a regulatory trials setting has remained unexplored.

The combined *p*-value function based on Fisher's method[26] is given by

$$p_{\mathrm{F}}(\mu) = 1 - \Pr\left(\chi^2_4 \leq -2[\log\{p_1(\mu)\} + \log\{p_2(\mu)\}]\right) \tag{12}$$

while the combined *p*-value function based on Pearson's method[37] is given by

$$p_{\mathrm{P}}(\mu) = \Pr\left(\chi^2_4 \leq -2[\log\{1 - p_1(\mu)\} + \log\{1 - p_2(\mu)\}]\right). \tag{13}$$

Pearson[38] proposed also another method based on the maximum of the test statistics underlying the *p*-values (12) and (13), but we will not consider this method here as its test statistic does not have an exact null distribution[39]. As with the two-trials rule and Tippett's method, the combined *p*-value functions of Fisher's and Pearson's methods are related in the sense that the *p*-value function based on Fisher's method and the alternative hypothesis "greater" is the same as one minus the *p*-value function based on Pearson's method and the alternative "less", and vice versa[28]. As we will show in the following, Fisher's method also acts in a similar anti-conservative way as Tippett's method, while Pearson's method acts in a similar conservative way as the two-trials rule.

3.2.1 | Effect estimation

CIs and point estimates based on Fisher's and Pearson's methods can in general not be obtained in closed-form but require numerical root-finding. However, a special case where a closed-form solution is available is when the effect estimates and standard errors are the same in both trials ($\hat{\theta}_1 = \hat{\theta}_2 = \hat{\theta}$ and $\sigma_1 = \sigma_2 = \sigma$). While this is unrealistic in practice, it serves as an important soundness check to investigate whether the methods produce reasonable estimates in the situation of identical trial results. In this case, we obtain the following closed-form combined estimation function for Fisher's method

$$\hat{\mu}_{\text{F}}(a) = \begin{cases} \hat{\theta} + \sigma\, z_{\exp\{-\chi^2_4(1-a)/4\}} & \text{for alternative = "greater"} \\ \hat{\theta} - \sigma\, z_{\exp\{-\chi^2_4(1-a)/4\}} & \text{for alternative = "less"} \end{cases} \tag{14}$$

and for Pearson's method

$$\hat{\mu}_{\text{P}}(a) = \begin{cases} \hat{\theta} + \sigma\, z_{\exp\{-\chi^2_4(a)/4\}} & \text{for alternative = "greater"} \\ \hat{\theta} - \sigma\, z_{\exp\{-\chi^2_4(a)/4\}} & \text{for alternative = "less"} \end{cases} \tag{15}$$

with $\chi^2_4(a)$ the $a \times 100\%$ quantile of the chi-squared distribution with four degrees of freedom. Importantly, the median estimates ($a = 1/2$) from both methods do not equal the observed estimate $\hat{\theta}$ but are shifted away from it by $z_{\exp\{-\chi^2_4(1/2)/2\}} \approx -0.17$ standard errors σ, similar to the two-trials rule and Tippett's method. Another similarity is that the CI is skewed since the distance between the lower and upper limits to the point estimate is not the same.

3.2.2 | Asymptotics

To understand the asymptotic behavior of Fisher's and Pearson's method, we may again examine their combined estimation functions for decreasing standard errors. When the true effects are equal ($\theta_1 = \theta_2 = \theta$), both Fisher's and Pearson's median estimates will converge toward it, which is clear from the theory of p-value functions but can also be informally seen from (14) and (15) shrinking toward the common effect estimate for a decreasing standard error. On the other hand, when the true effects are unequal, it can then be shown that

$$\operatorname*{plim}_{\sigma_1,\sigma_2 \downarrow 0} \hat{\mu}_{\text{F}}(a) = \begin{cases} \max\{\theta_1, \theta_2\} & \text{for alternative = "greater"} \\ \min\{\theta_1, \theta_2\} & \text{for alternative = "less",} \end{cases}$$

and

$$\operatorname*{plim}_{\sigma_1,\sigma_2 \downarrow 0} \hat{\mu}_{\text{P}}(a) = \begin{cases} \min\{\theta_1, \theta_2\} & \text{for alternative = "greater"} \\ \max\{\theta_1, \theta_2\} & \text{for alternative = "less",} \end{cases}$$

see Appendix B.4 for details. This means that the combined estimation functions converge toward the more extreme effect for Fisher's method (e.g., the maximum of two positively oriented effects), and the less extreme effect for Pearson's method (e.g., the minimum of two positively oriented effects). The behavior is similar to Tippett's method and the two-trials rule where one method acts anti-conservative (Fisher and Tippett's methods), while the other methods acts conservative (Pearson's method and the two-trials rule). However, the examples in panels B and D of Figure 2 suggest that in finite samples, Fisher's and Pearson's method remain closer to the weighted average compared to Tippett's method and the two-trials rule.

3.3 | Edgington's (sum) method

Edgington's method based on the sum of p-values[25] is yet another p-value combination method that can be used for obtaining a combined p-value function, and the last method that we will consider in this paper. It is given by

$$p_{\text{E}}(\mu) = \begin{cases} E^2/2 & \text{if } 0 \leq E \leq 1 \\ 1 - (2-E)^2/2 & \text{if } 1 < E \leq 2 \end{cases} \tag{16}$$

with $E = p_1(\mu) + p_2(\mu)$. An attractive feature is that two-sided CIs based on Edgington's method are orientation invariant, which is not the case for the other combined p-value functions considered so far. That is, CIs based on Edgington's method do not depend

on the orientation of the underlying one-sided p-values, so the same CI is obtained regardless whether one uses p-values with the alternative "greater" or "less"[28]. Edgington's method has previously been used in meta-analysis[28], to synthesize p-values from original and replication studies[40], and suggested as an alternative for the two-trials rule[22]. However, its estimation properties in the context of two trials remain unexplored.

3.3.1 | Effect estimation

The median estimate based on Edgington's method has an intuitive interpretation as the null value μ for which the sum of the p-values is one. It can be obtained in closed-form by

$$\hat{\mu}_{\mathrm{E}}(1/2) = \frac{\hat{\theta}_1/\sigma_1 + \hat{\theta}_2/\sigma_2}{1/\sigma_1 + 1/\sigma_2} \tag{17}$$

so is a weighted average of the two effect estimates, as the meta-analytic point estimate (9). However, the weights from Edgington's method are equal to the inverse standard errors, whereas the weights from meta-analysis are equal to the inverse squared standard errors. Thus, Edgington's method gives more weight to smaller studies (those with larger standard errors) compared to meta-analysis. Moreover, since the expectation of the median estimate (17) is again a weighted average of the true effects, it follows that Edgington's median estimate is unbiased when the true effects coincide ($\theta_1 = \theta_2 = \theta$), while in case they differ, the median estimate targets a weighted average of the true effects, though not the same weighted average as targeted by meta-analysis.

The standard error of Edgington's median estimate is given by

$$\sigma_{\mathrm{E}} = \frac{\sqrt{2}}{1/\sigma_1 + 1/\sigma_2} \tag{18}$$

and is always larger than the meta-analytic standard error (10), see Appendix B.2. Therefore, under effect homogeneity, Edgington's method is less efficient thanmet a-analysis at estimating the common effect. Under effect heterogeneity, however, the two methods target different estimands, so a comparison of their standard errors is not meaningful. Finally, Edgington's standard error is not always equal or smaller than either of the two trials' standard errors. This is only the case if the standard error ratio is $\sqrt{2} - 1 \leq \sigma_2/\sigma_1 \leq \sqrt{2} + 1$. For example, suppose $\sigma_1 = 0.5$ and $\sigma_2 = 2$, then Edgington's standard error is $\sqrt{2}/(2 + 0.5) = 0.566$, which is greater than σ_1

In general, CIs for Edgington's method do not have closed-form solutions and must be computed numerically. Nevertheless, as with Pearson's and Fisher's methods, a closed-form combined estimation function is available when the effect estimates and standard errors from both trials coincide ($\hat{\theta} = \hat{\theta}_1 = \hat{\theta}_2$ and $\sigma = \sigma_1 = \sigma_2$), which enables again analytical assessment of how the CI behaves in this important scenario. In this case, the combined estimation function is

$$\hat{\mu}_{\mathrm{E}}(a) = \begin{cases} \hat{\theta} + \sigma\, z_{\sqrt{a/2}} & \text{for } a \leq 1/2 \\ \hat{\theta} - \sigma\, z_{\sqrt{(1-a)/2}} & \text{for } a > 1/2 \end{cases} \tag{19}$$

for the alternative "greater" and with the plus (minus) after $\hat{\theta}$ replaced with minus (plus) in (19) for the alternative "less". We can see that CIs obtained from (19) are symmetric and centered around the observed effect estimate $\hat{\theta}$, similar to meta-analysis but unlike the CIs from the two-trials rule, Fisher's, and Pearson's methods. Yet, Edgington's CI is in this case narrower than the meta-analytic CI. For example, Edgington's 95% CI is 12.2% narrower than the corresponding meta-analytic 95% CI. Panel A of Figure 2 illustrates this as Edgington's CI is narrower than the meta-analytic CI, although both are centered around the same effect estimate. However, in case the trials' effect estimates are different, Edgington's CI can also be much wider. For instance, in panel B of Figure 2 where the trials produced very different results, Edgington's CI is much wider than any of the other methods. This suggests that Edgington's method reacts to heterogeneity by widening its CI to include both trial effect estimates.

3.3.2 | Asymptotics

Because the median estimate based on Edgington's method is a weighted average of two normally distributed effect estimates, it is also normally distributed

$$\hat{\mu}_{\mathrm{E}}(1/2) \sim \mathrm{N}\left(\frac{\theta_1}{1+\sqrt{c}} + \frac{\theta_2}{1+1/\sqrt{c}}, \frac{2}{(1/\sigma_1 + 1/\sigma_2)^2}\right)$$

with relative variance $c = \sigma_1^2/\sigma_2^2$. As the median estimate has its mean at the weighted average

$$\frac{\theta_1}{1+\sqrt{c}} + \frac{\theta_2}{1+1/\sqrt{c}}$$

it is clear that it will converge toward it as the standard errors decrease. Whether the CI shrinks to this weighted average depends on whether the true effects are equal. In case they are, it can be informally seen that the CI (19) will shrink to the common true effect, which is illustrated in panel C of Figure 2. However, when the true effects differ, the CI will not shrink to a point but remain an interval that always includes both true effects as the limiting combined estimation function is

$$\operatorname*{plim}_{\sigma_1,\sigma_2 \downarrow 0} \hat{\mu}_{\mathrm{E}}(a) = \begin{cases} \min\{\theta_1, \theta_2\} & \text{for } a < 1/2 \\ \dfrac{\theta_1}{1+\sqrt{c}} + \dfrac{\theta_2}{1+1/\sqrt{c}} & \text{for } a = 1/2 \\ \max\{\theta_1, \theta_2\} & \text{for } a > 1/2 \end{cases} \tag{20}$$

see Appendix B.4 for details. This means that CIs based on Edgington's method will asymptotically always include both true effects, even when the trials' sample sizes become arbitrarily large, see panel D of Figure 2 for an illustration. This behavior is strikingly different from meta-analysis whose CI shrinks to a point at the weighted average, even when the true effects are not the same.

