Exact meta-analysis approach for discrete data and its application to 2×2 tables with rare events

Dungang Liu, Regina Y. Liu and Minge Xie[1]

[1] Dungang Liu is Postdoctoral Associate, Department of Biostatistics, Yale University School of Public Health, New Haven, CT 06520 (E-mail: dungang.liu@yale.edu). Regina Y. Liu and Minge Xie are Professors, Department of Statistics and Biostatistics, Rutgers University, Piscataway, NJ 08854 (E-mails: rliu@stat.rutgers.edu and mxie@stat.rutgers.edu). This work is based on part of the first author's Ph.D. dissertation. The authors are thankful to the Associate Editor and two anonymous referees for their suggestions and comments that led to significant improvement of the paper. The first author acknowledges the Graduate Fellowship from Rutgers University and support from NSF grant DMS 0707053 and NIH grant R01 DA016750. The second and third authors acknowledge support from grants: NSF-DMS 1107012; 1007683; 0915139; 0707053, NSF-SES 0851521, and NSA-H98230:08-1-0104; 11-1-0157.

Exact meta-analysis approach for discrete data and its application to 2×2 tables with rare events

Abstract

This paper proposes a general exact meta-analysis approach for synthesizing inferences from multiple studies of discrete data. The approach combines the *p-value functions* (also known as *significance functions*) associated with the exact tests from individual studies. It encompasses a broad class of exact meta-analysis methods, as it permits broad choices for the combining elements, such as tests used in individual studies, and any parameter of interest. The approach yields statements that explicitly account for the impact of individual studies on the overall inference, in terms of efficiency/power and the type I error rate. Those statements also give rises to empirical methods for further enhancing the combined inference. Although the proposed approach is for general discrete settings, for convenience, it is illustrated throughout using the setting of meta-analysis of multiple 2×2 tables. In the context of rare events data, such as observing few, zero or zero total (i.e., zero events in both arms) outcomes in binomial trials or 2×2 tables, most existing meta-analysis methods rely on the large-sample approximations which may yield invalid inference. The commonly used corrections to zero outcomes in rare events data, aiming to improve numerical performance can also incur undesirable consequences. The proposed approach applies readily to any rare event setting, including even the zero total event studies without any artificial correction. While debates continue on whether or how zero total event studies should be incorporated in meta-analysis, the proposed approach has the advantage of automatically including those studies and thus making use of all available data. Through numerical studies in rare events settings, the proposed exact approach is shown to be efficient and, generally, outperform commonly used meta-analysis methods, including Mental-Haenszel and Peto methods.

Key Words:: combining information; confidence distribution; exact inference; p-value function; zero events.

1 Introduction

Meta-analysis is perhaps the most used methodology to synthesize findings from independent studies. It has been used extensively in many fields, and it has a rich literature (see, for example, the reviews by Sutton and Higgins (2008) and Finkelstein and Levin (2012), and the references therein). However, several important issues in meta-analysis remain unaddressed and new challenges continue to emerge as we have access to more diverse sources of data nowadays. For example, in discrete data settings, most existing meta-analysis methods rely on the large-sample approximations which may not necessarily produce valid inference. In case of rare event data such as observing few, zero or zero total (i.e., zero events in both arms) outcomes in binomial trials, it is also a common practice to apply correction of a constant to zero outcomes and/or remove zero total outcomes from the analysis. Either approach, however, may incur undesirable consequences to the final inference.

The goal of this paper to to propose a general and exact meta-analysis approach for combining the inferences from multiple studies from discrete settings. The approach combines the *p-value functions* (also known as significance functions) associated with the exact tests from individual studies. The simple structure of the combining formula in the proposed approach allows us to obtain mathematical expressions to explicitly account for the impact of individual studies on the overall combined inference in terms of efficiency/power and the type I error rate. Using those expressions, we can also develop and implement empirical methods to further enhance the combined inference to ensure certain desired levels of efficiency and accuracy. The accuracy here refers to the deviation between the achieved type I error rate and the nominal one.

Although the proposed approach is developed for general settings, much of the discussion focuses on the particular setting of discrete and rare events data. Such discrete and rare data are often seen in either small-sample survey studies or large-sample clinical trials with very low event rates, such as safety analysis in drug development (e.g., Nissen and Wolski, 2007). In case of rare events, a single study is inadequate for drawing a reliable conclusion, but the conclusion can often be strengthened by using meta-analysis to synthesize conclusions from a number of similar studies. The analysis of rare events data raises special statistical challenges, and has been intensely studied (Sweeting, Sutton, and Lambert, 2004; Bradburn et al., 2007; Finkelstein and Levin, 2012), and new methodological developments continue to emerge (Tian et al., 2009; Cai, Parast, and Ryan, 2010; Bhaumik et al., 2012). So far, most of the commonly used meta-analysis methods rely on the asymptotic distribution of the combined estimator to make inference. As concrete examples, the widely used inverse-variance weighted method combines point estimators from individual studies, assuming that the distributions of all the estimators can be well approximated by normal distributions. For discrete data, the well-known Mantel-Haenszel and Peto methods also rely on the normal approximation to the distribution of the combined estimator. However, it is known that the normal approximation is ill-suited for data sets which are discrete with small sample sizes. In particular, in the rare events setting, such as observing 1 or even 0 outcome out of 100 Bernoulli experiments, the normal approximation to the test statistic may yield an unacceptably low coverage probability of confidence intervals (Bradburn et al., 2007; Tian et al., 2009). Furthermore, the commonly used 0.5 correction (known as Haldane's bias correction) to zero events, which aims to improve the approximation, is shown with compelling evidence to have undesirable impact on inference outcomes, such as yielding a severe bias (Sweeting et al., 2004; Bradburn et al., 2007). All these shortcomings clearly show the need for the exact approach, rather than those using limiting distribution, to the analysis of discrete data, especially when the events of interest are rare. Here, similar to Agresti (2007), the term "exact" refers to using exact distributions in the inference, rather than achieving exact test size or coverage levels.

In this paper, instead of working on point estimators, we develop a new meta-analysis approach by combining functions, more specifically the *p-value functions* (also known as *significance functions*; cf. Fraser, 1991) obtained from the exact tests associated with individual studies. We show that such an approach can yield a broad class of methods for combining exact inference, and it subsumes as special cases all the existing p-value combination methods and the confidence interval combination method (i.e., Tian et al., 2009). Moreover, the idea of combining p-value functions applies to inference for any parameter of interest, including the odds ratio, risk ratio or risk difference in the analysis of 2×2 tables. This paper justifies the validity of the proposed approach, and demonstrates, by using the setting of multiple 2×2 tables with a common odds ratio as a working example, that the proposed approach applies easily and performs well even to difficult situations even when confronting zero total event studies. To sum up, the proposed approach applies readily to any rare event setting, including even the zero total event studies without any artificial correction. Overall, our proposed approach compares favorably to the existing methods, including Mental-Haenszel and Peto methods.

