Biometrics 72, 1378–1386
December 2016

DOI: 10.1111/biom.12497

Meta-Analysis Framework for Exact Inferences with Application to the Analysis of Rare Events

Guang Yang,[1,†] Dungang Liu,[2,*,†] Junyuan Wang,[3] and Min-ge Xie[4]

[1]Business Analytics, Dun & Bradstreet Inc., Short Hills, New Jersey 07078, U.S.A.
[2]University of Cincinnati Lindner College of Business, Cincinnati, Ohio 45221, U.S.A.
[3]Biostatistics and Programming, Merck Serono, Beijing 100022, China
[4]Department of Statistics and Biostatistics, Rutgers University, New Brunswick, New Jersey 08854, U.S.A.
[†]These authors contributed equally to this work.
[*]*email:* dungang.liu@uc.edu

Summary. The usefulness of meta-analysis has been recognized in the evaluation of drug safety, as a single trial usually yields few adverse events and offers limited information. For rare events, conventional meta-analysis methods may yield an invalid inference, as they often rely on large sample theories and require empirical corrections for zero events. These problems motivate research in developing *exact* methods, including Tian et al.'s method of combining confidence intervals (2009, *Biostatistics*, **10**, 275–281) and Liu et al.'s method of combining p-value functions (2014, *JASA*, **109**, 1450–1465). This article shows that these two exact methods can be unified under the framework of combining confidence distributions (CDs). Furthermore, we show that the CD method generalizes Tian et al.'s method in several aspects. Given that the CD framework also subsumes the Mantel–Haenszel and Peto methods, we conclude that the CD method offers a general framework for meta-analysis of rare events. We illustrate the CD framework using two real data sets collected for the safety analysis of diabetes drugs.

Key words: Clinical trial; Confidence distribution; Drug safety; Safety signal detecting; Zero event.

1. Introduction

During drug development and post-approval, it is important to enhance our understanding of drug safety (Barnes, 2007). Different from efficacy evaluation where a confirmatory clinical trial is often powered to make inference, safety evaluation often entails dealing with non-prespecified rare events and requires multiple trials for a reliable assessment. Under such circumstances, it is common to use meta-analysis, a systematic approach to integrating evidence from multiple studies, to strengthen inferences. The goal of this article is to show a general framework for *exact* meta-analysis, which is suited for analyzing rare events data and drawing conclusions on drug safety.

Our article is motivated by a recent safety evaluation of a diabetes drug in a class of medicines called SGLT2 inhibitors. Due to safety concerns, the Food and Drug Administration (FDA) did not approve its New Drug Application submission in December 2011 and requested additional data. The resubmission in 2013 provided data from 21 independent studies with respect to indexed bladder cancer. The observed cancer events are rare (see Web Table 1). The majority of the studies did not observe any cancer event and no study reported more than three cancer cases.

The aforementioned rare events setting is similar to another well-known safety study on rosiglitazone (Nissen and Wolski, 2007), where 48 clinical trials were analyzed to assess its risk of myocardial infarction and death from cardiovascular causes. While most of the trials reported few or zero adverse events, their meta-analysis concluded that rosiglitazone was associated with an elevated cardiovascular risk. This result led to FDA's action of issuing a black-box warning on the drug label and thousands of lawsuits filed against the manufacturer. On the other hand, this meta-analysis has been criticized as misleading and raised debates in the societies of medicine, law, and statistics (see Finkelstein and Levin, 2012 for a review of this case).

Although rare events are common in the studies of drug safety, it was only in recent years that scientists became fully aware of certain limitations of applying conventional statistical methods to such data. First, most meta-analysis methods, including Peto's method used in Nissen and Wolski (2007), rely on large sample theories to draw inferences. Such inference outcomes often do not remain valid in the rare events setting (Tian et al., 2009; Rücker et al., 2009; Liu et al., 2014). Second, when we observe a large portion of zero events in a cohort of clinical trials, commonly used methods require empirical continuity corrections for zero events to include all available data in the analysis. This approach is known to have an undesirable impact on the inference (Sweeting et al., 2004; Rücker et al., 2009; Liu et al., 2014). To solve these problems, Tian et al. (2009) and Liu et al. (2014) developed *exact* meta-analysis methods. The former is based on combining confidence intervals and the latter is based on combining p-value functions.

This article shows that Tian et al.'s and Liu et al.'s methods can be unified under the framework of combining confidence distributions (CDs; see Xie, Singh, and Strawderman, 2011), which implies that the two methods are equivalent in a certain

sense. We also show that the CD framework is more general in the senses that (1) Tian et al.'s method involves a choice of confidence levels, whereas the CD method can integrate the confidence intervals of all possible levels; (2) the CD method affords a variety of "transformation functions" in the combination, whereas Tian et al.'s specifically corresponds to the logit transformation.

A confidence distribution is referred to as a sample-dependent distribution function that can represent the confidence intervals of all levels for a parameter of interest (see, e.g., Cox (1958); Efron (1993); and the review in Xie and Singh (2013)). Cox (2013) stated that the CD approach aims to provide "simple and interpretable summaries of what can reasonably be learned from data (and an assumed model)". For example, consider a simple normal sample $\boldsymbol{x} = \{x_i, i = 1, \ldots, n\}$, where $x_i \sim \mathcal{N}(\mu, 1)$. To make an inference about μ, we may use a point estimate $\bar{x}_n = \sum_{i=1}^{n} x_i/n$ or an interval estimate $(\bar{x}_n - 1.96/\sqrt{n}, \bar{x}_n + 1.96/\sqrt{n})$. To make an inference based on a CD, we use $\mathcal{N}(\bar{x}_n, 1/n)$ or its cumulative distribution function $H(\mu) = \Phi(\sqrt{n}(\mu - \bar{x}_n))$, to estimate μ. The function $H(\mu)$ depends on both the sample $\boldsymbol{x}$ and the parameter μ. It is a distribution function on the parameter space of μ when given a sample $\boldsymbol{x}$. From this distribution function, we can derive commonly used inference outcomes. For example, $(H^{-1}(\alpha/2), H^{-1}(1-\alpha/2)) = (\bar{x}_n + \Phi^{-1}(\alpha/2)/\sqrt{n}, \bar{x}_n + \Phi^{-1}(1-\alpha/2)/\sqrt{n})$ provides a $(1-\alpha)100\%$ confidence interval for μ, for any $0 < \alpha \leq 1$. The mean/median of the distribution estimator $\mathcal{N}(\bar{x}_n, 1/n)$ provides a point estimator $\bar{x}_n$ for μ. The tail mass $H(b) = \Phi(\sqrt{n}(b - \bar{x}_n))$ provides a p-value for the one-sided hypothesis test $K_0 : \mu \leq b$ versus $K_1 : \mu > b$. This example illustrates the capacity of a CD as a vehicle carrying different types of inference outcomes. It is this capacity that explains why the CD concept is useful for developing a general meta-analysis framework.