4 | APPLICATIONS

We will now illustrate combined p-value functions, CIs, and median estimates on data from two different pairs of clinical trials.

4.1 | The RESPIRE trials

We first revisit the RESPIRE trials[6,7,8], which were presented as motivating example in Table 1 in the introduction. The trials investigated the effect of ciprofloxacin in the treatment of non-cystic fibrosis bronchiectasis. Each trial had two treatment groups (on/off treatment cycles of either 14 or 28 days for 48 weeks) and two corresponding control groups. RESPIRE 1 showed a substantial treatment effect in the 14-day treatment regimen (estimated log rate ratio of $\log \widehat{\mathrm{RR}} = -0.49$ with 95% CI from –0.85 to –0.13), while the benefit was less pronounced in RESPIRE 2 ($\log \widehat{\mathrm{RR}} = -0.18$ with 95% CI from –0.53 to 0.16). Surprisingly, this was reversed for the 28-day regimens, with RESPIRE 2 showing a much stronger treatment effect ($\log \widehat{\mathrm{RR}} = -0.6$ with 95% CI from –0.96 to –0.23) while RESPIRE 1 showed almost no benefit ($\log \widehat{\mathrm{RR}} = -0.02$ with 95% CI from –0.39 to 0.35). Figure 3 shows the p-value functions of the two studies (dashed lines) along with different combined p-value functions (solid lines) and corresponding point estimates and CIs (top). Table 3 shows the results in numerical form. Of note, all results were computed with our R package `twotrials` and Appendix A shows how the results for the 14-day treatment group can be reproduced.

Looking at the combined point estimates, we can see that for both the 14-day and 28-day regimens, the estimate based on Tippett's method is the smallest (i.e., most anti-conservative for alternative = "less"), while the estimate based on the two-trials rule is the largest (i.e., the most conservative). A similar but slightly attenuated pattern is seen for Fisher's (anti-conservative) and Pearson's (conservative) methods, whereas the estimates from meta-analysis and Edgington's method are almost identical and fall somewhere between the individual trials' effect estimates. All point estimates are thus consistent with the theoretically expected behavior of the methods.

It is interesting to consider the median estimate as a weighted average of the trial specific point estimates, and to determine the corresponding (implicit) weights. Table 3 reports the weight of the point estimate from RESPIRE 1 toward the median estimate

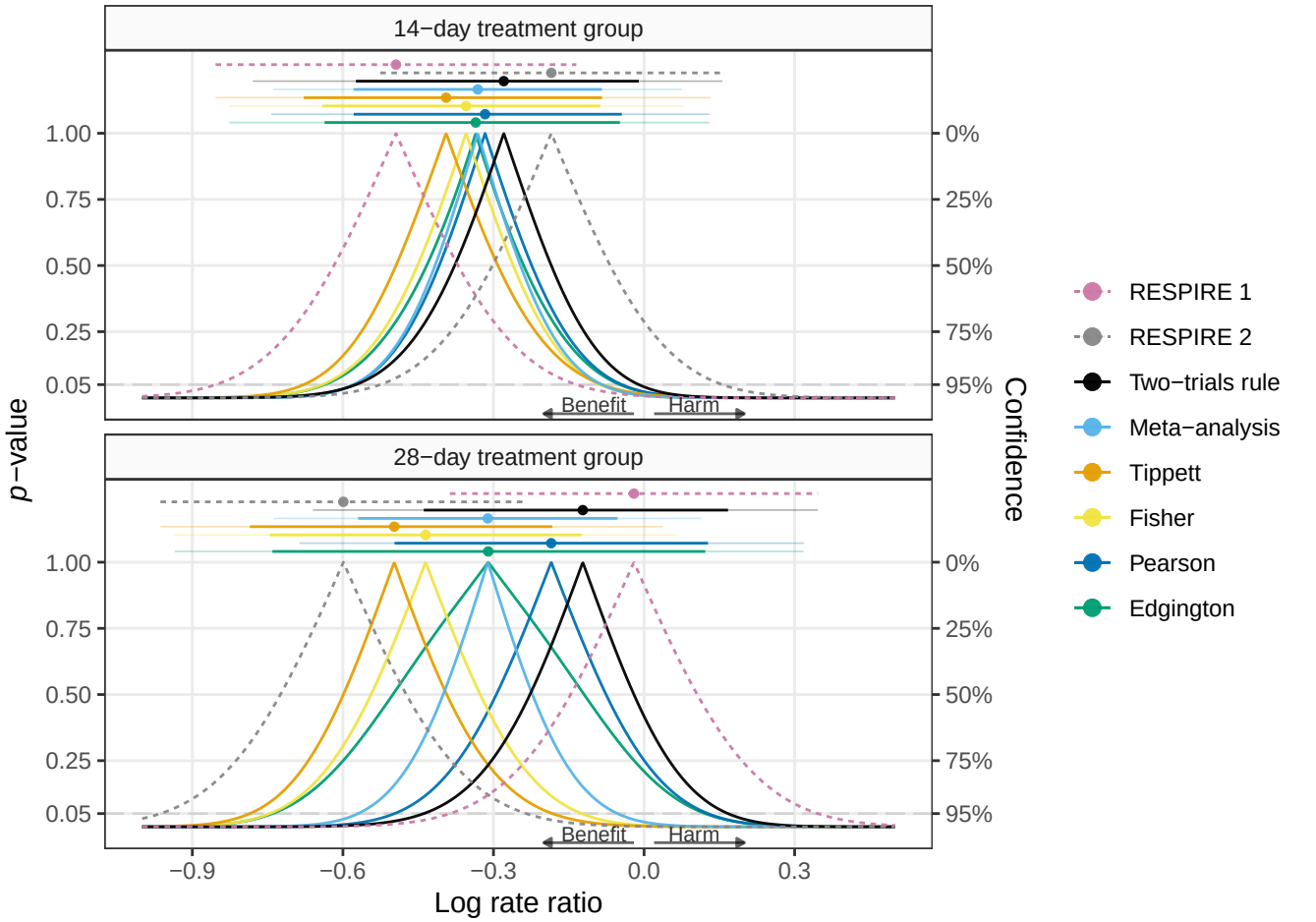


FIGURE 3 Results of the RESPIRE trials[6,7,8] for the effect of ciprofloxacin over 14 days (top) or 28 days (bottom) compared to placebo for the treatment of non-cystic fibrosis bronchiectasis. The two-sided p-value functions of the individual trials (dashed lines), and the combined p-value functions (solid lines) based on the two-trials rule, fixed-effect meta-analysis, Tippett's, Fisher's, Pearson's, and Edgington's methods are shown along with corresponding median estimates and CIs (95% and 99.875% via telescope lines). All p-values are based on the alternative "greater" and then converted to two-sided p-values via the centrality function $2\min\{p, 1-p\}$.

(the weight from RESPIRE 2 is one minus the weight from RESPIRE 1).It is visible that the more extreme estimate (RESPIRE 1 in the 14-day group, and RESPIRE 2 in the 28-day group) contributes more to Tippett's and Fisher's estimates and less to Pearson's and the two-trials rule estimates, which aligns with the expected behavior. Similarly, the weight of RESPIRE 1 is slightly larger for Edgington's method than meta-analysis because Edgington's estimate gives more weights to trials with larger standard errors due to its inverse standard error weighting.

Looking at the CIs, we can see that meta-analysis produces narrower CIs than the other methods for both treatment regimens. The widest CIs are produced by Edgington's method. For the 28-day regimen, Edgington's CI is the only method that includes both trial effect estimates, and as a result is even wider than the CIs from the individual trials, reflecting the apparent heterogeneity. Looking at the decision based on the CIs, we can see that for the 14-day regimens, all 95% CIs exclude a log rate ratio of zero, the value of no effect, while all 99.875% CIs include it. However, for the 28-day regimens, the 95% CIs from meta-analysis, Fisher's, and Tippett's methods exclude zero. The other method's 95% CIs include zero, but only Edgington's method includes also the point estimate from RESPIRE 2. Finally, the 99.875% CIs of all methods include zero, thus leading to identical decisions at the one-sided $0.025^2 = 0.000625$ level. Note that for each method, the decision based on the CI is compatible with the combined

TABLE 3 Point estimates (with implicit weights), 95% CIs (with widths), and *p*-values for the RESPIRE trials[6,7,8].

	Log rate ratio	Weight RESPIRE 1	95% CI	CI width	*P*-value (one-sided)
14-day treatment group					
RESPIRE 1	–0.49		–0.85 to –0.13	0.72	0.00351
RESPIRE 2	–0.18		–0.53 to 0.16	0.68	0.14400
Two-trials rule	–0.28	0.31	–0.57 to –0.01	0.56	0.02073
Meta-analysis	–0.33	0.47	–0.58 to –0.08	0.49	0.00432
Tippett	–0.39	0.68	–0.68 to –0.08	0.59	0.00701
Fisher	–0.35	0.55	–0.64 to –0.09	0.55	0.00434
Pearson	–0.32	0.43	–0.58 to –0.04	0.53	0.01138
Edgington	–0.34	0.49	–0.64 to –0.05	0.59	0.01088
28-day treatment group					
RESPIRE 1	–0.02		–0.39 to 0.35	0.73	0.45699
RESPIRE 2	–0.60		–0.96 to –0.23	0.73	0.00064
Two-trials rule	–0.12	0.82	–0.44 to 0.17	0.61	0.20884
Meta-analysis	–0.31	0.50	–0.57 to –0.05	0.52	0.00912
Tippett	–0.50	0.18	–0.79 to –0.18	0.60	0.00127
Fisher	–0.44	0.28	–0.75 to –0.12	0.62	0.00266
Pearson	–0.18	0.72	–0.50 to 0.13	0.62	0.12562
Edgington	–0.31	0.50	–0.74 to 0.12	0.86	0.10471

p-values in Table 3, for example, a 99.875% CI excludes a log rate ratio of zero only if also the combined one-sided *p*-value is less than 0.000625.

4.2 | The ORBIT trials

Another pair of clinical trials that investigated the effect of ciprofloxacin are the ORBIT 3 and ORBIT 4 trials[41]. The trials assessed the effect of inhaled liposomal ciprofloxacin compared to placebo in patients with non-cystic fibrosis bronchiectasis and chronic lung infection with *Pseudomonas aeruginosa*. Like the RESPIRE trials, the ORBIT trials also showed considerable heterogeneity. Figure 4 shows *p*-values, point estimates, and CIs for the primary endpoint (time to the first exacerbation; effect quantified with a log hazard ratio) and a secondary endpoint (frequency of exacerbations; effect quantified with a log rate ratio). Table 4 gives numerical summaries.

We see that there is substantial heterogeneity for the primary endpoint, with the point estimate from ORBIT 3 close to zero ($\log \widehat{\text{HR}} = -0.01$ with 95% CI from –0.34 to 0.32), whereas the estimate from ORBIT 4 indicates a more beneficial treatment effect ($\log \widehat{\text{HR}} = -0.33$ with 95% CI from –0.63 to –0.03). While the theoretically expected patterns of the different median estimates and CIs are visible, the qualitative decisions based on all the different combination methods are the same at both the 0.025 and 0.025^2 levels.