The rest of this paper is organized as follows. In Section 2, we present our p-value function combination approach as a general methodology for combining exact inference from independent studies. In Section 3, we use the combining formula to establish theoretical results to decompose the type I error rate as well as power/efficiency in the combined inference into the individual input from each study. This is established for both small-sample and large-sample settings. In Section 4, we present some empirical methods for improving the overall efficiency and accuracy, by choosing suitable weights and adjusting individual p-value functions. In Section 5, we present numerical studies on the performance of our approach and compare it with some commonly used existing approaches in the rare events setting. Finally, we present a discussion in Section 6 to address issues regarding the choices of the combining elements in our approach. In particular, we provide a general guideline in terms of steps to implement our approach. We also elaborate on the handling of rare events, and emphasize that our approach automatically incorporates all the available data in the analysis without requiring corrections to zero events.

2 Methodology

2.1 Problem setup

Assume that K ($K > 1$) independent studies are conducted to examine the same parameter ψ, and that a random sample is collected in each study for testing the hypotheses

$$H_0 : \psi = \psi^* \text{ versus } H_1 : \psi > \psi^*, \tag{1}$$

where ψ^* is an arbitrary but fixed value in the parameter space. The goal is to develop a general approach to combine the K individual test results from (1) to make exact and efficient inference for the K studies as a whole.

Our proposed approach is developed for combining testing results from multiple studies for all discrete data settings, but, for convenience, it is illustrated in this paper through the working example of combining multiple binary comparison studies (i.e., 2×2 tables). Specifically, consider K independent 2×2 tables formed by pairs of independent binomial random variables (X_i, Y_i) with sample sizes (n_i, m_i) and their associated event rates (π_{1i}, π_{0i}), for $i = 1, \ldots, K$. Assume that the task is, based on the observed numbers of events x_i and y_i, to make inference about the *effect measure*, such as *odds ratio*, *risk ratio*, or *risk difference* which are defined respectively as

$$\text{OR}_i \equiv \frac{\pi_{1i}/(1 - \pi_{1i})}{\pi_{0i}/(1 - \pi_{0i})}, \quad \text{RR}_i \equiv \frac{\pi_{1i}}{\pi_{0i}}, \quad \text{RD}_i \equiv \pi_{1i} - \pi_{0i}, \quad i = 1, \ldots, K.$$

Under a fixed effects model, it is often assumed that an effect measure has a common value across all the studies. Making inference on this common parameter, denoted by ψ as seen in (1) and throughout the paper, is often of importance. For example, Sweeting et al. (2004) reviewed meta-analysis methods for the common odds ratio $\text{OR}_i = \psi$, and Bradburn et al. (2007) compared methods for the common risk difference $\text{RD}_i = \psi$. The approach we develop in this paper provides a general meta-analysis procedure for making exact inference for ψ where ψ can be OR, RR, RD, any risk measure, or any common parameter in any discrete models.

2.2 The proposed exact meta-analysis approach

Before describing the proposed approach, we first describe the main tool we used in the approach, namely the p-value function.

A p-value function, is formed by computing p-values for a one-sided test with varying boundaries of the null hypothesis (e.g., Fraser, 1991). In the context of problem setting above, we let $z = (x, y)$ denote the sample, and $p = p(\psi^*; z)$ denote a p-value computed based on a given test for testing $H_0 : \psi = \psi^*$ versus $H_1 : \psi > \psi^*$. The p-value $p = p(\psi^*; z)$ depends on both the sample z and the value of ψ^*. Given the sample z, as the value of ψ^* varies, $p(\psi^*) \equiv p(\psi^*; z)$ is a function on the parameter space of ψ. This sample-dependent function $p(\cdot)$ is called a p-value function. Under some mild conditions, a p-value function is typically a distribution function on the parameter space. From the viewpoint of confidence distributions (see a review by Xie and Singh (2013) and the references therein), this p-value function is often viewed as a "distribution estimator" of the unknown parameter, in the sense that a sample-dependent distribution function, rather than a point or an interval, is used to estimate the parameter. The "distribution estimator" carries much more information than a point or an interval estimator, such as skewness of the exact distribution of statistics. Therefore, the p-value function seems to be an ideal device for combining exact inference from multiple studies.

To make exact inference on the common parameter ψ for the combined inference, we begin by carrying out for each study an exact test for the hypotheses $H_0 : \psi = \psi^*$ versus $H_1 : \psi > \psi^*$. We denote by $p_i(\psi^*; x_i, y_i)$ the p-value obtained from the test in the i-th study and $p_i(\cdot) \equiv p_i(\cdot; x_i, y_i)$ the corresponding p-value function. Figure 1 shows a p-value function (in a black solid curve) on testing the odds ratio from a study that observes $(x_i, y_i) = (1, 3)$ with the sample sizes $(n_i, m_i) = (15, 60)$. As a result from an exact test, this p-value function preserves all the intrinsic finite-sample properties of the test. In particular, it preserves the possible asymmetry from the distribution, as opposed to the common approaches from normal approximations which automatically result in symmetric normal-based p-value functions. It is also worth noting that such a p-value function $p_i(\cdot; x_i, y_i)$ is always obtainable regardless of whether the entries x_i and y_i are zeros or not, see some examples of $p_i(\cdot; x_i, y_i)$ with different settings of x_i and y_i in Figure 2 (a)-(d). This is a desirable feature, especially in the rare events setting, since it allows our approach to combine all studies from their corresponding p-value functions, even p-value functions from the zero total event studies with $(x_i, y_i) = (0, 0)$. This is not the case in many commonly used meta-analysis methods in practice. The issues related to zero total event studies are discussed further in Section 6.