The CD concept has gained interest in recent years (Xie and Singh, 2013). It has been used to develop meta-analysis methods for a variety of problems, such as how to incorporate expert opinions in clinical analysis (Xie et al., 2013), how to perform network analysis without using subjective prior information (Yang et al., 2014), how to make an inference on fixed study-specific parameters (Claggett et al., 2014), and how to efficiently integrate heterogeneous studies (Liu et al., 2015).

2. CD Framework for Meta-Analysis

In this section, we review the CD framework for meta-analysis. The idea is to combine functions derived from the sample of each study. The overall inference is based on an integrated function, rather than a combined point estimate. When making inferences for a common parameter Δ across K independent studies, conventional meta-analysis methods extract point estimates $\widehat{\Delta}_1, \ldots, \widehat{\Delta}_K$ from individual studies and make inferences based on the combined point estimate:

$$\widehat{\Delta}^{(c)} = (w_1\widehat{\Delta}_1 + \cdots + w_K\widehat{\Delta}_K)/(w_1 + \cdots + w_K), \qquad (1)$$

where $w_k \geq 0$ are weights with at least one $w_k \neq 0$.

Different from the conventional method, the CD method obtains "distribution estimates" $H_1(\Delta), \ldots, H_K(\Delta)$, i.e., CD functions, from individual studies. The overall inference is drawn from the integrated function (Singh et al., 2005):

$$H^{(c)}(\Delta) = G^{(c)}\{w_1\psi(H_1(\Delta)) + \cdots + w_K\psi(H_K(\Delta))\}, \qquad (2)$$

where $\psi(\cdot)$ is a monotonic "transformation function" (e.g., the inverse of a continuous cumulative distribution function), $w_k \geq 0$ are generic weights, and $G^c(\cdot)$ is the distribution function of $w_1\psi(U_1) + \cdots + w_K\psi(U_K)$ where $U_1, \ldots, U_K$ are independent random variables following the uniform distribution U(0,1). In the rest of this article, we will use examples to illustrate how to derive CDs and integrate them to perform meta-analysis.

The CD combination method (2) leads to a meta-analysis framework (Xie et al., 2011). Xie et al. (2011) showed that this framework subsumes point estimates combination method and p-values combination method. This article shows that the CD method also subsumes Tian et al.'s method of combining confidence intervals. The reason that the CD method can unify these methods is that a CD function is loaded with point estimates, interval estimates, and p-values. As a result, the integrated CD function in (2) still carries the combined point/interval estimate and the combined p-value.

2.1. *Point Estimates Combination*

Most meta-analyses are performed by combining point estimates derived from individual studies. The point estimates combination is a special case of the CD combining framework (2) (c.f., Xie et al., 2011). To see this, we consider a commonly used model where it is assumed that

$$\widehat{\Delta}_k \overset{\text{ind}}{\sim} \mathcal{N}(\Delta, \sigma_k^2). \qquad (3)$$

Here, $\widehat{\Delta}_k$ is a summary statistic from the k-th study and $\sigma_k^2 = \text{var}(\widehat{\Delta}_k)$ is assumed to be known. This general model (3) is often used in the traditional domain of meta-analysis. It applies to both fixed-effects meta-analysis (where σ_k^2 represents the within-study variation s_k^2) and random-effects meta-analysis (where σ_k^2 represents s_k^2 plus the between-study variation τ^2). Model (3) applies to any parameter of interest such as risk difference, relative risk, and odds ratio. See Normand (1999) for a review.

Given (3), the point estimates combination method (1) yields a combined point estimate

$$\widehat{\Delta}^{(c)} = \frac{\sum_{k=1}^{K} \widehat{\Delta}_k/\sigma_k^2}{\sum_{k=1}^{K} 1/\sigma_k^2}, \qquad (4)$$

using weight $w_k = 1/\sigma_k^2$. The inference is based on the normality of the combined point estimate $\widehat{\Delta}^{(c)}$, coupled with its variance $\sigma_c^2 = 1/\{\sum_{k=1}^{K} 1/\sigma_k^2\}$.

Under the same model assumption (3), we can show that the distribution function $H_k(\Delta) = \Phi\left(\frac{\Delta - \widehat{\Delta}_k}{\sigma_k}\right)$ is a CD for Δ. Applying the CD combination method (2) yields a combined function

$$H^{(c)}(\Delta) = \Phi\left(\frac{\Delta - \widehat{\Delta}^{(c)}}{\sigma_c}\right), \qquad (5)$$

Table 1
Typical layout of a 2×2 *table*

	Event	No event	Total
Treatment	X	$n_T - X$	n_T
Control	Y	$n_C - Y$	n_C
Total	T	$N - T$	N

using the transformation function $\psi(\cdot) = \Phi^{-1}(\cdot)$ and weight $w_k = 1/\sigma_k$. The resulting function $H^{(c)}(\Delta)$ is a distribution function on the parameter space of Δ, and it is used as a distribution estimate of Δ in the domain of frequentist inference. From $H^{(c)}(\Delta)$, we can draw other inference outcomes. For example, the mean of $H^{(c)}(\Delta)$ can be used as a point estimate for Δ. In this case, it is the same as $\widehat{\Delta}^{(c)}$ in (4) obtained from the point estimates combination. This observation implies that the point estimates combination in (4) is a special case of the CD combination.