Looking at the secondary endpoint, there is also considerable heterogeneity between the results from ORBIT 3 ($\log \widehat{\text{RR}} = -0.16$ with 95% CI from –0.43 to 0.11) and ORBIT 4 ($\log \widehat{\text{RR}} = -0.46$ with 95% CI from –0.73 to –0.2) leading to some more noticeable qualitative differences between the methods. That is, the 99.875% CIs from meta-analysis and Fisher's method exclude a log rate ratio of zero while the remaining methods include it, leading to different decisions at the 0.025^2 level. Again, Edgington's CI is much wider than the others due to the substantial heterogeneity.

In summary, the analyses of the RESPIRE and ORBIT trials showed how combined *p*-value functions allow us to obtain point estimates, CIs, and *p*-values that are inherently compatible. They also showed that different combination methods can lead to different inferences and decisions, especially in the presence of between-trial heterogeneity, highlighting the need to think about the estimand of interest.

5 | EXTENSION TO MORE THAN TWO TRIALS

The methods discussed so far have focused on the setting where only two trials are available, but in practice it may happen that investigators want to assess the combined evidence from more than two trials. In this context, Rosenkranz[32] suggested that

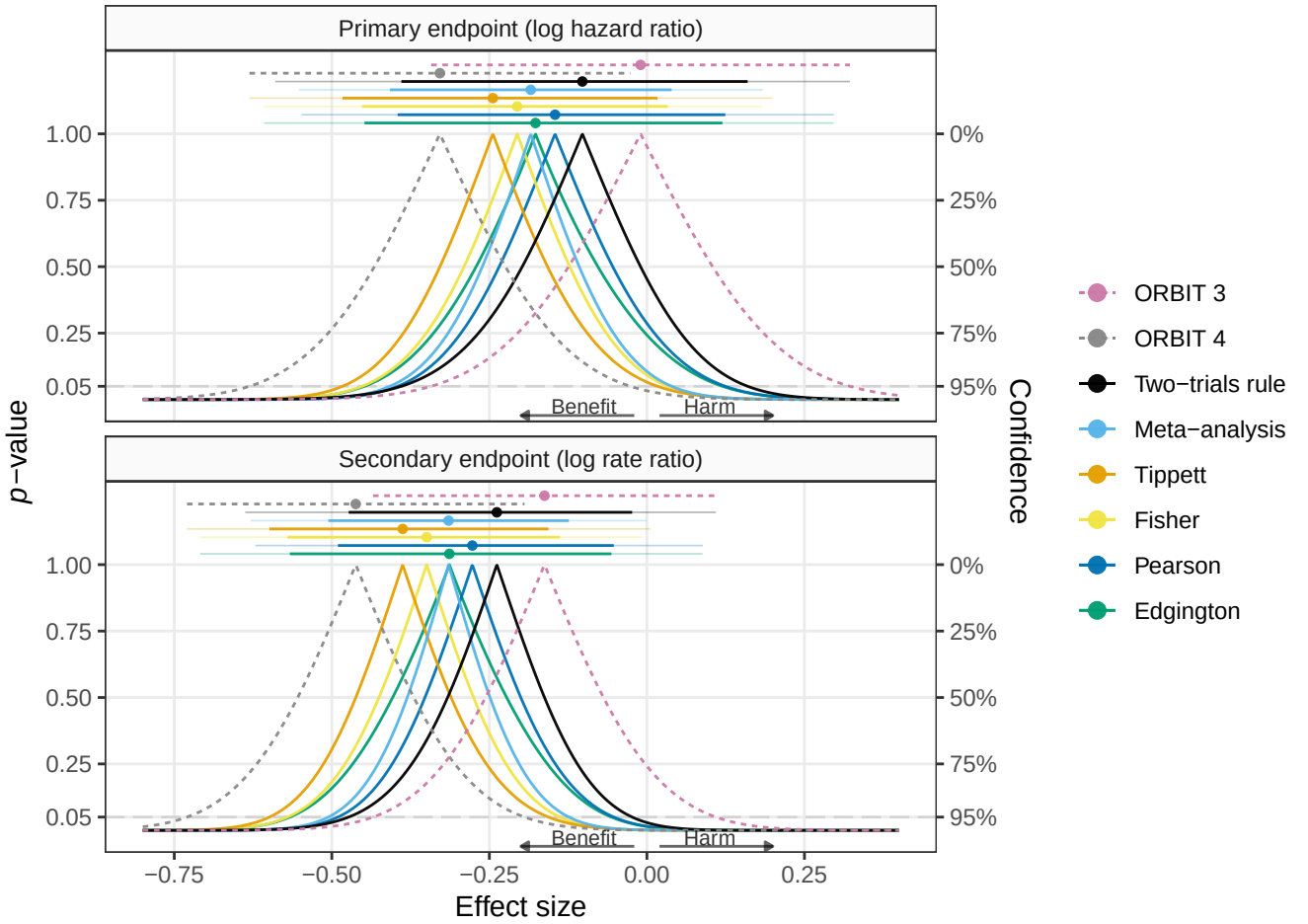


FIGURE 4 Results of the ORBIT trials[41] for the effect of ciprofloxacin in patients with non-cystic fibrosis bronchiectasis and chronic lung infection with *Pseudomonas aeruginosa*. The two-sided *p*-value functions of the individual trials (dashed lines), and the combined *p*-value functions (solid lines) based on the two-trials rule, fixed-effect meta-analysis, Tippett's, Fisher's, Pearson's, and Edgington's methods are shown along with corresponding median estimates and CIs (95% and 99.875% via telescope lines). All *p*-values are based on the alternative "greater" and then converted to two-sided *p*-values via the centrality function $2\min\{p, 1-p\}$.

decision rules should maintain the type I error rate of the two-trials rule for two studies α^2, even if there are more than two studies. This can be implemented using combined *p*-value functions, as all methods considered before can be generalized to more than two trials[13,28]. A decision rule can then be based on the combined one-sided *p*-value for the null hypothesis of no effect or a $(1-2\alpha^2)\times 100\%$ CI obtained from a combined *p*-value function. In addition, a point estimate and 95% CI can be used to summarize the combined evidence.

While all point estimates and confidence intervals in this setting can be computed numerically, some of the analytical results derived earlier generalize to more than two studies. Specifically, closed-form median estimates and confidence intervals remain available for the two-trials rule, Tippett's method, and meta-analysis, whereas such closed-form solutions are not available for Fisher's, Pearson's, and Edgington's methods[28]. In particular, for Edgington's method, one might expect the median estimate (17) to generalize by incorporating additional effect estimates with inverse standard error weights. However, a comparison with numerically computed median estimates showed that this is not the case. Thus, the inverse standard error weighted average in (17) corresponds to Edgington's median estimate only in the setting of two trials.

Figure 5 shows *p*-value functions that combine all four results from the two RESPIRE trials, as also done by Chotirmall and Chalmers[8] with fixed-effect meta-analysis. Looking at the median estimates, we see the same patterns as when the methods

TABLE 4 Point estimates (with implicit weights), 95% CIs (with widths), and *p*-values for the ORBIT trials[41].

	Log rate ratio	Weight ORBIT 3	95% CI	CI width	*P*-value (one-sided)
Primary endpoint (log hazard ratio)					
ORBIT 3	–0.01		–0.34 to 0.32	0.66	0.47636
ORBIT 4	–0.33		–0.63 to –0.03	0.60	0.01657
Two-trials rule	–0.10	0.71	–0.39 to 0.16	0.55	0.22692
Meta-analysis	–0.18	0.45	–0.41 to 0.04	0.45	0.05305
Tippett	–0.24	0.26	–0.48 to 0.02	0.50	0.03286
Fisher	–0.21	0.39	–0.45 to 0.03	0.49	0.04610
Pearson	–0.15	0.57	–0.40 to 0.12	0.52	0.14328
Edgington	–0.18	0.48	–0.45 to 0.12	0.57	0.12149
Secondary endpoint (log rate ratio)					
ORBIT 3	–0.16		–0.43 to 0.11	0.54	0.12083
ORBIT 4	–0.46		–0.73 to –0.19	0.54	0.00036
Two-trials rule	–0.24	0.75	–0.47 to –0.02	0.45	0.01460
Meta-analysis	–0.31	0.49	–0.51 to –0.12	0.38	0.00062
Tippett	–0.39	0.25	–0.60 to –0.16	0.44	0.00072
Fisher	–0.35	0.38	–0.57 to –0.14	0.43	0.00048
Pearson	–0.28	0.62	–0.49 to –0.05	0.44	0.00765
Edgington	–0.31	0.50	–0.57 to –0.06	0.51	0.00734

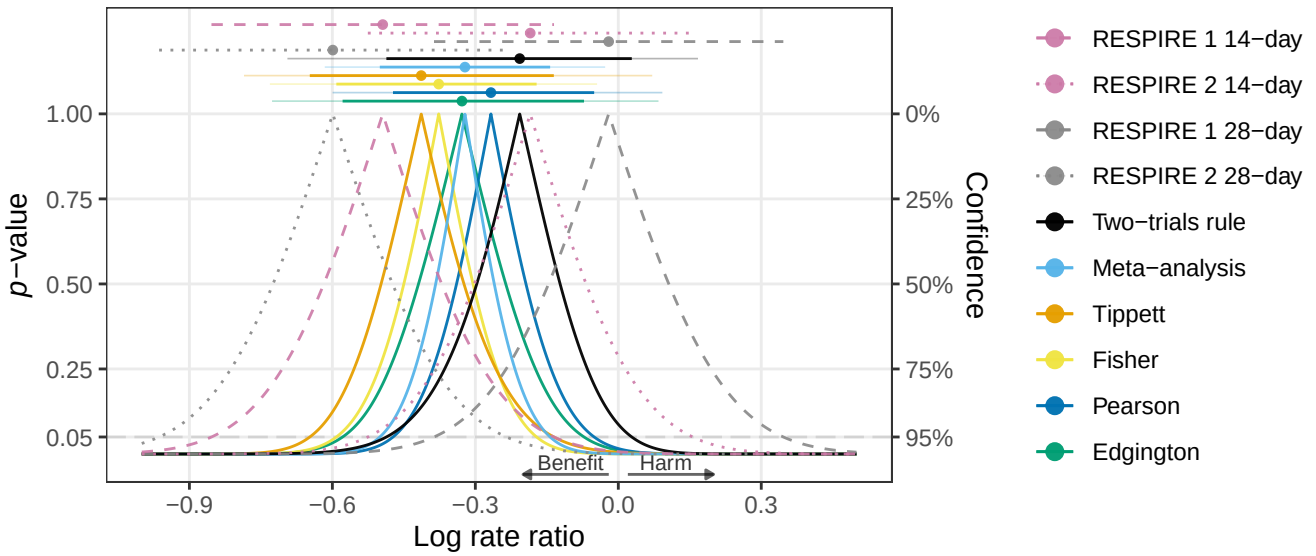


FIGURE 5 Combined *p*-value function based on combining all four estimates from the RESPIRE trials[6,7,8] along with corresponding median estimates and CIs (95% and 99.875% via telescope lines).

are applied to only two trials; The median estimates from meta-analysis and Edgington's method are somewhere in between the trials' individual estimates. The median estimates based on Fisher's and Tippett's method are the most anti-conservative, and the median estimates based on Pearson's method and the two-trials rule are the most conservative. A decision rule that maintains the $\alpha^2 = 0.025^2$ type I error rate of the two trials rule could now be defined by flagging drug efficacy when the $(1-2\alpha^2) \times 100\% = 99.875\%$ excludes the null value. Following this rule, we can see that Fisher's method and meta-analysis flag efficacy while the remaining methods do not. In sum, this example illustrates how combined *p*-value functions can be applied to more than two trials, while allowing to maintain the same type I error rate as for two trials.