After obtaining the p-value functions from all K studies, namely, $\{p_i(\cdot; x_i, y_i), i = 1, \cdots, K\}$, we proceed to combine them using the following recipe. Let $p_i(\cdot) \equiv p_i(\cdot; x_i, y_i)$ and $p_{(c)}(\cdot)$ be the overall combined p-value function, then

$$p_{(c)}(\psi) \equiv F_{(c)} \left[w_1 h(p_1(\psi)) + \cdots + w_K h(p_K(\psi)) \right]. \tag{2}$$

Here, $h(\cdot)$ is a "transformation function" which can be any monotonically increasing function, and $F_{(c)}(\cdot) = h^{-1}(\cdot/w_1) * \cdots * h^{-1}(\cdot/w_k)$, where $*$ stands for convolution. Throughout this paper we

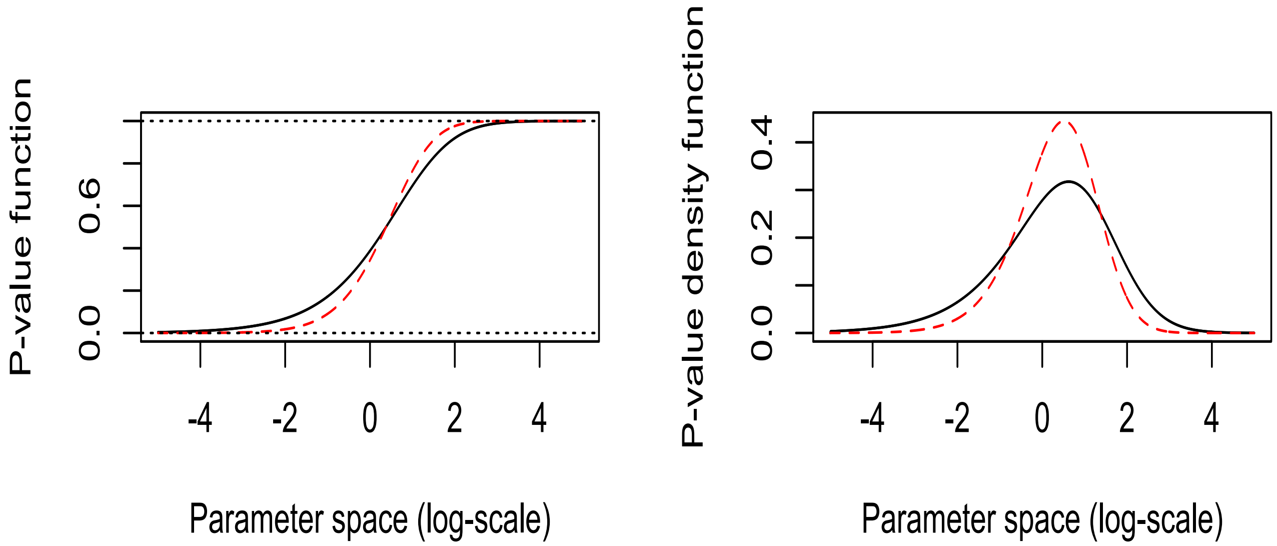

Figure 1: Illustrations of an individual p-value function (black solid curve) and a combined p-value function (red dashed curve). The individual p-value function is obtained using the mid-p adaptation of Fisher exact test on odds ratio in a study that observes $x_i = 1$ and $y_i = 3$ with sample sizes $(n_i, m_i) = (15, 60)$. The combined p-value function is from combining two independent copies of the individual p-value function, assuming that the two studies happen to have the same observations. The x-axes are in the logarithm scale.

use $\Phi(\cdot)$, the cumulative distribution function of the standard normal distribution, as the inverse transformation function $h^{-1}(\cdot)$ and thus $F_{(c)}(\cdot) = \Phi(\cdot/w_1) * \cdots * \Phi(\cdot/w_k) = \Phi(\cdot/(\sum_{i=1}^{K} w_i^2)^{1/2})$ and w_i's are weights subject to $\sum_{i=1}^{K} w_i = 1$. More discussion on the choice of the transformation function $h(\cdot)$ is in Section 6. The combining recipe (2) has also been proposed in Xie et al. (2011) for continuous or large sample settings. Note that Stouffer's method (Stouffer et al., 1949) can be viewed as a special case of (2) with weights $w_i \equiv 1$ for all i. As noted in Xie et al. (2011), the use of non-trivial weights makes the recipe (2) more advantageous than the traditional p-value combination approaches such as Stouffer's method. We show in this paper that using the more informative weights, our approach can achieve asymptotic efficiency in the overall inference as well as enhance finite sample efficiency. The choice of weights is discussed in detail in Section 4.1.

In Figure 1 the red dashed curve is the combined p-value function $p_{(c)}(\cdot)$ that combines two independent copies of the individual p-value function (plotted in a solid curve), assuming that the two studies happen to have the same observations. Figure 1 shows that the combined p-value function curve retains the skewness in the individual p-value function curves. Additional examples