Example 1. *(Meta-analysis of* 2×2 *tables) In clinical research, the outcomes of interest are often reported in the form of a* 2×2 *table as shown in Table 1. To examine the difference between the event rates of treatment* (p_T) *and control* (p_C) *groups, we consider making inferences for certain risk measures, such as the log odds ratio (LOR) defined as* $\Delta = \log\left(\frac{p_T/(1-p_T)}{p_C/(1-p_C)}\right)$. *It is known that the point estimate* $\widehat{\Delta} = \log\left(\frac{X/(n_T-X)}{Y/(n_C-Y)}\right)$ *follows approximately normal distribution* $\mathcal{N}(\Delta, s^2)$, *as both* n_T *and* $n_C \to \infty$, *where* $s^2 = \frac{1}{X} + \frac{1}{n_T-X} + \frac{1}{Y} + \frac{1}{n_C-Y}$. *Thus,* $H_A(\Delta) = \Phi\left((\Delta - \widehat{\Delta})/s\right)$ *is a CD for* Δ *asymptotically. In the presence of multiple* 2×2 *tables with a common LOR* Δ, *we use Figure 1 to illustrate the relationship between combining point estimates as in (4) and combining CD functions as in (5). Specifically, the mean/median/mode of the combined confidence distribution (dashed curve) yields a combined point estimate* $\widehat{\Delta}^{(c)}$ *(triangle), which is the same as the one obtained from the point estimates combination.*

2.2. *P-Values (Functions) Combination*

Another classical approach for meta-analysis is to combine p-values obtained from testing a common hypothesis associated with multiple studies. Consider testing a left-sided hypothesis

$$Q_0 : \Delta \le \Delta_0 \quad \text{versus} \quad Q_1 : \Delta > \Delta_0, \tag{6}$$

for some particular value Δ_0 of interest. The classical approach combines the p-value p_i derived from each of the studies and draws the overall inference based on the combined p-value $p^{(c)} \equiv p^{(c)}(p_1, \ldots, p_K)$. Commonly used methods include Fisher's combination (Fisher, 1932)

$$p^{(c)} = \Pr\left\{\chi^2_{2K} \ge -2\sum_{k=1}^{K}\log(p_k)\right\}, \tag{7}$$

where χ^2_{2K} denotes a χ^2-distributed random variable with $2K$ degrees of freedom, and Stouffer's combination (Stouffer et al.,

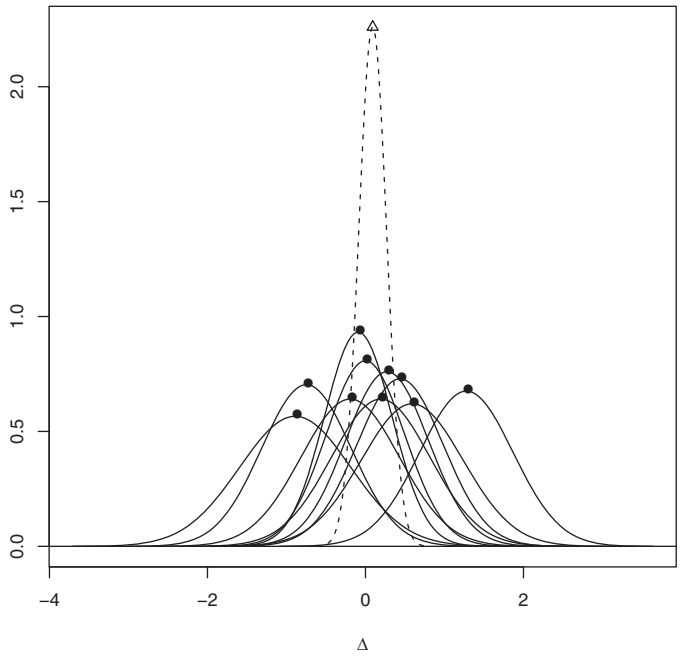

Figure 1. An illustration of meta-analysis by combining confidence distributions in the context of Example 1. Each of the $K = 10$ solid curves represents a confidence density function derived from a 2×2 table simulated from the setting where $\Delta = 1.2, n_T = n_C = 100$ and $p_C \sim U(0.05, 0.1)$. The solid circles represent the point estimates $\widehat{\Delta}_k$ from individual studies. The combined confidence density function is plotted using a dashed curve, together with the combined point estimate $\widehat{\Delta}^{(c)}$ depicted using a triangle.

1949)

$$p^{(c)} = \Phi\left(\frac{1}{\sqrt{K}}\left\{\Phi^{-1}(p_1) + \Phi^{-1}(p_2) + \cdots \Phi^{-1}(p_K)\right\}\right). \tag{8}$$

Additional p-values combination methods can be found in Marden (1991).

It is shown in Xie et al. (2011) that the p-values combination is a special case of the CD framework (2). In fact, Liu et al.'s method of combining p-value functions can also be unified under the same framework. To see this, we consider varying the value of Δ_0 in the hypothesis testing problem (6). Then, $p_k \equiv p_k(\Delta_0)$ becomes a function over the parameter space of Δ, and we denote it by $p_k(\cdot)$. This function is known as a *significance function* (Fraser, 1991) or *p-value function* (see, e.g., Singh et al., 2005; Xie et al., 2011; Liu et al., 2014). Such a function $p_k(\cdot)$ is typically a distribution function on the parameter space—it is increasing and bounded below by 0 and above by 1. The p-value function $p_k(\cdot)$ is usually a CD (at least asymptotically) (Singh et al., 2005). If we let $H_k(\cdot) = p_k(\cdot)$, p-values combination methods, such as the ones in (7) and (8), are special cases of the CD framework (2). To see this, let $\psi(\cdot) = \log(\cdot)$ and $w_k = 1$ for all k in (2); then, the combined function $H^{(c)}(\cdot)$ can yield a p-value $H^{(c)}(\Delta_0)$, which is the same as $p^{(c)}$ in Fisher's combination (7). Similarly, let $\psi(\cdot) = \Phi^{-1}(\cdot)$ and $w_k = 1$, and then, $H^{(c)}(\cdot)$ can yield a p-value $H^{(c)}(\Delta_0)$, which is the same as $p^{(c)}$ in Stouffer's combination (8). For other forms of p-values combination that can be written as special cases of CDs combination, we refer readers to Xie et al. (2011).

The CDs combination method generalizes the classical p-values combination methods (Xie et al., 2011; Liu et al., 2014). First, the classical methods combine the observed p-values, a set of single values from individual studies, whereas the CD method combines p-value functions, a set of functions (on the parameter space) from individual studies. Second, the classical methods use equal weights in the combination (as seen in formulas (7) and (8)), which may lead to a decline in the Fisher efficiency. The CD method (2) offers a flexible choice of weights. The use of non-trivial weights can improve the Fisher efficiency and yield robust inferences in the presence of heterogeneous studies (see, e.g., Singh et al., 2005; Xie et al., 2011; Liu et al., 2014).

Incorporating the classical p-values combination into the CD framework leads to a class of *exact* meta-analysis methods for discrete data (Liu et al., 2014). As stated in the introduction, exact inferences are desired in meta-analysis of rare events where typical normal approximations (such as those relied on for model (3)) may not work well. The key idea of Liu et al. (2014) is to derive p-value functions from exact test results. Such p-value functions preserve small-sample properties, which are inherited from the exact distributions of test statistics. This point is illustrated in the following example.