6 | DISCUSSION

The two-trials rule has been widely discussed in the literature but discussions have mostly focused on hypothesis testing characteristics, such as power or type I error rate[9,3,4,36,5,42,30,43,44,10,31,32,22]. In this paper, we took a different perspective, systematically examining the two-trials rule and alternative methods in terms of effect estimation. By casting them in a combined *p*-value function framework, we derived compatible *p*-values, confidence intervals, and point estimates. These quantities are compatible in the sense that the two-sided *p*-value for a null value is less than α if and only if the null value is excluded by the $(1-\alpha)\times 100\%$ CI, and that the point estimate is contained in the CI at any confidence level. While meta-analytic effect estimates, CIs, and *p*-values have been well studied, our novel results enable computation of CIs and effect estimates based on the two-trials rule. Investigators could therefore report not only individual trial *p*-values (essentially the two-trials rule) and meta-analytic estimates but also point estimates and CIs based on the two-trials rule.

Our findings also clarify how different *p*-value combination methods implicitly target different estimands. Reassuringly, under effect homogeneity (i.e., the same true effect in both trials), all methods yield consistent point estimates and CIs that shrink toward the true effect as standard errors decrease, although some show bias. Theoretically, meta-analysis has the smallest variance among all unbiased estimators (attaining the Cramér-Rao lower bound) and may be preferred. However, under effect heterogeneity – arguably the more realistic scenario – it is less clear which method should be recommended. The two-trials rule and Pearson's method are conservative (targeting the less extreme effect), Fisher's and Tippett's methods are anti-conservative (targeting the more extreme effect), while Edgington's method and meta-analysis are balanced (targeting a weighted average). This raises an important question: What kind of effect is of scientific interest when the true trial effects differ? If the investigators are interested in the less extreme effect – arguably a sensible choice when the effects relate to a medical treatment with potential side effects – then the two-trials rule and Pearson's method seem reasonable. On the other hand, a weighted average effect, as targeted by meta-analysis and Edgington's method, might be a relevant estimand if it is representative for a larger population[34]. Finally, the more extreme effect might be the relevant estimand if the maximum achievable benefit of a treatment is of scientific interest, in which case Tippett's and Fisher's methods might be reasonable choices. This parallels the findings of Heard and Rubin-Delanchy[21], who showed that many *p*-value combination methods are equivalent to a likelihood ratio test for specific alternative hypotheses. This means that each such method can be most powerful under certain conditions. Therefore, researchers must carefully reflect which alternative hypothesis is most relevant to their application – just as they need to reflect on choosing an appropriate estimand – to select a suitable combination method.

Beyond theoretical considerations, practical issues must be addressed. A major concern is that if the effect estimates from both trials are the same, the two-trials rule, Tippett's, Fisher's, and Pearson's methods all produce counterintuitive effect estimates that differ from the one observed in both trials. Such point estimates are unintuitive and difficult to communicate to non-statisticians. Moreover, only Edgington's method and meta-analysis produce the same combined estimate and two-sided confidence interval in case the alternative of combined *p*-values is changed[28], which seems another practically desired property. From this perspective, Edgington's method and meta-analysis may be preferable. In particular, Edgington's method can also account for effect heterogeneity by widening its CI when there is heterogeneity and asymptotically always includes both effects. However, this is traded off with a less efficient median estimate under effect homogeneity, whose standard error can even be larger than those from both trials if they greatly differ. Finally, another practical challenge is aligning decisions based on a one-sided combined *p*-value thresholded at α^2 with two-sided CIs. This requires using a $(1-2\alpha^2)\times 100\%$ confidence level. However, in many fields, researchers are not used to such confidence levels, so we suggest to report both a more conventional level (e.g., 95%) along with $(1-2\alpha^2)\times 100\%$ via telescope-style CIs, as well as the underlying *p*-value function, as in Figures 3–5.

A broader issue is the question of whether two trials are actually necessary. If the designs of the two trials are so similar that they can be considered exchangeable ("direct replications"[45]), there are various arguments in favor of conducting one large trial instead of two smaller ones[9,10]. Also our study demonstrates that having two trials instead of one makes estimation more complicated. Conversly, if the trial designs differ significantly ("conceptual replications"[45], e.g., if they use different endpoints or populations), achieving success in both trials may provide more robust evidence of treatment efficacy. From this perspective, it is sensible to design the trials differently to some extent[10]. However, there is perhaps a limit to how different the trials can be, as when there is too much heterogeneity, combining the effect estimates into a single number would no longer be meaningful.

Our results have broader implications beyond the two-trial setting. Methods for combining *p*-values are also used in adaptive trials, where they enable combination of stage-wise *p*-values[46,47,48]. Combined *p*-value functions can be generalized to more than two studies, making them applicable to meta-analysis[13,28]. They can also be applied to replication and real-world evidence studies, where the two-trials rule (under different names such as significance criterion or vote-counting) is used to assess the replicability

of original findings[49,50,40]. In all these scenarios, we may consider combined p-value functions for parameter estimation, but in each application researchers must also decide which combination method has the statistical properties to estimate the scientific effect of interest. Future research may also examine other combination methods beyond the ones considered here, such as the inverse chi-square method[51,17], the harmonic mean χ^2 test[52], the Cauchy combination test[53], random-effects meta-analysis[33], and combining p-value functions that are based on the exact distribution of the data rather than normality e.g., the p-value function based on Fisher's exact test with mid-p correction as considered in [54,28]. Additionally, fixed-effect meta-analysis has a Bayesian interpretation, corresponding to posterior inferences assuming equal true study effects and a flat prior distribution. Investigating whether other p-value combination methods have similar Bayesian justifications could be an interesting avenue for future work. To sum up, combined p-value functions provide a unified approach for combining results from two trials that can be further developed theoretically. Moreover, our software implementation allows researchers to conveniently apply these methods in practice.

ACKNOWLEDGMENTS

We thank the editor, associate editor, Stephen Senn, and another anonymous reviewer for valuable comments that led to numerous additions and improvements. The acknowledgment of these individuals does not imply their endorsement of the paper.

CONFLICT OF INTEREST

We declare no conflict of interest.

SOFTWARE AND DATA

Data from the RESPIRE trials were extracted from Table 3 in De Soyza et al.[7] and Table 3 in Aksamit et al.[6]. Data from the ORBIT trials were extracted from page 219 in Haworth et al.[41]. Code and data to reproduce all numbers, tables, and figures are openly available at https://github.com/SamCH93/twotrials. Spreadsheets containing the numbers from Tables 3 and 4 with higher precision are also available at the repository. A snapshot of the repository at the time of writing is available at https://doi.org/10.5281/zenodo.15017483. We used the statistical programming language R version 4.4.1 (2024-06-14) for analyses[55] along with the `confMeta`[56], `ggplot2`[57], `kableExtra`[58], `dplyr`[59], and `knitr`[60] packages.

REFERENCES

1. FDA . Providing clinical evidence of effectiveness for human drug and biological products.. Website; 1998. www.fda.gov/regulatory-information/search-fda-guidance-documents/providing-clinical-evidence-effectiveness-human-drug-and-biological-products.
2. FDA . Demonstrating Substantial Evidence of Effectiveness for Human Drug and Biological Products. Website; 2019. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/demonstrating-substantial-evidence-effectiveness-human-drug-and-biological-products.
3. Fisher LD. Carvedilol and the Food and Drug Administration (FDA) Approval Process: The FDA Paradigm and Reflections on Hypothesis Testing. *Controlled Clinical Trials.* 1999;20(1):16 - 39. doi: 10.1016/S0197-2456(98)00054-3
4. Lu HL, Huque M. Understanding on the Pooled Test for Controlled Clinical Trials. *Biometrical Journal.* 2001;43(7):909–923. doi: 10.1002/1521-4036(200111)43:7<909::aid-bimj909>3.0.co;2-o
5. Maca J, Gallo P, Branson M, Maurer W. Reconsidering some aspects of the two-trials paradigm. *Journal of Biopharmaceutical Statistics.* 2002;12(2):107–119. doi: 10.1081/bip-120006450
6. Aksamit T, De Soyza A, Bandel TJ, et al. RESPIRE 2: a phase III placebo-controlled randomised trial of ciprofloxacin dry powder for inhalation in non-cystic fibrosis bronchiectasis. *European Respiratory Journal.* 2018;51(1):1702053. doi: 10.1183/13993003.02053-2017
7. De Soyza A, Aksamit T, Bandel TJ, et al. RESPIRE 1: a phase III placebo-controlled randomised trial of ciprofloxacin dry powder for inhalation in non-cystic fibrosis bronchiectasis. *European Respiratory Journal.* 2018;51(1):1702052. doi: 10.1183/13993003.02052-2017
8. Chotirmall SH, Chalmers JD. RESPIRE: breathing new life into bronchiectasis. *European Respiratory Journal.* 2018;51(1):1702444. doi: 10.1183/13993003.02444-2017
9. Fisher LD. One Large, Well-Designed, Multicenter Study as an Alternative to the Usual FDA Paradigm.. *Drug Information Journal.* 1999;33(1):265–271. doi: 10.1177/009286159903300130
10. Senn S. *Statistical Issues in Drug Development*. Wiley, 2021
11. Zhang AD, Puthumana J, Downing NS, Shah ND, Krumholz HM, Ross JS. Assessment of Clinical Trials Supporting US Food and Drug Administration Approval of Novel Therapeutic Agents, 1995-2017. *JAMA Network Open.* 2020;3(4):e203284. doi: 10.1001/jamanetworkopen.2020.3284
12. Bender R, Berg G, Zeeb H. Tutorial: Using Confidence Curves in Medical Research. *Biometrical Journal.* 2005;47(2):237–247. doi: 10.1002/bimj.200410104
13. Xie M, Singh K. Confidence Distribution, the Frequentist Distribution Estimator of a Parameter: A Review. *International Statistical Review.* 2013;81(1):3–39. doi: 10.1111/insr.12000
14. Fraser DAS. The p-value Function and Statistical Inference. *The American Statistician.* 2019;73(sup1):135–147. doi: 10.1080/00031305.2018.1556735
15. Infanger D, Schmidt-Trucksäss A. P value functions: An underused method to present research results and to promote quantitative reasoning. *Statistics in Medicine.* 2019;38(21):4189-4197. doi: 10.1002/sim.8293
16. Marschner IC. Confidence distributions for treatment effects in clinical trials: Posteriors without priors. *Statistics in Medicine.* 2024;43(6):1271–1289. doi: 10.1002/sim.10000