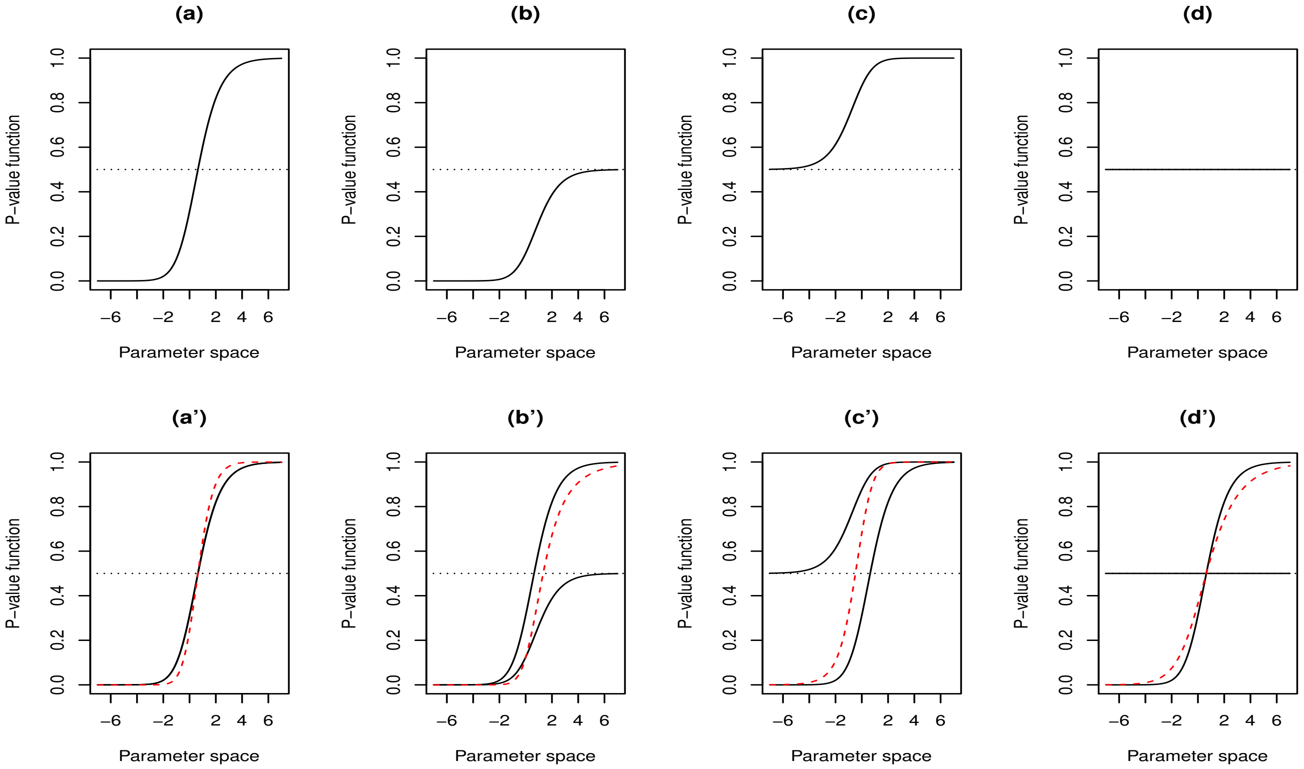

Figure 2: Illustrations of individual p-value functions (upper row) and combined p-value functions (lower row). The individual p-value functions (black solid curves) in the upper row are for the cases: (a) $x_i = 2, y_i = 1$; (b) $x_i = 2, y_i = 0$; (c) $x_i = 0, y_i = 2$; and (d) $x_i = 0, y_i = 0$, all with the same sample sizes $n_i = m_i = 100$. These functions are obtained using mid-p adaption of Fisher's exact test. The combined p-value functions (red dashed curves) shown in the lower row are from combining the p-value function in (a) with the individual p-value function in (a), (b), (c) and (d) respectively, using equal weights. The x-axes are in the logarithm scale.

of the combined p-value functions are the red dashed curves in Figure 2 (a')-(d'), where in (b')-(d') at least one entries of x_i and y_i are zeros in one of the individual p-value functions. Note that the p-value functions in Figure 2 are obtained using the mid-p of Fishers exact test, and that different tests may yield different p-value functions. In any case, once the combined p-value function $p_{(c)}(\cdot)$ is obtained, it is ready for the overall inference for ψ. More specifically, $p_{(c)}(\psi)$, as a function on the parameter space, can be used to derive all forms of inference outcomes. For instance, $p_{(c)}(\psi^*)$ can naturally be used as the overall p-value for testing hypothesis (1), the median of $p_{(c)}(\cdot)$ (i.e., $M_n \equiv p_{(c)}^{-1}(1/2)$) as a point estimator for ψ, and the intervals $(p_{(c)}^{-1}(\alpha), \infty)$ and $(p_{(c)}^{-1}(\alpha/2), p_{(c)}^{-1}(1-\alpha/2))$ as a $100(1-\alpha)\%$ one-sided and two-sided confidence intervals for ψ, respectively. Here $p_{(c)}^{-1}(\cdot)$ is the inverse function of $p_{(c)}(\cdot)$ (Its corresponding upper or lower limit versions are used in the case when $p_{(c)}(\cdot)$ is not continuous).

The remainder of this paper is devoted to justifying the validity of using the combined p-value function $p_{(c)}(\cdot)$ for the exact overall inference. In particular, it aims to achieve the following two distinct but closely related goals.

First, we show that the idea of combining "functions" proposed in Singh et al. (2005) is valid in case of small-sample and discrete data analysis. For continuous or large-sample data, Xie et al. (2011) employed the same idea in which the elements for combining are required to be (or at least asymptotically) *confidence distribution functions*. In the context of this paper, this is equivalent to requiring that the statistic $p_i(\psi_0) \equiv p_i(\psi_0; X_i, Y_i)$ follow (or at least asymptotically follow) U(0,1) distribution, i.e., $\Pr(p_i(\psi_0) \le s) = s$, for any $0 \le s \le 1$, where ψ_0 is the true value of ψ. However, such a requirement is not fulfilled in the discrete setting, especially with the case of small samples. In fact, in the analysis of rare events data with zero events, the deviation of the distribution of $p_i(\psi_0)$ from U(0,1) distribution is often non-negligible. It measures the difference between the achieved and the nominal type I error rates, and thus the loss of accuracy in inference. This deviation is due to the intrinsic discrete nature of the underlying distribution, and, for certain exact tests where p-values are derived by maximizing over the nuisance parameters, it may not even diminish asymptotically. We show in Section 3 that the simple combining formula in (2) enables us to derive an explicit expression (cf. Theorem 1) for assessing precisely this deviation and the loss of accuracy in the overall inference. This loss of accuracy has been difficult to ascertain when other approaches of combining inferences are used.

Second, we show that the idea of combining p-value functions, as a general approach for combining exact inferences, allows a variety of choices of tests, weights, transformation functions ($h(\cdot)$ in (2)), and thus yields a broad class of exact methods for combining inferences. For example, it subsumes the traditional p-value combination approaches. It has also been shown that the interval combination method proposed in Tian et al. (2009) is a special case of ours by using essentially a logistic function as the transformation function $h(\cdot)$ in the combining formula (2) (Yang et al., 2012). In Section 4.1, we examine the effect of the choice of weights on the combined inference. In Section 6, we discuss some guidelines for narrowing down choices for tests, weights, "transformation" functions and other elements for the combining procedure. If ever a particular choice is required or preferred, we can assess this particular choice using the formulas in Section 3. Moreover, if it is desired to improve the efficiency of the overall inference, we can further apply the adjustment scheme on the individual p-value functions developed in Section 4.2 to achieve the goal.

3 Theoretical properties

In this section, we show how our combining formula (2) with its simple structure enables us to derive theoretical properties of the combined inference for both small-sample and large-sample settings. Specifically, we establish an explicit formula in Theorem 1 that decomposes the achieved type I error of the combined test into those from the individual studies. This and other properties on the power and efficiency of the combined inference in turn can be used develop empirical methods to further improve the power/efficiency for the combined inference in the rare events setting. This latter point will be elaborated in Section 4.

3.1 Small-sample properties

For a study that involves discrete data, the discrete nature of the underlying distribution and the possible presence of unknown nuisance parameters generally prevent any exact test from achieving the nominal type I error rate. As a matter of fact, in the rare events setting the achieved type I error rate may be far below the nominal one. In other words, the p-value function $p_i(\psi)$ for an individual study, when assuming its value at $\psi = \psi_0$ and as a function of random sample (X_i, Y_i), may not follow exactly the U(0,1) distribution. In contrast to the combining approach in Xie et al. (2011) which require $p_i(\psi_0) \sim$ U(0,1) for continuous data and large-sample inference, we justify here that our approach of combining p-value functions, despite of their not following U(0,1), is valid for discrete data and small-sample inference.