Example 1 (Continued) *(Analysis of rare events) In the setting of rare events, the probability of observing the event of interest is very low. Consider Table 1 as an example. Suppose we observe $x = 4$ events out of 300 cases in the treatment group ($\hat{p}_T = 1.33\%$) and $y = 1$ out of 100 controls ($\hat{p}_C = 1.00\%$). To illustrate Liu et al.'s exact method, we plot in Figure 2(a), a p-value function $p(\cdot)$ for the log odds ratio Δ (in the solid curve) obtained from the mid-p adaption of Fisher's exact test. This p-value function is presented in its density form (i.e., its derivative function). As a reference, we also include in Figure 2(a), a confidence density function (in the dashed curve) obtained from the normal approximation to $\widehat{\Delta}_k$ as seen in model (3). This dashed curve is derived in the same way as the curves in Figure 1. The comparison between the solid and dashed curves shows that the p-value function $p(\cdot)$ based on the exact test preserves the skewness in the exact distributions of the statistics in Table 1.*

In the setting of rare events, the normal approximation may not work well even when it is applied to log-transformed statistics. In our example, parameter Δ and its estimate have already been log-transformed. The inadequacy of applying log-transformation is exemplified in Figure 2(b), where we

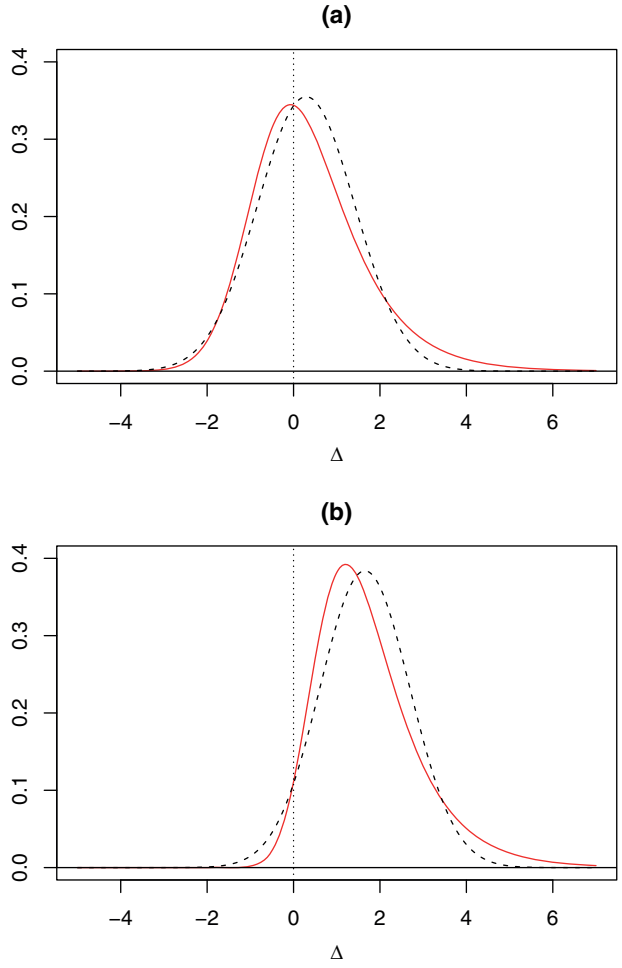

Figure 2. An illustration of p-value functions derived from exact test results (in solid curves) and normal approximations (in dashed curves). Displayed are two cases using: (a) the data in Table 2; and (b) the data in Table 2 with the observed events in the treatment group $x = 4$ replaced by $x = 15$.

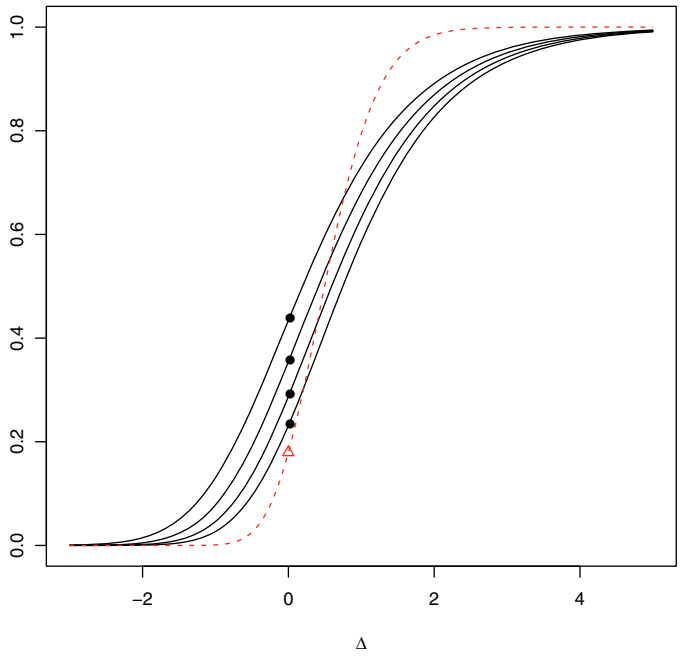

Figure 3. An illustration of combining p-value functions in the rare events setting. Plotted in solid curves are individual p-value functions (in its cumulative form) obtained from four 2×2 tables, all having the data layout of Table 2 with X takes different realizations, i.e., $x = 4, 5, 6,$ and 7. The dashed curve represent the combined p-value function. The solid circle and triangle denote the individual and combined p-value, respectively, for testing the hypothesis $Q_0 : \Delta \leq 0$ versus $Q_1 : \Delta > 0$

*change the observed events x in the treatment group to $x = 15$ ($\hat{p}_T = 5\%$). This is the only difference between Figure 2(b) and (a). We observe in Figure 2(b) that the tails of the solid and dashed curves are distinct on both sides. This implies that inferences based on the normal approximations could be different from that from exact tests, considering that a p-value for testing the hypothesis (6) is calculated as the area under the curve to the left of the dotted vertical line. In this example, the distinct behavior in tails yields different p-values (*0.034 *and* 0.056 *from the solid and dashed curves, respectively) and results in disparate conclusions.*

Figure 3 illustrates the combination of p-value functions. The solid curves represent individual p-value functions for four 2×2 tables, all having the data layout of Table 2 where X takes different realizations ($x = 4, 5, 6,$ and 7). The solid circles represent individual p-values for testing the hypothesis $Q_0 : \Delta \leq 0$ versus $Q_1 : \Delta > 0$. Plotted in the dashed curve is the combined p-value function, from which we can draw a combined p-value (shown as the triangle in Figure 3) and a combined confidence interval. In the following section, we show that under the CD framework, such a combined confidence interval is, in a certain sense, equivalent to the one proposed by Tian et al. (2009).