17. Hedges LV, Olkin I. *Statistical Methods for Meta-Analysis*. Elsevier, 1985
18. Singh K, Xie M, Strawderman WE. Combining information from independent sources through confidence distributions. *The Annals of Statistics.* 2005;33(1). doi: 10.1214/009053604000001084
19. Cousins RD. Annotated Bibliography of Some Papers on Combining Significances or *p*-values. 2007. doi: 10.48550/arXiv.0705.2209
20. Xie M, Singh K, Strawderman WE. Confidence Distributions and a Unifying Framework for Meta-Analysis. *Journal of the American Statistical Association.* 2011;106(493):320–333. doi: 10.1198/jasa.2011.tm09803
21. Heard NA, Rubin-Delanchy P. Choosing between methods of combining *p*-values. *Biometrika.* 2018;105(1):239–246. doi: 10.1093/biomet/asx076
22. Held L. Beyond the two-trials rule. *Statistics in Medicine.* 2024. doi: 10.1002/sim.10055
23. Wilkinson B. A Statistical Consideration in Psychological Research. *Psychological Bulletin.* 1951;48:156–158. doi: 10.1037/h0059111
24. Stouffer SA, Suchman EA, Devinney LC, Star SA, Williams J. *The American soldier: Adjustment during army life. (Studies in social psychology in World War II)*. Princeton Univ. Press: Cambridge University Press, 1949.
25. Edgington ES. An Additive Method for Combining Probability Values from Independent Experiments. *The Journal of Psychology.* 1972;80(2):351–363. doi: 10.1080/00223980.1972.9924813
26. Fisher RA. *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd. 4 ed., 1932.
27. Birnbaum A. Combining Independent Tests of Significance. *Journal of the American Statistical Association.* 1954;49(267):559. doi: 10.2307/2281130
28. Held L, Hofmann F, Pawel S. A comparison of combined *p*-value functions for meta-analysis. *Research Synthesis Methods.* 2025. doi: 10.1017/rsm.2025.26
29. Fraser DAS. *p*-Values: The Insight to Modern Statistical Inference. *Annual Review of Statistics and Its Application.* 2017;4(1):1–14. doi: 10.1146/annurev-statistics-060116-054139
30. Shun Z, Chi E, Durrleman S, Fisher L. Statistical consideration of the strategy for demonstrating clinical evidence of effectiveness—one larger vs two smaller pivotal studies. *Statistics in Medicine.* 2005;24(11):1619–1637. doi: 10.1002/sim.2015
31. Zhan SJ, Kunz CU, Stallard N. Should the two-trial paradigm still be the gold standard in drug assessment?. *Pharmaceutical Statistics.* 2022;22(1):96–111. doi: 10.1002/pst.2262
32. Rosenkranz GK. A Generalization of the Two Trials Paradigm. *Therapeutic Innovation & Regulatory Science.* 2023;57:316–320. doi: 10.1007/s43441-022-00471-4
33. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods.* 2010;1(2):97–111. doi: 10.1002/jrsm.12
34. Rice K, Higgins JPT, Lumley T. A re-evaluation of fixed effect(s) meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society).* 2018;181(1):205-227. doi: 10.1111/rssa.12275
35. Tippett LHC. *Methods of Statistics*. Williams Norgate, 1931.
36. Rosenkranz G. Is It Possible to Claim Efficacy if One of Two Trials is Significant While the other Just Shows a Trend?. *Drug Information Journal.* 2002;36(4):875–879. doi: 10.1177/009286150203600416
37. Pearson K. On a Method of Determining Whether a Sample of Size n Supposed to Have Been Drawn from a Parent Population Having a Known Probability Integral has Probably Been Drawn at Random. *Biometrika.* 1933;25:379-410. doi: 10.1093/biomet/25.3-4.379
38. Pearson K. On a New Method of Determining "Goodness of Fit". *Biometrika.* 1934;26(4):425–442. doi: 10.1093/biomet/26.4.425
39. Owen AB. Karl Pearson's meta-analysis revisited. *The Annals of Statistics.* 2009;37(6B):3867–3892. doi: 10.1214/09-aos697
40. Held L, Pawel S, Micheloud C. The assessment of replicability using the sum of *p*-values. *Royal Society Open Science.* 2024. doi: 10.1098/rsos.240149
41. Haworth CS, Bilton D, Chalmers JD, et al. Inhaled liposomal ciprofloxacin in patients with non-cystic fibrosis bronchiectasis and chronic lung infection with *Pseudomonas aeruginosa* (ORBIT-3 and ORBIT-4): two phase 3, randomised controlled trials. *The Lancet Respiratory Medicine.* 2019;7(3):213–226. doi: 10.1016/s2213-2600(18)30427-2
42. Shao J, Chow S. Reproducibility probability in clinical trials. *Statistics in Medicine.* 2002;21(12):1727–1742. doi: 10.1002/sim.1177
43. van Ravenzwaaij D, Ioannidis JPA. A simulation study of the strength of evidence in the recommendation of medications based on two trials with statistically significant results. *PLOS ONE.* 2017;12(3):e0173184. doi: 10.1371/journal.pone.0173184
44. Kennedy-Shaffer L. When the Alpha is the Omega: *P*-values, "Substantial Evidence," and the 0.05 Standard at FDA. *Food and Drug Law Journal.* 2017;72(4):595–635.
45. Nosek BA, Errington TM. Making sense of replications. *eLife.* 2017;6. doi: 10.7554/elife.23383
46. Bauer P, Kohne K. Evaluation of Experiments with Adaptive Interim Analyses. *Biometrics.* 1994;50(4):1029–1041. doi: 10.2307/2533441
47. Lehmacher W, Wassmer G. Adaptive Sample Size Calculations in Group Sequential Trials. *Biometrics.* 1999;55(4):1286–1290. doi: 10.1111/j.0006-341x.1999.01286.x
48. Bauer P, Kieser M. Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine.* 1999;18(14):1833–1848. doi: 10.1002/(sici)1097-0258(19990730)18:14<1833::aid-sim221>3.0.co;2-3
49. Bartlett VL, Dhruva SS, Shah ND, Ryan P, Ross JS. Feasibility of Using Real-World Data to Replicate Clinical Trial Evidence. *JAMA Network Open.* 2019;2(10):e1912869. doi: 10.1001/jamanetworkopen.2019.12869
50. Wang SV, Sreedhara SK, Schneeweiss S, et al. Reproducibility of real-world evidence studies using clinical practice data to inform regulatory and coverage decisions. *Nature Communications.* 2022;13(1). doi: 10.1038/s41467-022-32310-3
51. Lancaster HO. The combination of probabilities: an application of orthonormal functions. *Australian Journal of Statistics.* 1961;3(1):20–33. doi: 10.1111/j.1467-842x.1961.tb00058.x
52. Held L. The harmonic mean χ^2-test to substantiate scientific findings. *Journal of the Royal Statistical Society, Series C.* 2020;69(3):697–708. doi: 10.1111/rssc.12410
53. Liu Y, Xie J. Cauchy Combination Test: A Powerful Test With Analytic p-Value Calculation Under Arbitrary Dependency Structures. *Journal of the American Statistical Association.* 2019;115(529):393–402. doi: 10.1080/01621459.2018.1554485
54. Schweder T, Hjort NL. Discussion of "Confidence Distribution, the Frequentist Distribution Estimator of a Parameter: A Review". *International Statistical Review.* 2013;81(1):56–68. doi: 10.1111/insr.12004
55. R Core Team . *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; Vienna, Austria: 2017.

56. Hofmann F, Held L, Pawel S. *confMeta: Confidence Curves and P-Value Functions for Meta-Analysis*. ; : 2023. R package version 0.3.4, https://github.com/felix-hof/confMeta.
57. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.
58. Zhu H. *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. ; : 2024. R package version 1.4.0.
59. Wickham H, François R, Henry L, Müller K, Vaughan D. *dplyr: A Grammar of Data Manipulation*. ; : 2023. R package version 1.1.4.
60. Xie Y. *Dynamic Documents with R and knitr*. Boca Raton, Florida: Chapman and Hall/CRC. 2nd ed., 2015. ISBN 978-1498716963.
61. Nadarajah S, Kotz S. Exact Distribution of the Max/Min of Two Gaussian Random Variables. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*. 2008;16(2):210–212. doi: 10.1109/tvlsi.2007.912191

□

APPENDIX

A THE R PACKAGE TWOTRIALS

We have developed the `twotrials` R package for easily conducting combined p-value function inference based on the parameter estimates (with standard errors) from two trials. The package can be installed from the Comprehensive R Archive Network (CRAN) via the R command `install.packages("twotrials")`.

For every p-value combination method discussed in this paper, the package provides a combined p-value function (e.g., `pEdgington`) and a combined estimation function (e.g., `muEdgington`). While these can be used to manually compute p-values and parameter estimates, the convenience function `twotrials` automatically computes estimates and p-values based on all methods, and allows for easy printing and plotting of the results. The following code chunk illustrates its usage by reproducing the results for the 14-day treatment group from Table 3.

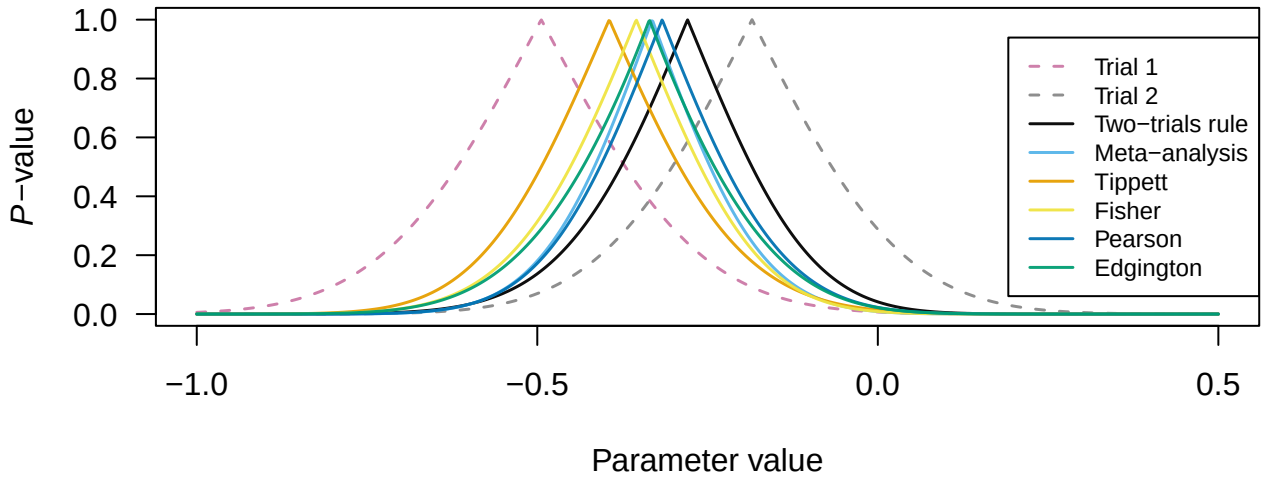
```
library(twotrials) # load package

## combine logRR estimates from RESPIRE trials
results <- twotrials(null = 0, t1 = -0.4942, t2 = -0.1847, se1 = 0.1833,
                     se2 = 0.1738, alternative = "less", level = 0.95)
print(results, digits = 2) # print summary of results

## INDIVIDUAL RESULTS
##    Trial Lower CL Estimate Upper CL P-value
##  Trial 1    -0.85    -0.49    -0.13  0.0035
##  Trial 2    -0.53    -0.18     0.16  0.1440
##
## COMBINED RESULTS
##           Method Lower CL Estimate Upper CL P-value   W1   W2
##  Two-trials rule    -0.57    -0.28   -0.011  0.0207 0.31 0.69
##    Meta-analysis    -0.58    -0.33   -0.084  0.0043 0.47 0.53
##          Tippett    -0.68    -0.39   -0.084  0.0070 0.68 0.32
##           Fisher    -0.64    -0.35   -0.087  0.0043 0.55 0.45
##          Pearson    -0.58    -0.32   -0.044  0.0114 0.43 0.57
##        Edgington    -0.64    -0.34   -0.048  0.0109 0.49 0.51
##
## NOTES
## Confidence level: 95%
## Null value: 0
## Alternative: less
```