In what follows, we examine the combined effect of our approach with respect to both the type I error and testing power. To present the results under a proper framework, we define a *power function* for our proposed approach for any fixed ψ^* as:

$$R_{(c)}(s; \psi^*) \equiv \Pr\{p_{(c)}(\psi^*) \leq s\}, \quad \text{for } \; 0 \leq s \leq 1, \qquad (3)$$

which is the cumulative distribution function of the statistic $p_{(c)}(\psi^*)$ and $\Pr(\cdot)$ is the probability evaluated under the true model. At $s = \alpha$ and when $\psi_0 > \psi^*$, $R_{(c)}(\alpha; \psi^*)$ gives the power of our approach for test (1). Similarly, we define power functions for the individual tests as $R_i(s;\psi^*) \equiv \Pr\{p_i(\psi^*) \leq s\}$ for all i.

When $\psi^* = \psi_0$, the power function in (3) becomes a *type I error rate function*, since $R_{(c)}(\alpha;\psi_0)$ gives the type I error rate of our approach for the hypothesis testing in (1). The theorem below shows that the achieved type I error rate function of our overall test $R_{(c)}(s;\psi_0)$ can be expressed explicitly in terms of the individual type I error rates functions $R_i(s;\psi_0)$, $i = 1,\ldots,K$.

Theorem 1. *The overall type I error rate function $R_{(c)}(s;\psi_0)$ can be expressed as*

$$R_{(c)}(s;\psi_0) = s + \sum_{i=1}^{K} d_i(s), \tag{4}$$

where

$$d_i(s) = E\left(D_i\left[\Phi\left\{\left(1+\sum_{j\neq i}\frac{w_j^2}{w_i^2}\right)^{1/2}\Phi^{-1}(s) - \sum_{j\neq i}\frac{w_j}{w_i}\Phi^{-1}(B_{ij})\right\}\right]\right). \tag{5}$$

Here, the functions $D_i(s) \equiv R_i(s;\psi_0) - s$, and the expectation E is taken with respect to the random variables B_{ij} which are independent and of the following distributions: for any $0 \leq t \leq 1$, $\Pr(B_{ij} \leq t) = t$ if $j \leq i$, and $\Pr(B_{ij} \leq t) = R_i(t;\psi_0)$ if $j > i$.

Theorem 1 immediately yields the following corollary which shows that the overall deviation of the type I error rate $\{R_{(c)}(s;\psi_0) - s\}$ can be bounded using the bounds of the individual deviations $\{R_i(s;\psi_0) - s\}$, $i = 1,\ldots,K$. As a special case, Corollary 1 suggests that, if the test of each study has a deflated (inflated) type I error rate, so will the overall test.

Corollary 1. *Suppose there exist fixed lower and upper bounds l_i and u_i such that*

$$l_i \leq R_i(s;\psi_0) - s \leq u_i, \quad i = 1,\ldots,K,$$

for any $0 \leq s \leq 1$. Then the overall type I error rate function $R_{(c)}(s;\psi_0)$ satisfies

$$\sum_{i=1}^{K} l_i \leq R_{(c)}(s;\psi_0) - s \leq \sum_{i=1}^{K} u_i, \tag{6}$$

for any $0 \leq s \leq 1$. Specifically, if $R_i(s;\psi_0) \leq s$ for all i and $0 \leq s \leq 1$, then $R_{(c)}(s;\psi_0) \leq s$. Similarly, if $R_i(s;\psi_0) \geq s$ for all i and $0 \leq s \leq 1$, then $R_{(c)}(s;\psi_0) \geq s$.

The results in Theorem 1 and Corollary 1 can be used to evaluate the effect of our combination in terms of the type I error. Specifically, Theorem 1 shows that the type I error rate of the overall test can be traced down to the individual ones. Hence, if we can evaluate the functions $R_i(s;\psi_0)$'s, then $R_{(c)}(s;\psi_0)$ can be evaluated from Theorem 1. In Appendix Part I, we propose an empirical method for estimating π_{1i} and π_{0i}, which allows us to evaluate $R_i(s;\psi_0)$, and then following Theorem 1 to evaluation $R_{(c)}(s;\psi_0)$. In case only the bounds for $\{R_i(s)-s\}$ are available for all i, the bounds for $\{R_{(c)}(s;\psi_0)-s\}$ can be derived from Corollary 1. Furthermore, the results in Theorem 1 and Corollary 1 also imply that we can improve the overall test by raising the type I error rates of the individual tests, if they are deflated, to be closer to the nominal level. This adjustment helps mend the possible inaccuracy caused by individual $p_i(\psi_0)$ not following closely U(0,1) in some studies, and it is discussed in details in Section 4.2. In summary, once the properties of the type I error of the individual tests are known, the corresponding properties of the type I error of the combined test can be derived accordingly.

When the alternative hypothesis in (1) holds (i.e., $\psi_0 > \psi^*$), the power function $R_{(c)}(s;\psi^*)$ gives the power of the combined test for rejecting the null hypothesis. Next theorem shows that a lower bound for the overall power function $R_{(c)}(s;\psi^*)$ can be derived from the lower bounds for the individual power function $R_i(s;\psi^*)$. In the theorem and also the later discussion, $N_i \equiv N_i(n_i,m_i)$ is a function of n_i and m_i, which may be views as a generic sample size of the i-th study; depending on the problem under consideration, N_i may assume n_i+m_i, $\min\{n_i,m_i\}$ or $1\big/\big[\{n_i\pi_{1i}(1-\pi_{1i})\}^{-1}+\{m_i\pi_{0i}(1-\pi_{0i})\}^{-1}\big]$, etc.