3. From Combining CIs to Combining CDs

Tian et al. (2009) proposed an exact meta-analysis method, which can include in the analysis all available data without using continuity corrections to zero events. Their idea is to combine confidence intervals derived from certain exact procedures. In this section, we show that (1) Tian et al.'s method of combining confidence intervals (CIs) is a special case of the CD framework; and (2) the CD method generalizes Tian et al.'s method in several aspects.

3.1. *Tian et al.'s Method of Combining CIs*

Consider constructing a $100(1-\alpha)\%$ one-sided CI $(a, +\infty)$ for Δ. Tian et al. (2009) proposed to consider study-specific one-sided η-level CIs $(a_k, +\infty)$ for Δ, for a given η. To integrate K study-specific CIs, they proposed to, for any fixed Δ, say Δ_0, examine how many CIs include Δ_0 as an interior point. If Δ_0 is the true value of Δ, then Δ_0 should be included in at least $100\eta\%$ of the K CIs under consideration. This fact is used as a criterion to examine all possible values of Δ. The points satisfying the criterion form a combined CI $(a, +\infty)$ for Δ. An explicit expression of such a $100(1-\alpha)\%$ CI is

$$\mathrm{CI}_\eta = \left\{ \Delta \,\middle|\, \sum_{k=1}^{K} w_k \{ \mathbf{I}(\Delta > a_k) - \eta \} \geq d_{\alpha,\eta} \right\}, \qquad (9)$$

where $\mathbf{I}(\cdot)$ is an indicator function and w_k is the weight assigned to the k-th study. The value of $d_{\alpha,\eta}$ is chosen such that $\Pr\left\{ \sum_{k=1}^{K} w_k (B_k - \eta) < d_{\alpha,\eta} \right\} \leq \alpha$, where $B_1, \ldots, B_K$ are K independent Bernoulli random variables with a success probability of η. Note that if Δ_0 is the true value of Δ, the random variable $\mathbf{I}(\Delta_0 > a_k)$ has the same distribution as B_k.

To improve inference efficiency, Tian et al. proposed to use multiple levels of CIs. Generally, they consider M levels $\{\eta_1, \ldots, \eta_M\}$ of confidence intervals $J_{km} = (a_{km}, \infty)$ $(m = 1, \ldots, M)$ for the k-th study, where $0 < \eta_1 < \ldots < \eta_M < 1$ and $a_{k1} \geq \cdots \geq a_{kM}$. Then, the aggregated $100(1-\alpha)\%$ confidence

interval is

$$\mathrm{CI}^{agg}_{\{\eta_m\}} = \left\{ \Delta \left| \sum_{k=1}^{K} w_k \sum_{m=1}^{M} \widetilde{w}_m \{\mathbf{I}(\Delta > a_{km}) - \eta_m\} \geq d_\alpha \right. \right\}, \tag{10}$$

where $\widetilde{w}_m$ is the within-study weight assigned to the η_m-level interval. Tian et al. suggested using $\widetilde{w}_m$ that is proportional to $\{\eta_m(1-\eta_m)\}^{-1}$.

3.2. *Tian et al.'s Method under the CD Framework*

In this subsection, we show that Tian et al.'s method is a special case of the CD framework. The intuition comes from the fact that a CD can *represent the confidence intervals of all levels* (i.e., for all $\eta \in (0,1)$) for the parameter of interest Δ. In fact, the classical way (Cox, 1958) of defining a CD is to invert the upper (lower) end of a one-sided level-η confidence interval. In the current context, consider the level-η_m CI $J_{km} = (a_{km}, \infty)$ for Δ in the k-th study for example. The lower end $a_{km} \equiv a_{km}(\eta_m)$ of the interval J_{km} is a function of η_m, as depicted on the left-side of Web Figure 1. A confidence distribution for Δ can be obtained by inverting the function $a_{km}(\eta_m)$ represented by the curve in Web Figure 1, such that $a_{km} = H_k^{-1}(1-\eta_m)$ for any $\eta_m \in (0,1)$. In other words, a CD for Δ can be defined as a function such that $H_k(a_{km}) = 1-\eta_m$, as depicted on the right-side of Web Figure 1. Such a function $H_k(\cdot)$ encompasses a set of CIs for all possible confidence levels.

Formally, we make the following assumption.

Assumption 1. *For any given sample $\boldsymbol{x}_k$ in the k-th study and any level-η_m one-sided confidence interval $J_{km} = (a_{km}, \infty)$ for the parameter Δ, we assume that there exists a continuous and increasing function $H_k(\cdot)$, defined on the parameter space of Δ, such that $H_k(a_{km}) = 1-\eta_m$.*

When Assumption 1 holds, the function $H_k(\cdot)$ is a CD for Δ. We show below that the CD framework (2) can yield a combined confidence interval that is the same as Tian et al.'s aggregated interval $\mathrm{CI}^{agg}_{\{\eta_m\}}$ in (10). Specifically, we take the transformation function in (2) to be

$$\psi_M(u) = \sum_{m=1}^{M} \widetilde{w}_m(\mathbf{I}(u > 1-\eta_m) - \eta_m). \tag{11}$$

Here, $\psi_M(u)$ is a non-decreasing step function. Then, the combined CD derived from (2) becomes

$$H^{(c)}_{\psi_M}(\Delta) = G^{(c)}_{\psi_M}\left\{ \sum_{k=1}^{K} w_k \sum_{m=1}^{M} \widetilde{w}_m (\mathbf{I}\{H_k(\Delta) > 1-\eta_m\} - \eta_m) \right\}. \tag{12}$$

Here, $G^{(c)}_{\psi_M}(t) = \Pr\left\{\sum_{k=1}^{K} w_k\, \psi_M(U_k) \leq t\right\}$ by definition (Section 2), where $U_1, \ldots, U_K$ are independent random variables following U(0,1). Plugging in the specific $\psi_M(\cdot)$ in (11), we have $G^{(c)}_{\psi_M}(t) = \Pr\left\{ \sum_{k=1}^{K} w_k \sum_{m=1}^{M} \widetilde{w}_m (B_{km} - \eta_m) \leq t \right\}$, where the random variables $B_{km} = \mathbf{I}(U_k > 1-\eta_m)$.