Note that, for each combined estimate, the function also returns the weights w_1 (`W1`) and w_2 (`W2`). These represent the implicit linear weights of the point estimates from trials 1 and 2 towards the combined estimate, i.e., $\hat{\mu}(1/2) = w_1\hat{\theta}_1 + w_2\hat{\theta}_2$. Such weights aid interpretation by indicating how close each trial estimate is to the combined estimate. Finally, applying the plot function to the resulting object makes it easy to display the combined p-value functions, as demonstrated below.

```
plot(results, xlim = c(-1, 0.5), two.sided = TRUE) # plot p-value functions
```



B TECHNICAL DETAILS

This appendix contains technical details on the derivation of results from the main paper.

B.1 Expectation of the combined estimation function

Consider the random variables

$$X = \min\{\hat{\theta}_1 + \sigma_1\, q, \hat{\theta}_2 + \sigma_2\, q\} \qquad \text{and} \qquad Y = \max\{\hat{\theta}_1 - \sigma_1\, q, \hat{\theta}_2 - \sigma_2\, q\}. \tag{B1}$$

Table B1 shows that for certain choices of the constant q, X and Y are equal to the combined estimation functions from the two-trials rule (3) and Tippett's method (11) in the main paper, and the approximate combined estimation functions from Fisher's (B5), Pearson's (B6), and Edgington's methods (B7) and (B8) discussed below. Note that for Edgington's method (with $a = 1/2$) and meta-analysis, the distribution of the combined estimation function is normal and its expectation is therefore known and need not be derived here.

TABLE B1 Constants q for which X or Y are equal to the combined estimation function of a specific method.

Method	Alternative "greater"	Alternative "less"
Two-trials rule	X with $q = z_{\sqrt{a}}$	Y with $q = z_{\sqrt{a}}$
Tippett	Y with $q = z_{\sqrt{1-a}}$	X with $q = z_{\sqrt{1-a}}$
Fisher (approximate)	Y with $q = -z_{\exp\{-\chi^2_4(1-a)/2\}}$	X with $q = -z_{\exp\{-\chi^2_4(1-a)/2\}}$
Pearson (approximate)	X with $q = -z_{\exp\{-\chi^2_4(a)/2\}}$	Y with $q = -z_{\exp\{-\chi^2_4(a)/2\}}$
Edgington (approximate, $a < 1/2$)	X with $q = z_{\sqrt{2a}}$	Y with $q = -z_{\sqrt{2a}}$
Edgington (approximate, $a > 1/2$)	Y with $q = -z_{\sqrt{2(1-a)}}$	X with $q = z_{\sqrt{2(1-a)}}$

According to the assumptions stated at the beginning of Section 2, $\hat{\theta}_1$ and $\hat{\theta}_2$ are independent normal random variables with means θ_1, θ_2 and variances σ_1^2, σ_2^2. We can therefore use the results from Nadarajah and Kotz[61] on closed-form expressions for the moments of minima and maxima of bivariate Gaussian random vectors. That is, using their equations (9) and (11), it can be

shown that the expectations of X and Y are given by

$$E(X) = (\theta_1 + \sigma_1 q) \times \Phi\left(\frac{\theta_2 - \theta_1 + q(\sigma_2 - \sigma_1)}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right) + (\theta_2 + \sigma_2 q) \times \Phi\left(\frac{\theta_1 - \theta_2 + q(\sigma_1 - \sigma_2)}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right) - \sqrt{\sigma_1^2 + \sigma_2^2} \times \phi\left(\frac{\theta_2 - \theta_1 + q(\sigma_2 - \sigma_1)}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right)$$

and

$$E(Y) = (\theta_1 - \sigma_1 q) \times \Phi\left(\frac{\theta_1 - \theta_2 + q(\sigma_2 - \sigma_1)}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right) + (\theta_2 - \sigma_2 q) \times \Phi\left(\frac{\theta_2 - \theta_1 + q(\sigma_1 - \sigma_2)}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right) + \sqrt{\sigma_1^2 + \sigma_2^2} \times \phi\left(\frac{\theta_1 - \theta_2 + q(\sigma_2 - \sigma_1)}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right),$$

respectively. The expectation of the combined estimation function from a specific method are thus obtained by setting the constant q to the corresponding value. For example, the expectation of the median estimate ($a = 1/2$) from the two-trials rule with alternative "greater" and $\sigma_1 = \sigma_2 = \sigma$ in equation (6) is obtained from $E(X)$ with $q = z_{\sqrt{1/2}}$.

B.2 Median estimate standard errors

Since the median estimates from meta-analysis and Edgington's method are simple linear combinations of the trial effect estimates, their standard errors can be straightforwardly derived to be

$$\sigma_{\text{MA}} = \frac{1}{\sqrt{1/\sigma_1^2 + 1/\sigma_2^2}} \qquad \text{and} \qquad \sigma_{\text{E}} = \frac{\sqrt{2}}{1/\sigma_1 + 1/\sigma_2}.$$

By applying algebraic manipulations to $\sigma_{\text{MA}} \leq \sigma_{\text{E}}$, one can see that the meta-analytic standard error is never larger than Edgington's standard error, with equality if and only if the trial standard error coincide ($\sigma_1 = \sigma_2$). Similarly, by applying algebraic manipulations to $\sigma_{\text{E}} \leq \sigma_1$ and $\sigma_{\text{E}} \leq \sigma_2$, one can see that Edgington's standard error is only equal or smaller than either trial standard error if $\sqrt{2} - 1 \leq \sigma_2/\sigma_1 \leq 1/(\sqrt{2} - 1) = \sqrt{2} + 1$.

For the remaining methods, the (approximate) median estimates are given by X and Y in equation (B1) with $a = 1/2$. We can therefore use the results from Nadarajah and Kotz[61] to obtain their second moments

$$E(X^2) = \{\sigma_1^2 + (\theta_1 + \sigma_1 q)^2\} \Phi\left(\frac{\theta_2 - \theta_1 + q(\sigma_2 - \sigma_1)}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right) + \{\sigma_2^2 + (\theta_2 + \sigma_2 q)^2\} \Phi\left(\frac{\theta_1 - \theta_2 + q(\sigma_1 - \sigma_2)}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right) - \{\theta_1 + \theta_2 + q(\sigma_1 + \sigma_2)\} \sqrt{\sigma_1^2 + \sigma_2^2}\, \phi\left(\frac{\theta_2 - \theta_1 + q(\sigma_2^2 - \sigma_1^2)}{\sqrt{\sigma_1 + \sigma_2}}\right)$$

and

$$E(Y^2) = \{\sigma_1^2 + (\theta_1 - \sigma_1 q)^2\} \Phi\left(\frac{\theta_1 - \theta_2 + q(\sigma_2 - \sigma_1)}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right) + \{\sigma_2^2 + (\theta_2 - \sigma_2 q)^2\} \Phi\left(\frac{\theta_2 - \theta_1 + q(\sigma_1 - \sigma_2)}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right) + \{\theta_1 + \theta_2 - q(\sigma_1 + \sigma_2)\} \sqrt{\sigma_1^2 + \sigma_2^2}\, \phi\left(\frac{\theta_1 - \theta_2 + q(\sigma_2 - \sigma_1)}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right)$$

and corresponding standard errors. For example, assuming equal standard errors ($\sigma_1 = \sigma_2 = \sigma$), equal true effects ($\theta_1 = \theta_2 = \theta$) and the alternative "greater", the standard error of the two-trials rule median estimate (X with $q = z_{\sqrt{1/2}}$) simplifies to (2.1.1).

Figure B1 shows the standard errors from the median estimates of the two-trials rule, meta-analysis, Tippett's, and Edgington's methods for various scenarios of true trial effects and trial standard errors that should resemble typical ranges for standardized mean difference parameters. The standard errors from Fisher's and Pearson's methods are very close to the ones from Tippett's method and the two-trials rule, and therefore not shown to make the plot easier to read.

We can see that the standard error from meta-analysis is always the lowest and is always smaller than the minimum of the two standard errors ($\sigma_{\text{MA}} \leq \min\{\sigma_1, \sigma_2\}$). The standard error from Edgington's method is equal (when $\sigma_1 = \sigma 2$) to or larger than the meta-analytic one, and can exceed the minimum standard error from the two trials (e.g., for $\sigma_1 = 0.1$ and $\sigma_2 = 0.3$, it