Theorem 2. *For a fixed $\psi^* < \psi_0$, assume that there exist a positive constant (maybe a function of N_i) $a_i > 0$ such that the individual power function $R_i(s;\psi^*)$ has the lower bound:*

$$R_i(s;\psi^*) = \Pr\{p_i(\psi^*) \le s\} \ge 1-(1-s)/a_i, \quad i=1,\ldots,K, \tag{7}$$

Then, the overall power function $R_{(c)}(s;\psi^*)$ has the following lower bound:

$$R_{(c)}(s;\psi^*) = \Pr\{p_{(c)}(\psi^*) \le s\} \ge 1-(1-s)\bigg/\bigg\{\prod_{i=1}^{K} a_i\bigg\}. \tag{8}$$

In the special case of $a_i \equiv 1$ for all i, Theorem 2 implies that if the individual p-value $p_i(\psi^*)$ is stochastically less than a U(0,1) distributed random variable (i.e., $R_i(s;\psi^*) \ge s$) for all i, so will the combined p-value $p_{(c)}(\psi^*)$. Theorem 2 also implies that combining test results from independent studies may lead to significant gain in power for the overall inference. Note that when the alternative hypothesis in (1) holds (i.e., $\psi_0 > \psi^*$), the power of an individual test, measured by $R_i(s;\psi^*)$, typically approaches 1 as the sample sizes n_i and m_i increase (i.e., as $N_i \to \infty$). Thus, let us assume the difference $\{1-R_i(s;\psi^*)\}$ is bounded by $(1-s)/a_i$, for some $a_i = O(N_i^c)$, $c > 0$. In this case, Theorem 2 suggests that the difference $\{1-R_{(c)}(s;\psi^*)\}$ for our combined test is bounded by $(1-s)/\{\prod_{i=1}^K a_i\}$. Since $\prod_{i=1}^K a_i$ can be much greater than any individual a_i, the lower bound in (8) for the overall power function $R_{(c)}(s;\psi^*)$ can be much higher than the lower bound in (7) for any individual power function $R_i(s;\psi^*)$.

3.2 Large-sample properties

Although our combining method is developed mainly for exact inference in the rare events or small sample setting, it also applies to the general meta-analysis setting where large-sample approximations may be reasonable. We provide the large-sample properties and theoretical justification for our proposal in a large sample setting when it is applicable. More importantly, the asymptotic results here help develop useful guidelines for choosing proper weights in the rare events setting where exact inference is desired, see Section 4.1.

For the combined test, we first consider the limiting type I error rate, defined as

$$R^L_{(c)}(s;\psi_0) \equiv \lim_{\substack{n_i,m_i\to\infty\\ i=1,\ldots,K}} R_{(c)}(s;\psi_0) = \lim_{\substack{n_i,m_i\to\infty\\ i=1,\ldots,K}} \Pr\{p_{(c)}(\psi_0) \le s\}.$$

Similar to Theorem 1, we can show that $R^L_{(c)}(s;\psi_0)$ can be expressed explicitly in terms of the limiting type I error rate functions for individual tests $R^L_i(s;\psi_0) \equiv \lim_{n_i,m_i\to\infty} R_i(s;\psi_0) = \lim_{n_i,m_i\to\infty}\Pr\{p_i(\psi_0) \le s\}$, $i = 1,\ldots,K$.

Theorem 3. *The statement of Theorem 1 holds with the following modifications: replacing the overall type I error function $R_{(c)}(s;\psi_0)$ by its limiting form $R^L_{(c)}(s;\psi_0)$ and replacing the individual type I error functions $R_i(s;\psi_0)$ by their corresponding limiting forms $R^L_i(s;\psi_0), i = 1,\ldots,K$.*

As a direct consequence of Theorem 3, Corollary 2 below states that, if the individual p-value function $p_i(\psi)$ yields no loss of inference accuracy (i.e., $p_i(\psi_0) \sim$ U(0,1)) asymptotically for all i, then the same holds for the combined p-value function $p_{(c)}(\psi)$. In this case, the test or the confidence interval derived from $p_{(c)}(\psi)$ achieves, respectively, the nominal type I error rate and coverage probability asymptotically.

Corollary 2. *If the individual limiting type I error rate functions $R^L_i(s;\psi_0) \equiv s$ for all $i, 1 \le i \le K$ and $0 \le s \le 1$, then the overall limiting type I error rate function $R^L_{(c)}(s;\psi_0) \equiv s$.*

It is worth pointing out that Theorem 3 also holds for general asymptotic settings without requiring $R^L_i(s;\psi_0) \equiv s$. This property is useful, considering the fact that the distribution of a p-value may not follow an U(0,1) distribution under the null hypothesis, even asymptotically. This can be the case when unknown nuisance parameters are present, as observed in (Robins, van der Vaart, and Ventura, 2000), which is also the case often seen in the analysis of 2×2 tables.

Consider the limiting power/asymptotic efficiency of the combined test. Theorem 4 below states that, if the exact test associated with each study is asymptotically equivalent to Wald test (see, e.g., Fraser, 1991), then our combining approach can achieve asymptotic efficiency with suitably chosen weights.

Theorem 4. *Suppose that the p-value function $p_i(\psi)$ obtained from the exact test associated with the i-th study can be expressed as*

$$p_i(\psi) = \Phi\left[(\psi - \hat{\psi}_{i,MLE}) \Big/ \left\{\widehat{aVar(\hat{\psi}_{i,MLE})}\right\}^{1/2}\right] + o_p(1), \quad i = 1, \ldots, K, \tag{9}$$

where $\hat{\psi}_{i,MLE}$ is the maximum likelihood estimate (MLE) of ψ based on the i-th study, and $\hat{\psi}_{i,MLE}$ has the limiting variance $aVar(\hat{\psi}_{i,MLE})$ with the corresponding estimate $\widehat{aVar(\hat{\psi}_{i,MLE})}$ satisfying that the ratio $\widehat{aVar(\hat{\psi}_{i,MLE})}/aVar(\hat{\psi}_{i,MLE})$ converges to 1 in probability. Let the weights in the combining recipe (2) satisfy

$$w_i \propto \{aVar(\hat{\psi}_{i,MLE})\}^{-1/2}, \quad i = 1, \ldots, K. \tag{10}$$

Then the median of the combined distribution function $p_{(c)}(\psi)$, namely $\hat{\psi}_c = p_{(c)}^{-1}(1/2)$, is consistent and asymptotically normally distributed as follows:

$$\left\{\sum_{i=1}^{K} \frac{1}{aVar(\hat{\psi}_{i,MLE})}\right\}^{1/2} (\hat{\psi}_c - \psi_0) \to N(0, 1). \tag{11}$$

The result above clearly indicates the asymptotically "optimal" choices of the weights and transformation function for achieving Fisher efficiency, and also explains why they are used in our combining formula (2). Specifically, using our approach with the weights in (10) and the transformation function $\Phi^{-1}(\cdot)$ can yield asymptotically efficient inference, just as the one using the maximum likelihood approach. This can be seen from the fact that the square of the normalizing constant in (11) satisfies

$$\sum_{i=1}^{K} \frac{1}{\mathrm{aVar}(\hat{\psi}_{i,MLE})} = \frac{1}{\mathrm{aVar}(\hat{\psi}_{MLE})},$$

where $\hat{\psi}_{MLE}$ is the MLE obtained based on all the K studies (Lin and Zeng, 2010). Our simulation results also confirm that such asymptotically "optimal" choices of the weights and transformation function also lead to the most efficient inference among other sensible choices in the setting of rare events data.