Based on the combined CD $H^{(c)}_{\psi_M}(\Delta)$ in (12), we can derive a one-sided $100(1-\alpha)\%$ confidence interval $\mathrm{CI}_{\psi_M} = \left\{ \Delta \left| H^{(c)}_{\psi_M}(\Delta) \geq \alpha \right. \right\}$, which is

$$\begin{aligned}
\mathrm{CI}_{\psi_M} &= \left\{ \Delta \left| G^{(c)}_{\psi_M} \left\{ \sum_{k=1}^{K} w_k \right. \right. \right. \\
&\qquad \left. \left. \left. \sum_{m=1}^{M} \widetilde{w}_m \{\mathbf{I}(H_k(\Delta) > 1-\eta_m) - \eta_m\} \right\} \geq \alpha \right. \right\} \\
&= \left\{ \Delta \left| \sum_{k=1}^{K} w_k \sum_{m=1}^{M} \widetilde{w}_m \{\mathbf{I}(\Delta > a_{km}) - \eta_m\} \right. \right. \\
&\qquad \left. \left. \geq G^{(c)^{-1}}_{\psi_M}(\alpha) \right. \right\}.
\end{aligned}$$

The last equation shows that CI_{ψ_M} is the same as Tian et al.'s interval in (10). Note that d_α in (10) is $G^{(c)^{-1}}_{\psi_M}(\alpha)$. Thus, we have established the following result.

Proposition 1. *If Assumption 1 holds, then Tian et al.'s aggregated confidence interval in (10) corresponds to a special case of combining confidence distributions $H_1(\cdot), \ldots, H_K(\cdot)$ as in the framework (2), with the transformation function being $\psi_M(u) = \sum_{m=1}^{M} \widetilde{w}_m(\mathbf{I}(u > 1-\eta_m) - \eta_m)$. More specifically, $CI^{agg}_{\{\eta_m\}} \equiv CI_{\psi_M}$.*

3.3. *Integrating the CIs for All Confidence Levels*

To apply Tian et al.'s method, we have to specify a number of confidence levels $\{\eta_m, m = 1, \ldots, M\}$. It is unknown that without compromising efficiency, how many levels should be used in practice and what specific levels should be included in the analysis. We show that the CD method can eliminate the ad hoc choosing of confidence levels and achieve the integration of the CIs of all possible confidence levels.

We have considered the transformation function in (11) to establish Proposition 1. We now derive the limiting form of $\psi_M(u)$ as the number of confidence levels M goes to infinity and the adjacent levels become sufficiently close to each other. Let $\delta_m = \eta_m - \eta_{m-1}$ and define

$$\begin{aligned}
\psi_\infty(u) &= \lim_{\max \delta_m \to 0; M \to \infty} \delta_m \psi_M(u) \\
&= \lim_{\max \delta_m \to 0; M \to \infty} \sum_{m=1}^{M} \delta_m \widetilde{w}_m (\mathbf{I}(u > 1-\eta_m) - \eta_m).
\end{aligned}$$

Inputting the within-study weight $\widetilde{w}_m = \{\eta_m(1-\eta_m)\}^{-1}$ as used in Tian et al. (2009), we get

$$\psi_\infty(u) = \int_0^1 \{\mathbf{I}(u > 1-\eta) - \eta\}/\{\eta(1-\eta)\}d\eta$$
$$= \ln\left(\frac{u}{1-u}\right) = \text{logit}(u). \quad (13)$$

The result shows that the transformation function $\psi_M(u)$, which is a step-function, converges to a continuous logit function as the confidence levels become sufficiently dense. For illustration purposes, we plot in Web Figure 2 the step-function $\psi_M(u)$ (scaled by a factor of $\delta = \delta_m$) in a solid curve along with its limiting form $\psi_\infty(u) = \text{logit}(u)$ in a dashed curve. The figure shows that $\psi_M(u)$ is an approximation to the logit function, and the approximation may be adequate except in the two tails when $M = 20$.

Using the transformation function $\psi_\infty(u)$ in the CD formula (2), we can derive a new combined CD

$$H^{(c)}_{\psi_\infty}(\Delta) = G^{(c)}_{\psi_\infty}\left\{\sum_{k=1}^K w_k \psi_\infty(H_k(\Delta))\right\}$$
$$= G^{(c)}_{\psi_\infty}\left\{\sum_{k=1}^K w_k \ln\left(\frac{H_k(\Delta)}{1-H_k(\Delta)}\right)\right\}, \quad (14)$$

where $G^{(c)}_{\psi_\infty}(t) = \Pr\left\{\sum_{k=1}^K w_k\, \psi_\infty(U_k) \le t\right\}$. The CD function $H^{(c)}_{\psi_\infty}(\Delta)$ yields a one-sided $100(1-\alpha)\%$ confidence interval $\text{CI}_{\psi_\infty} = \left\{\Delta \,\middle|\, H^{(c)}_{\psi_\infty}(\Delta) \ge \alpha\right\}$. The proposition below is a result of the continuity and monotonicity of $H_k(\Delta)$'s. It implies that combining CIs for all possible confidence levels is equivalent to combining CDs.

Proposition 2. *If Assumption 1 holds and the within-study weight is given as* $\widetilde{w}_m = \{\eta_m(1-\eta_m)\}^{-1}$*, then the lower end of Tian et al.'s aggregated confidence interval* $CI^{agg}_{\{\eta_m\}}$ *in (10) converges to the lower end of the confidence interval* CI_{ψ_∞} *derived from the CD function* $H^{(c)}_{\psi_\infty}(\Delta)$ *in (14), as* $M \to \infty$ *and* $\max \delta_m \to 0$.

An implication of Proposition 2 is that the CD method offers a tractable means to study the impact of within-study weights in Tian et al.'s method. Such impact can be explained by the impact of the transformation functions in the CD framework. Recall that Tian et al.'s within-study weight $\widetilde{w}_m = \{\eta_m(1-\eta_m)\}^{-1}$ results in the limiting form of a logit function $\psi_\infty(u) = \text{logit}(u)$. Generally, (13) implies that a different weighting scheme $\widetilde{w} = \widetilde{w}(\eta)$ will lead to a different transformation function,

$$\widetilde{\psi}^{w}_\infty(u) = \int_0^1 \{\mathbf{I}(u > 1-\eta) - \eta\}\widetilde{w}(\eta)d\eta, \quad (15)$$

provided that the integration exists for any u. In view of (15), it is equivalent to study the effect of the weight $\widetilde{w}(\eta)$ or that of the transformation function $\widetilde{\psi}^{w}_\infty(u)$. While we are not aware of any systematic investigation on the choice of $\widetilde{w}(\eta)$, which may be difficult because of the correlation among intervals, there exist a series of works on studying the impact of transformation functions in (2). For example, the double exponential function $\psi(u) = \{1 + \text{sgn}(u)(1-e^{-u})\}/2$ yields an efficient combination in Bahadur's sense (Singh et al., 2005). The normal distribution function $\psi(\cdot) = \Phi^{-1}(\cdot)$ yields an efficient combination in Fisher's sense when $H_k(\cdot)$s are approximately normal (Xie et al., 2011; Liu et al., 2014). These results can be used as guidelines for applications.