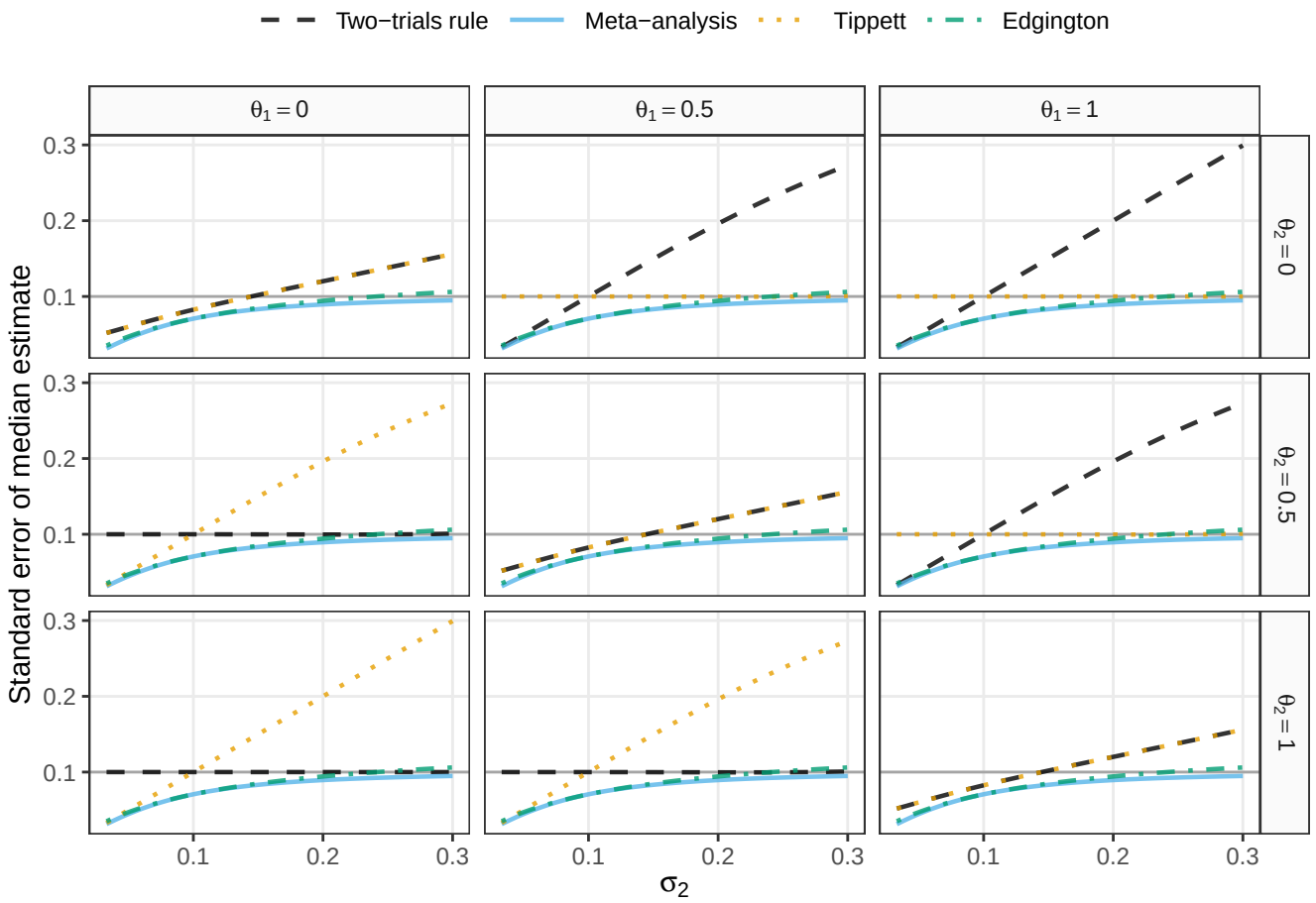


FIGURE B1 Comparison of median estimate standard errors across different scenarios of true trial effects θ_1 and θ_2 and trial standard error from the second trial σ_2. The standard error of the first trial is $\sigma_1 = 0.1$ across all scenarios. The standard errors from Fisher's and Pearson's methods are close to Tippett's method the two-trials rule, and not shown to make the plot easier to read.

is $\sigma_E = 0.106$). The standard errors for both Edgington's method and meta-analysis only depend on the trials' standard error, but not on the true effects, so their standard errors are the same across all panels in Figure B1. This is not the case case for the two-trials rule and Tippett's method, which show a more irregular behavior. When, the true trial effects coincide ($\theta_1 = \theta_2$; panels on the diagonal), the combined standard error decreases with decreasing standard error from the second trial σ_2. However, when the true effect from the first trial is smaller than the one from the second trial ($\theta_1 < \theta_2$; upper off-diagonal panels), the combined standard error from Tippett's method remains constant whereas the standard error from the two-trials rule changes drastically with with changing σ_2. The opposite occurs when the effect from the first trial is larger ($\theta_1 > \theta_2$; lower off-diagonal panels). This is plausible, as these methods target the minimum (two-trials rule) or maximum (Tippett) effect, meaning that the standard error of the trial with minimum or maximum effect mainly drives the standard error of the combined estimate.

B.3 Limiting combined estimation functions

Consider again the random variables Y and X as defined in Appendix B.1. Their cumulative distribution functions can be derived to be

$$\begin{aligned}\Pr(Y \leq y) &= \Pr(\max\{\hat{\theta}_1 - \sigma_1\, q, \hat{\theta}_2 - \sigma_2\, q\} \leq y) \\ &= \Pr(\hat{\theta}_1 - \sigma_1\, q \leq y, \hat{\theta}_2 - \sigma_2\, q \leq y) \\ &= \Pr(\hat{\theta}_1 - \sigma_1\, q \leq y) \times \Pr(\hat{\theta}_2 - \sigma_2\, q \leq y) \\ &= \Phi\left(\frac{y-\theta_1}{\sigma_1} + q\right) \times \Phi\left(\frac{y-\theta_2}{\sigma_2} + q\right)\end{aligned}$$

and

$$\begin{aligned}\Pr(X \leq x) &= \Pr(\min\{\hat{\theta}_1 + \sigma_1\, q, \hat{\theta}_2 + \sigma_2\, q\} \leq x) \\ &= 1 - \Pr(\hat{\theta}_1 + \sigma_1\, q > x, \hat{\theta}_2 + \sigma_2\, q > x) \\ &= 1 - \{\Pr(\hat{\theta}_1 + \sigma_1\, q > x) \times \Pr(\hat{\theta}_2 + \sigma_2\, q > x)\} \\ &= 1 - \left\{\Phi\left(\frac{\theta_1 - x}{\sigma_1} + q\right) \times \Phi\left(\frac{\theta_2 - x}{\sigma_2} + q\right)\right\}.\end{aligned}$$

Letting the standard errors σ_1 and σ_2 got to zero, this leads to

$$\begin{aligned}\lim_{\sigma_1,\sigma_2 \downarrow 0} \Pr(Y \leq y) &= 1_{[\theta_1,+\infty)}(y) \times 1_{[\theta_2,+\infty)}(y) \\ &= 1_{[\max\{\theta_1,\theta_2\},+\infty)}(y)\end{aligned} \tag{B2}$$

and

$$\begin{aligned}\lim_{\sigma_1,\sigma_2 \downarrow 0} \Pr(X \leq x) &= 1 - \{1_{(-\infty,\theta_1]}(x) \times 1_{(-\infty,\theta_2]}(x)\} \\ &= 1_{[\min\{\theta_1,\theta_2\},+\infty)}(x)\end{aligned} \tag{B3}$$

where $1_A(x) = 1$ if $x \in A$ and 0 otherwise, and the constant q vanishes. Since (B2) and (B3) is the cumulative distribution function of a degenerate random variable at $\max\{\theta_1, \theta_2\}$ and $\min\{\theta_1, \theta_2\}$, respectively, this implies that the combined estimation functions given by X and Y converge in probability to $\max\{\theta_1, \theta_2\}$ and $\min\{\theta_1, \theta_2\}$, respectively, for any constant q. Thus, all combined estimation functions from Table B1 converge in probability to $\max\{\theta_1, \theta_2\}$ or $\min\{\theta_1, \theta_2\}$ as σ_1 and σ_2 decrease.

B.4 Approximate combined estimation functions

Suppose that the trials' individual p-value functions

$$p_1(\mu) = \begin{cases} 1 - \Phi\left(\frac{\hat{\theta}_1 - \mu}{\sigma_1}\right) & \text{for alternative = "greater"} \\ \Phi\left(\frac{\hat{\theta}_1 - \mu}{\sigma_1}\right) & \text{for alternative "less"} \end{cases} \qquad p_2(\mu) = \begin{cases} 1 - \Phi\left(\frac{\hat{\theta}_2 - \mu}{\sigma_2}\right) & \text{for alternative = "greater"} \\ \Phi\left(\frac{\hat{\theta}_2 - \mu}{\sigma_2}\right) & \text{for alternative = "less"} \end{cases}$$

are "well-separated" in the sense that in the region where $p_1(\mu)$ changes from 0 to 1, $p_2(\mu)$ stays almost constant at 0 or 1, see the dotted lines in Figure B2 for an example. This happens when either the estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ are far apart and/or the standard errors σ_1 and σ_2 are small relative to the estimates (provided the estimates are not equal). Note that asymptotically the individual p-value functions approach the step functions

$$\lim_{\sigma_1 \downarrow 0} p_1(\mu) = \begin{cases} 1_{[\theta_1,+\infty)}(\mu) & \text{for alternative = "greater"} \\ 1_{(-\infty,\theta_1]}(\mu) & \text{for alternative = "less"} \end{cases} \qquad \lim_{\sigma_2 \downarrow 0} p_2(\mu) = \begin{cases} 1_{[\theta_2,+\infty)}(\mu) & \text{for alternative = "greater"} \\ 1_{(-\infty,\theta_2]}(\mu) & \text{foralternative = "less"} \end{cases}$$

Hence, with decreasing standard errors, the trials' p-value functions eventually become well-separated whenever the true effects θ_1 and θ_2 are unequal.

In case of well-separated p-value functions, we can approximate the combined p-value function from Fisher's, Pearson's, and Edgington's method by setting one of the p-values to 0 or 1, depending on alternative and combination method, and derive an approximate but closed-form combined estimation function. For example, in Figure B2 the combined p-value from Fisher's

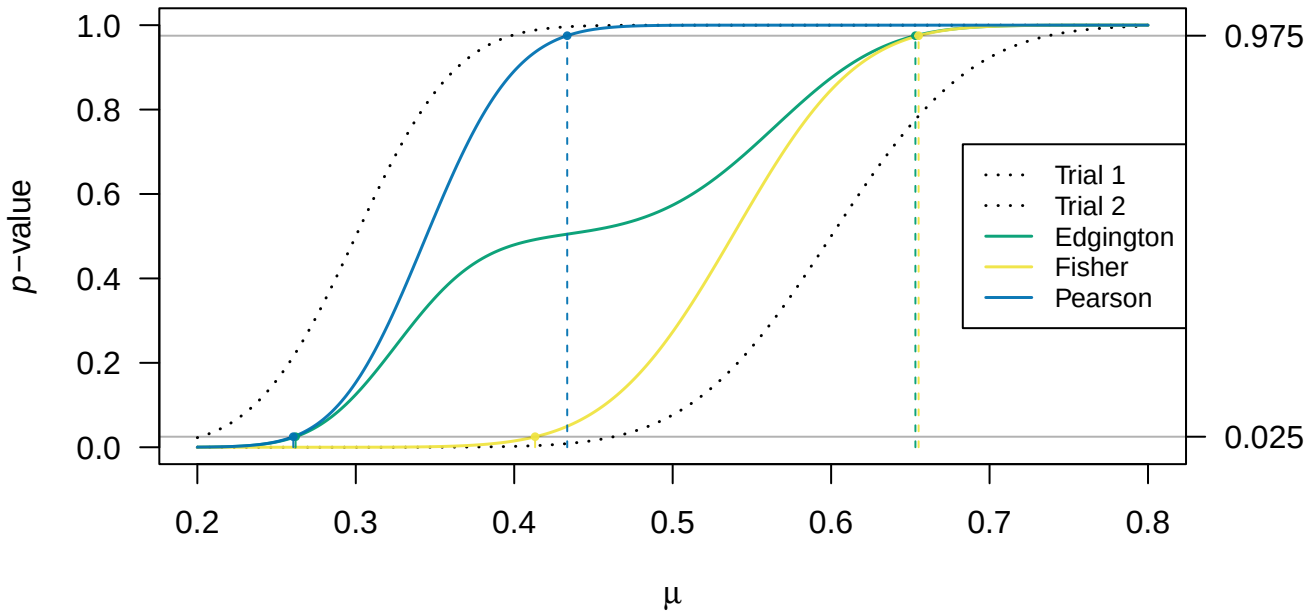


FIGURE B2 Two well-separated p-value functions (with alternative "greater") and the associated combined p-value functions based on Fisher's, Pearson's, and Edgington's methods. The dashed vertical lines and points denote the 95% CI limits computed with the approximate combined estimation functions (B5), (B6), and (B7). The effect estimates are $\hat{\theta}_1 = 0.3$ and $\hat{\theta}_2 = 0.6$ while the standard errors are $\sigma_1 = 0.05$ and $\sigma_2 = 0.07$.

method (12) remains virtually constant for increasing μ when the first individual p-function increases and only starts to increase as the second p-value function increases.