To illustrate the statements in Corollary 2 and Theorem 4, we consider an example with the individual p-value function $p_i(\psi)$ obtained from the mid-p adaptation of Fisher exact test for the odds ratio. It can be shown that $p_i(\psi_0; X_i, Y_i)$ converges to U(0,1) in distribution as $n_i \to \infty$ and $m_i \to \infty$, provided that n_i/m_i is bounded away from 0 and ∞. Thus, by Corollary 2, the combined p-value function $p_{(c)}(\psi)$ provides asymptotically accurate inference for the odds ratio, in the sense that the achieved type I error rate converges to the nominal one. Under the same condition, the individual p-value function $p_i(\psi)$ can be expressed in the form of (9) (see Breslow, 1981; Kou and Ying, 1996). Thus, by Theorem 4, the combined p-value function $p_{(c)}(\psi)$ with the weights in (10) leads to asymptotically efficient inference.

4 Empirical methods for improving accuracy and efficiency

4.1 The choice of weights

One distinctive advantage of our approach is that it affords great flexibility in the choice of weights. In this subsection, we examine the impact of different weighting schemes on the combined inference, especially in the rare events setting, and make our recommendation.

For exact inference, the traditional p-value combination method and the confidence interval combination method by Tian et al. (2009) (both can be viewed as special cases of our approach) use (i) equal weight and (ii) the inverse of the sample size (i.e., $1/(n_i + m_i)$) as the weight, respectively. We consider here the third choice by using (iii) the explicit formula of $\{\text{aVar}(\hat{\psi}_{i,MLE})\}^{-1/2}$ as the weight in the combining formula (2). For example, when ψ is the common odds ratio, $\text{aVar}(\hat{\psi}_{i,MLE}) = \psi_0^2 \left[\{n_i\pi_{1i}(1-\pi_{1i})\}^{-1} + \{m_i\pi_{0i}(1-\pi_{0i})\}^{-1}\right]$ (see, e.g., Breslow (1981)). Thus, we can use the following weights

$$w_i \propto \left[\{n_i\pi_{1i}(1-\pi_{1i})\}^{-1} + \{m_i\pi_{0i}(1-\pi_{0i})\}^{-1}\right]^{-1/2}, \quad i = 1, \ldots, K, \tag{12}$$

to implement our approach for the odds ratio. We recommend this weighting scheme based on the following two reasons. First, it approximates the most efficient inference when the sample size are sufficiently large, as shown in Theorem 4. Second, in the rare events setting, our numerical result shows that it gains significant efficiency over the weighting schemes (i) and (ii). In fact, the order of efficiency is typically (i)<(ii)<(iii). Such a result is not surprising, since the weighting scheme (iii) incorporates the sample sizes and the event rates of the studies, both of which are important factors in determining the amount of information contained in a study.

To use the weights in (iii), we first need to estimate the unknown parameters π_{1i} and π_{0i} from the data. Clearly, in the rare events setting, the naive estimates of π_{1i} and π_{0i} using the sample proportions are not reliable. We propose an empirical method to estimate π_{1i} and π_{0i} in the i-th study by borrowing information from the other studies, which is similar to the idea of borrowing strength in Efron (1996). As a result, for instance, if $x_i = y_i = 0$ in the i-th study, our estimation method still yields positive estimates of π_{1i} and π_{0i} if the other studies observe non-zero events. The magnitude of these non-zero estimates is determined jointly by these two sources: a) the information borrowed from the other studies, roughly speaking, the "average level" of the event rates in the other studies; and b) the information provided by the i-th study itself, namely the sample sizes n_i and m_i. The details on our empirical estimation method are provided in Part I of the Appendix. Although the unknown parameters in the weights for combining are estimated, we would still consider the approach an exact method (or at least a good compromise), since all calculations involved are based on exact formulas and distributions.

In the rare events setting, our simulation study (results not reported in this paper) shows that the empirical weights $\hat{w}_i$, (namely, using the empirical estimate of π_{1i} and π_{0i} in the weights (12)) yields results similar to those from the fixed weights (12) when the true values of π_{1i} and π_{0i} are used. Specifically, the empirical weights $\hat{w}_i$ substantially improve the efficiency over the fixed weights (i) and (ii). More importantly, such a gain of efficiency is achieved while maintaining the type I error. In fact, for the type I error rate of our combined inference, we can establish a claim similar to that of Theorem 1, when the fixed weight w_i is replaced by its empirical version $\hat{w}_i = w_i + O_p(1/\sqrt{N_i})$ where N_i is the generic sample size as defined in Section 3.1. Write $N_{min} = \min\{N_1, \dots, N_K\}$ and let $R^{em}_{(c)}(s;\psi_0)$ be the overall type I error rate function when we use the empirical weight $\hat{w}_i$, $\eta_w = \sum_{i=1}^K w_i \Phi^{-1}(p_i(\psi_0))/\sqrt{\sum_{i=1}^K w_i^2}$ and $\eta_{\hat{w}} = \sum_{i=1}^K \hat{w}_i \Phi^{-1}(p_i(\psi_0))/\sqrt{\sum_{i=1}^K \hat{w}_i^2}$. We have the following corollary.

Corollary 3. *Suppose* $\mathrm{E}\{\Phi^{-1}(p_i(\psi_0))\}^2 < C$, *$i = 1, \dots, K$, for a constant C. For a fixed* $0 < s < 1$, *if* $\Pr\{|\eta_w - \Phi^{-1}(s)| \le \delta\} = O(\delta)$ *for any small $\delta > 0$, then we have*

$$R^{em}_{(c)}(s;\psi_0) = s + \sum_{i=1}^K d_i(s) + O\left(1/\sqrt{N_{min}}\right).$$

Here, $d_i(s)$ has the same expression as in Equation (5).

4.2 Adjustment on individual p-value functions

From Section 3 we see that the level of accuracy of individual inference is well reflected in the combined inference through our proposed approach. If the individual inference has high accuracy (with $p_i(\psi_0)$ distributed close to U(0,1)), the combined inference will also have high accuracy (with $p_{(c)}(\psi_0)$ distributed close to U(0,1)). In case individual inference is overly conservative (with $R_i(\alpha, \psi_0) \equiv \Pr(p_i(\psi_0) \le \alpha)$ far smaller than α), the combined inference will suffer the same. Given that our approach applies regardless of whether or not $p_i(\psi_0)$ follows exactly the U(0,1) distribution and it permits a wide range of p-value functions, we can seek adjustment to make the individual $p_i(\psi_0)$ follow U(0,1) closely and thus achieve high accuracy for the combined inference. In this section, we propose a simple adjustment to each individual exact test to reduce the difference between the distribution of its $p_i(\psi_0)$ and U(0,1). Our simulation studies in Section 5 show that combining the adjusted p-value functions can lead to significant power improvement in the rare events setting.