4. Case Studies

We consider two real data sets collected for safety analyses of two diabetes drugs. Both data sets have a considerable portion of trials where few events of interest are observed. We analyze the data sets using the methods presented in Sections 2 and 3. Since these methods can be unified under the CD framework, we implement the analysis using an R package gmeta, developed by Yang and Xie (2016) as a computing tool to perform the CD inference for meta-analysis.

For the rosiglitazone case introduced in Section 1, we consider its potential risk in increasing CVD-related mortality. For this endpoint, there are 25 studies (out of the total of 48) where no death was observed in either arm. Using the risk difference $(p_T - p_C)$ as the risk measure, we obtain the results shown in Table 2. The Mantel–Haenszel (MH) method shows a marginally significant result with a p-value of 0.05. This method relies on a large-sample theory and it effectively uses 23 studies (48%) in the analysis. In this case, the MH method may yield invalid inference outcomes (Tian et al., 2009; Liu et al., 2014). Unlike the MH method, the other three methods in Table 2 make use of an exact test (see, Tian et al., 2009). As a result, these methods can yield confidence intervals with a proper coverage probability and/or testing results with a proper type I error rate. Table 2 shows that both the traditional p-values combination methods (Fisher's and Stouffer's) yield a p-values of 1, which implies that these two methods may be overly conservative. For Tian et al.'s method, Table 2 shows that as the number of CI levels (M) increases, the combined CIs become narrower and the p-values smaller, which indicates that the inference becomes more efficient. Recall that when $M \to \infty$, Tian et al.'s method is equivalent to the CD combination (or Liu et al.'s method) with a logit transformation. Table 2 shows that the latter method improves inference efficiency, with a narrower CI $(-0.10\%, 0.20\%)$ and a smaller p-value 0.71. It also shows that other transformations (normal/double exponential) yield similar results, whereas the normal transformation seems slightly more efficient. Both Tian et al.'s and Liu et al.'s methods effectively use all available data (including the 25 studies where no event was observed). The two methods do not require continuity corrections to zero events, which can lead to severe inference bias (Sweeting et al., 2004; Tian et al., 2009; Liu et al., 2014).

Now, we consider safety analysis for a new diabetes drug in a class of medicines called SGLT2 inhibitors that work independently of insulin to help remove excess sugar from the

Table 2
Meta-analysis results for assessing rosiglitazone's risk in increasing CVD-related mortality (with risk difference used as the risk measure)

	95% CI (%)	*P*-value
Point estimates combination		
Mantel–Haenszel	(0.00, 0.21)	0.05
P-values combination		
Fisher's	—	1.00
Stouffer's	—	1.00
Tian et al.'s method of combining CIs		
M=5	(−0.26, 0.31)	0.96
M=10	(−0.14, 0.27)	0.85
M=20	(−0.14, 0.23)	0.81
Liu et al.'s method of combining p-value functions (CDs)		
$\psi(u) = \ln\{u/(1-u)\}$	(−0.10, 0.20)	0.71
$\psi(u) = \Phi(u)$	(−0.09, 0.20)	0.69
$\psi(u) = \{1 + \text{sgn}(u)(1 - e^{-u})\}/2$	(−0.11, 0.21)	0.77

body in adult patients. See Section 1.2 of FDA Briefing Book (2013) for its review history. Web Table 1 presents the safety data from 21 independent studies with respect to indexed bladder cancer. The events are rarer than the rosiglitazone example. Only 7 studies (33%) reported bladder cancer and no study observed more than 3 such events. Without patients' exposure time and time to event data, we cannot obtain the hazard ratio (HR) as FDA has done. Using the risk difference as the risk measure, we obtain the results in Web Table 2. The results are similar to those in the rosiglitazone case. Thus, a similar interpretation follows and detailed discussions are omitted. One caveat here is that we have implicitly assumed that the risk difference remains constant regardless of the varying exposure time in the 21 studies. This should not be overlooked in our interpretation.

5. Discussion

We have shown that for exact meta-analysis of rare events, the CD framework unifies Tian et al.'s method of combining confidence intervals and Liu et al.'s method of combining p-value functions. In view of the duality of confidence intervals and hypothesis testings, our results demonstrate that the two exact methods are equivalent in the sense as elaborated in Section 3. We show in the Web Appendix that the CD framework also subsumes the Mantel–Haenszel and Peto methods. The key is to define CD-like functions so that even if individual functions are not CD functions, the combined function can be proved to be a CD function. Thus far, we can conclude that the CD method offers a unifying framework for the meta-analysis of rare events.

The CD framework eliminates the need to select the number and levels of confidence intervals, as required by Tian et al.'s method. However, to implement the CD method, users still need to specify transformation functions in practice. As far as the length of confidence intervals (as a measure of efficiency) is concerned, we recommend using the normal distribution function $\psi(\cdot) = \Phi^{-1}(\cdot)$, based on the theoretical result summarized at the end of Section 3 and the numerical results in Section 4. The recommendation concerns the efficiency, rather than the validity of the CD combination. The CD method remains valid, regardless of the choice of transformation functions.

To achieve an overall inference, the CD framework requires users to specify an exact test (or a set of exact tests) for each individual study. The choice of exact tests depends on what risk measure is of interest. For the risk difference, Tian et al. (2009) used the mid-p adaption of Chan and Zhang (1999)'s method. For the odds ratio, Liu et al. (2014) used the mid-p adaption of Fisher's exact test. The mid-p method reduces the conservatism of individual exact tests and improves the efficiency of the overall inference. Although, theoretically, the mid-p method cannot guarantee that the inference has a proper type I error rate, it works well in our numerical studies. A known result is that if each individual test has a proper type I error rate, so does the overall test derived from the CD combination method (see Corollary 1 in Liu et al., 2014).