We may hence approximate Fisher's combined p-value by

$$p_{\mathrm{F}}(\mu) = \begin{cases} 1 - \Pr\left[\chi^2_4 \leq -2\log\left\{1 - \Phi\left(\max\left\{\frac{\hat{\theta}_1-\mu}{\sigma_1}, \frac{\hat{\theta}_2-\mu}{\sigma_2}\right\}\right)\right\}\right] & \text{for alternative = "greater"} \\ 1 - \Pr\left[\chi^2_4 \leq -2\log\left\{\Phi\left(\min\left\{\frac{\hat{\theta}_1-\mu}{\sigma_1}, \frac{\hat{\theta}_2-\mu}{\sigma_2}\right\}\right)\right\}\right] & \text{for alternative = "less".} \end{cases} \tag{B4}$$

The corresponding combined estimation function can then be obtained by equating (B4) to a and solving for μ, which leads to

$$\hat{\mu}_{\mathrm{F}}(a) = \begin{cases} \max\{\hat{\theta}_1 + \sigma_1 z_{\exp\{-\chi^2_4(1-a)/2\}}, \hat{\theta}_2 + \sigma_2 z_{\exp\{-\chi^2_4(1-a)/2\}}\} & \text{for alternative = "greater"} \\ \min\{\hat{\theta}_1 - \sigma_1 z_{\exp\{-\chi^2_4(1-a)/2\}}, \hat{\theta}_2 - \sigma_2 z_{\exp\{-\chi^2_4(1-a)/2\}}\} & \text{for alternative = "less".} \end{cases} \tag{B5}$$

The dashed yellow vertical lines in Figure B2 show the limits of a 95% CI computed via (B5), demonstrating that the approximation is accurate in this case, despite the finite standard errors.

In an analogous fashion, the combined p-value function based on Pearson's method can be approximated by

$$p_{\mathrm{P}}(\mu) = \begin{cases} \Pr\left[\chi^2_4 \leq -2\log\left\{\Phi\left(\min\left\{\frac{\hat{\theta}_1-\mu}{\sigma_1}, \frac{\hat{\theta}_2-\mu}{\sigma_2}\right\}\right)\right\}\right] & \text{for alternative = "greater"} \\ \Pr\left[\chi^2_4 \leq -2\log\left\{1 - \Phi\left(\max\left\{\frac{\hat{\theta}_1-\mu}{\sigma_1}, \frac{\hat{\theta}_2-\mu}{\sigma_2}\right\}\right)\right\}\right] & \text{for alternative = "less"} \end{cases}$$

leading to the approximate combined estimation function

$$\hat{\mu}_{\mathrm{P}}(a) = \begin{cases} \min\{\hat{\theta}_1 - \sigma_1 z_{\exp\{-\chi^2_4(a)/2\}}, \hat{\theta}_2 - \sigma_2 z_{\exp\{-\chi^2_4(a)/2\}}\} & \text{for alternative = "greater"} \\ \max\{\hat{\theta}_1 + \sigma_1 z_{\exp\{-\chi^2_4(a)/2\}}, \hat{\theta}_2 + \sigma_2 z_{\exp\{-\chi^2_4(a)/2\}}\} & \text{for alternative = "less".} \end{cases} \tag{B6}$$

The functions (B5) and (B6) based on Fisher's and Pearson's methods have a striking similarity to the combined estimation functions based on Tippett's method (11) and the two-trials rule (3), respectively, as they again involve shifted maxima/minima of the trial estimates.

In a similar way, Edgington's combined p-value function can be approximated by

$$p_{\mathrm{E}}(\mu) = \begin{cases} \left\{1 - \Phi\left(\min\left\{\frac{\hat{\theta}_1-\mu}{\sigma_1}, \frac{\hat{\theta}_2-\mu}{\sigma_2}\right\}\right)\right\}^2 / 2 & \text{if } \mu < \dfrac{\hat{\theta}_1/\sigma_1 + \hat{\theta}_2/\sigma_2}{1/\sigma_1 + 1/\sigma_2} \\ \dfrac{1}{2} & \text{if } \mu = \dfrac{\hat{\theta}_1/\sigma_1 + \hat{\theta}_2/\sigma_2}{1/\sigma_1 + 1/\sigma_2} \\ 1 - \left\{\Phi\left(\max\left\{\frac{\hat{\theta}_1-\mu}{\sigma_1}, \frac{\hat{\theta}_2-\mu}{\sigma_2}\right\}\right)\right\}^2 / 2 & \text{else} \end{cases}$$

for the alternative "greater" and with

$$p_{\mathrm{E}}(\mu) = \begin{cases} \left\{\Phi\left(\max\left\{\frac{\hat{\theta}_1-\mu}{\sigma_1}, \frac{\hat{\theta}_2-\mu}{\sigma_2}\right\}\right)\right\}^2 / 2 & \text{if } \mu > \dfrac{\hat{\theta}_1/\sigma_1 + \hat{\theta}_2/\sigma_2}{1/\sigma_1 + 1/\sigma_2} \\ \dfrac{1}{2} & \text{if } \mu = \dfrac{\hat{\theta}_1/\sigma_1 + \hat{\theta}_2/\sigma_2}{1/\sigma_1 + 1/\sigma_2} \\ 1 - \left\{1 - \Phi\left(\min\left\{\frac{\hat{\theta}_1-\mu}{\sigma_1}, \frac{\hat{\theta}_2-\mu}{\sigma_2}\right\}\right)\right\}^2 / 2 & \text{else} \end{cases}$$

for the alternative "less". Consequently, the approximate combined estimation function is

$$\hat{\mu}_{\mathrm{E}}(a) = \begin{cases} \min\{\hat{\theta}_1 + \sigma_1 z_{\sqrt{2a}}, \hat{\theta}_2 + \sigma_2 z_{\sqrt{2a}}\} & \text{for } a < 1/2 \\ \dfrac{\hat{\theta}_1/\sigma_1 + \hat{\theta}_2/\sigma_2}{1/\sigma_1 + 1/\sigma_2} & \text{for } a = 1/2 \\ \max\{\hat{\theta}_1 - \sigma_1 z_{\sqrt{2(1-a)}}, \hat{\theta}_2 - \sigma_2 z_{\sqrt{2(1-a)}}\} & \text{for } a > 1/2 \end{cases} \tag{B7}$$

for the alternative "greater" and

$$\hat{\mu}_{\mathrm{E}}(a) = \begin{cases} \max\{\hat{\theta}_1 - \sigma_1 z_{\sqrt{2a}}, \hat{\theta}_2 - \sigma_2 z_{\sqrt{2a}}\} & \text{for } a < 1/2 \\ \dfrac{\hat{\theta}_1/\sigma_1 + \hat{\theta}_2/\sigma_2}{1/\sigma_1 + 1/\sigma_2} & \text{for } a = 1/2 \\ \min\{\hat{\theta}_1 + \sigma_1 z_{\sqrt{2(1-a)}}, \hat{\theta}_2 + \sigma_2 z_{\sqrt{2(1-a)}}\} & \text{for } a > 1/2 \end{cases} \tag{B8}$$

for the alternative "less". The combined estimation functions (B7) and (B8) also include the closed-form solution for the median estimate ($a = 1/2$) from (17) as this value does not require any approximation. Surprisingly, a $(1-\alpha) \times 100\%$ CI constructed from (B7) or (B8) always includes the individual estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ since the lower limit ($a = \alpha/2$) is always smaller than the minimum of the two effect estimates, and the upper limit ($a = 1 - \alpha/2$) is always larger than the maximum of the two. This demonstrates that Edgington's method reacts to heterogeneity by widening its CI to include both trial effect estimates.

All of these approximations become more accurate with decreasing standard errors as the individual p-value functions become more separated. Since all approximate combined estimation functions (B5)–(B8) are essentially shifted minima and maxima (apart from the median estimate of Edgington's method), the results from Appendix B.3 apply. That is, as the standard errors σ_1 and σ_2 decrease toward zero, all minima converge in probability to $\min\{\theta_1, \theta_2\}$ while all maxima converge to $\max\{\theta_1, \theta_2\}$.

COMPUTATIONAL DETAILS

```
cat(paste(Sys.time(), Sys.timezone(), "\n"))

## 2025-06-16 15:33:04.443793 Europe/Zurich

sessionInfo()

## R version 4.4.1 (2024-06-14)
## Platform: x86_64-pc-linux-gnu
## Running under: Ubuntu 24.04.2 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.12.0
## LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.12.0
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=de_CH.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=de_CH.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=de_CH.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=de_CH.UTF-8 LC_IDENTIFICATION=C
##
## time zone: Europe/Zurich
## tzcode source: system (glibc)
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] dplyr_1.1.4      kableExtra_1.4.0 twotrials_0.6    ggplot2_3.5.2
## [5] confMeta_0.4.2   knitr_1.48
##
## loaded via a namespace (and not attached):
##  [1] generics_0.1.3          xml2_1.3.6              stringi_1.8.4
##  [4] lattice_0.22-6          lme4_1.1-36             hms_1.1.3
##  [7] digest_0.6.37           magrittr_2.0.3          evaluate_0.24.0
## [10] grid_4.4.1              meta_8.0-2              fastmap_1.2.0
## [13] CompQuadForm_1.4.3      Matrix_1.7-2            purrr_1.0.4
## [16] viridisLite_0.4.2       scales_1.3.0            numDeriv_2016.8-1.1
## [19] reformulas_0.4.0        Rdpack_2.6.2            cli_3.6.4
## [22] rlang_1.1.5             rbibutils_2.3           ReplicationSuccess_1.3.3
## [25] munsell_0.5.1           splines_4.4.1           withr_3.0.2
## [28] tools_4.4.1             tzdb_0.4.0              nloptr_2.1.1
## [31] minqa_1.2.8             metafor_4.8-0           colorspace_2.1-1
## [34] mathjaxr_1.6-0          boot_1.3-30             vctrs_0.6.5
## [37] R6_2.6.1                lifecycle_1.0.4         stringr_1.5.1
## [40] MASS_7.3-64             pkgconfig_2.0.3         pillar_1.10.1
## [43] gtable_0.3.6            glue_1.8.0              Rcpp_1.0.14
## [46] systemfonts_1.1.0       xfun_0.49               tibble_3.2.1
## [49] tidyselect_1.2.1        highr_0.11              rstudioapi_0.16.0
## [52] farver_2.1.2            htmltools_0.5.8.1       nlme_3.1-167
## [55] patchwork_1.3.0         labeling_0.4.3          svglite_2.1.3
## [58] rmarkdown_2.27          readr_2.1.5             compiler_4.4.1
## [61] metadat_1.5-0
```