Meta-analyses are used in both exploratory (hypothesis-generating) analyses and confirmatory (hypothesis-testing) analyses in clinical trials (Berlin et al., 2013). For studying drug safety, exploratory analyses can be viewed as a discovery process in which decision makers are hoping to recognize safety issues as early as possible, such as in phase II studies during drug development. To achieve an early detection, we can perform a meta-analysis of the existing trials using the CD exact method. Since the concern may be better directed at trying not to miss a signal (Berlin et al., 2013; Thompson et al., 2011), we may consider more aggressive methods to mitigate the conservativeness of the exact test used in the CD method. An example is the so-called beta adjustment proposed by Liu et al. (2014). We may also explore the data using different types of exact tests or risk measures to maximize the likelihood of discovery (Thompson et al., 2011). In contrast, the choice of analyses is different in confirmatory analyses where decision makers have a pre-specified hypothesis prior to the trials and the goal is to confirm the hypothesized association. In this situation, a priori analysis plan is essential (Thompson et al., 2011). To use the CD method, we may specify in advance what risk measure/exact test to use and whether or not to make an adjustment to the exact test. Once the data from confirmatory trials are available, we

may use the CD method to meta-analyze the new data, or more contentiously, with the data used in exploratory analyses to capture the totality of evidence. We refer readers to Thompson et al. (2011) and Berlin et al. (2013) for practical issues in meta-analyses of clinical trials.

6. Supplementary Materials

Web Appendix, referenced in Sections 1, 3.2, 3.3, 4, and 5, is available with this article at the *Biometrics* website on Wiley Online Library.

Acknowledgements

The authors are grateful to the Editor, the Associate Editor and two anonymous reviewers for their comments and suggestions, which led to a significant improvement of the presentation. The authors acknowledge Pixu Shi for her discussion and assistance in the early phase of the research. The research is supported in part by the research fund for new faculty at Lindner College of Business (Liu) and NSF grants DMS1513483 and DMS1107012 (Xie).

References

Barnes, J. (2007). Drug safety. *Nature Reviews Drug Discovery* **6**, 937.

Berlin, J. A., Crowe, B. J., Whalen, E., Xia, H. A., Koro, C. E., and Kuebler, J. (2013). Meta-analysis of clinical trial safety data in a drug development program: Answers to frequently asked questions. *Clinical Trials* **10**, 20–31.

Chan, I. S. and Zhang, Z. (1999). Test-based exact confidence intervals for the difference of two binomial proportions. *Biometrics* **55**, 1202–1209.

Claggett, B., Xie, M., and Tian, L. (2014). Meta analysis with fixed, unknown, study-specific parameters. *Journal of the American Statistical Association* **109**, 1660–1671.

Cox, D. (2013). Discussion of "Confidence distribution, the frequentist distribution estimator of a parameter — A Review". *International Statistical Review* **81**, 40–41.

Cox, D. R. (1958). Some problems connected with statistical inference. *The Annals of Mathematical Statistics* **29**, 357–372.

Efron, B. (1993). Bayes and likelihood calculations from confidence intervals. *Biometrika* **80**, 3–26.

Finkelstein, M. O. and Levin, B. (2012). Meta-analysis of sparse data: Perspectives from the avandia cases. *Jurimetrics Journal* **52**, 123–153.

Fisher, R. (1932). *Statistical Methods for Research Workers*. London: Oliver and Boyd, 4th edition.

Fraser, D. A. S. (1991). Statistical inference: Likelihood to significance. *Journal of the American Statistical Association* **86**, 258–265.

Liu, D., Liu, R., and Xie, M. (2014). Exact meta-analysis approach for discrete data and its application to 2 × 2 tables with rare events. *Journal of the American Statistical Association* **109**, 1450–1465.

Liu, D., Liu, R. Y., and Xie, M. (2015). Multivariate meta-analysis of heterogeneous studies using only summary statistics: Efficiency and robustness. *Journal of the American Statistical Association* **110**, 326–340.

Marden, J. I. (1991). Sensitive and sturdy p-values. *The Annals of Statistics* **19**, 918–934.

Nissen, S. E. and Wolski, K. (2007). Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *New England Journal of Medicine* **356**, 2457–2471.

Normand, S. (1999). Tutorial in biostatistics meta-analysis: Formulating, evaluating, combining, and reporting. *Statistics in Medicine* **18**, 321–359.

Rücker, G., Schwarzer, G., Carpenter, J., and Olkin, I. (2009). Why add anything to nothing? the arcsine difference as a measure of treatment effect in meta-analysis with zero cells. *Statistics in Medicine* **28**, 721–738.

Singh, K., Xie, M., and Strawderman, W. E. (2005). Combining information from independent sources through confidence distributions. *The Annals of Statistics* **33**, 159–183.

Stouffer, S. A., DeVinney, L. C., and Suchmen, E. A. (1949). *Adjustment During Army Life*, volume 1. Princeton: Princeton University Press.

Sweeting, M., Sutton, A., and Lambert, P. (2004). What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in Medicine* **23**, 1351–1375.

Thompson, J. R., Attia, J., and Minelli, C. (2011). The meta-analysis of genome-wide association studies. *Briefings in Bioinformatics* **12**, 259–269.

Tian, L., Cai, T., Pfeffer, M. A., Piankov, N., Cremieux, P.-Y., and Wei, L. (2009). Exact and efficient inference procedure for meta-analysis and its application to the analysis of independent 2× 2 tables with all available data but without artificial continuity correction. *Biostatistics* **10**, 275–281.

Xie, M., Liu, R. Y., Damaraju, C., and Olson, W. H. (2013). Incorporating external information in analyses of clinical trials with binary outcomes. *The Annals of Applied Statistics* **7**, 342–368.

Xie, M. and Singh, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: A review (with discussion). *International Statistical Review* **81**, 2–39.

Xie, M., Singh, K., and Strawderman, W. (2011). Confidence distributions and a unifying framework for meta-analysis. *Journal of the American Statistical Association* **106**, 320–333.

Yang, G., Liu, D., Liu, R. Y., Xie, M., and Hoaglin, D. C. (2014). Efficient network meta-analysis: A confidence distribution approach. *Statistical Methodology* **20**, 105–125.

Yang, G. and Xie, M. (2016). gmeta: Meta-Analysis via a Unified Framework of Confidence Distribution. R package version 2.2-4. http://CRAN.R-project.org/package=gmeta.

Received April 2015. *Revised December* 2015.
Accepted January 2016.