As noted in Boschloo (1970) and Crans and Shuster (2008), we may consider pursuing adjustment on the test result when the exact test is overly conservative, namely, the actual type I error rate falls far below the nominal rate (i.e., $R_i(\alpha, \psi_0) - \alpha \ll 0$). Such conservatism tends to be passed onto the combined inference (Tian et al., 2009), leading to $(R_{(c)}(\alpha, \psi_0) - \alpha \ll 0)$. It follows from Theorem 1 and Corollary 1 that, if the difference $\{R_i(s, \psi_0) - s\}$ for each test is reduced, so will the difference

$\{R_{(c)}(s, \psi_0) - s\}$ for the combined test. To achieve an overall improvement, we propose a simple adjustment on the entire p-value function from each study. The idea is to align the distribution of $p_i(\psi_0)$ as closely to U(0,1) as possible. It is well known that the probability function transformation of a continuous random variable follows U(0,1). More precisely, $F_Z(Z) \sim U(0,1)$ if $F_Z(\cdot)$ is the cumulative distribution function of the continuous random variable Z. Although this property does not hold exactly if Z is discrete, $F_Z(\cdot)$ can still be used as a baseline for adjusting $F_Z(Z)$ to become close to $U(0,1)$. Now consider $p_i(\psi_0)$ as the Z in our discrete setup. Here Z as well as $F_Z(Z)$ both follow discrete distributions. We propose to use a smooth function $G_i(\cdot)$ to approximate $F_Z(\cdot)$, and then impose $G_i(\cdot)$ on Z. For example, for the mid-p adaptation of Fisher exact test, we propose the following adjustment function

$$G_i(s) = \begin{cases} F_{beta}\left\{s; 1 + \frac{\lambda}{m_i\pi_{0i}(1-\pi_{0i})}, 1 + \frac{\lambda}{m_i\pi_{0i}(1-\pi_{0i})}\right\} & \text{if } s \leq 1/2, \\ F_{beta}\left\{s; 1 + \frac{\lambda}{n_i\pi_{1i}(1-\pi_{1i})}, 1 + \frac{\lambda}{n_i\pi_{1i}(1-\pi_{1i})}\right\} & \text{if } s > 1/2, \end{cases} \tag{13}$$

Here λ is a positive tuning parameter, and $F_{beta}(s; \beta_1, \beta_2)$ is the cumulative distribution function of the beta distribution with the parameters β_1 and β_2. If λ is zero or close to zero, there is no or little adjustment. The impact of the adjustment (13) increases as λ increases. An example of this adjustment is illustrated in Figure 3, with the cumulative distribution function of $p_i(\psi_0)$ (in solid step function) and its adjusted version $G_i(p_i(\psi_0))$ (in a dashed step function). There $G_i(\cdot)$ is the dotted S-shaped curve, which mimics the shape of the original cumulative distribution function of $p_i(\psi_0)$. Clearly, the adjusted cumulative distribution function $G_i(p_i(\psi_0))$ shown in Figure 3 is closer to the diagonal line than the original one.

We impose the adjustment function $G_i(\cdot)$ on the entire p-value function from each study and denote it by $p_i^a(\cdot)$, i.e., $p_i^a(\psi) \equiv G_i\{p_i(\psi)\}$. Such an adjustment can typically bring the type-I error rate closer to the nominal level and, at the same time, lead to power improvement for each individual test. Consequently, the type-I error rate of the combined test can be also closer to the nominal level and the power of the combined test can be improved.

The validity of the proposed adjustment can be formally justified by the following three properties. First, the adjustment effect is discernable only when $m_i\pi_{0i}(1 - \pi_{0i})$ and $n_i\pi_{1i}(1 - \pi_{1i})$ are small, and it diminishes when the sample sizes become sufficiently large. The latter can be easily seen by noting that $G_i(\cdot)$ becomes an identity transformation when $m_i\pi_{0i}(1-\pi_{0i})$ and $n_i\pi_{1i}(1-\pi_{1i})$ are sufficiently large. This first property is important because an ideal adjustment approach should magnify the adjustment effect when the sample size is small and thus correct the conservatism of the test when it is overly conservative, and it should also avoid over-adjustment by diminishing the adjustment effect when the sample size is large. Second, this adjustment procedure can be considered robust in the sense that even if the adjustment on an individual test is overly aggressive in such a case that the error bound $-l_i \leq R_i(s; \psi_0) - s < 0$ becomes $-l_i/2 \leq R_i(s; \psi_0) - s < l_i/2$ (namely,

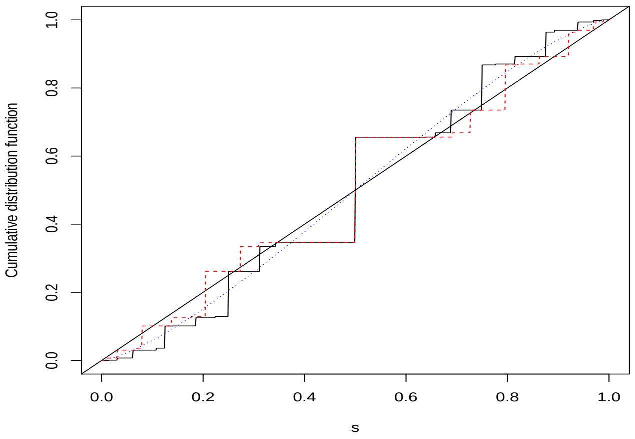

Figure 3: The cumulative distribution function $\Pr\{p_i^a(\psi_0) \leq s\}$ for a single study before ($\lambda = 0$; solid step function) and after ($\lambda = 0.4$; dashed step function) the beta adjustment. The illustration is for the case where $\psi_0 = 1$ and $\pi_{0i} = 0.01$ with the sample sizes $n_i = m_i = 100$. The corresponding beta adjustment function $G_i(s)$ is shown in the dotted curve hugging the diagonal line.

roughly symmetric around zero) after adjustment, the resulting error bound for $\{R_{(c)}(s; \psi_0) - s\}$ for the combined inference will not exceed much above zero. The latter is due to a "smoothing" effect of our proposed combining, which is further illustrated in Figure 4. There we plot respectively, in a solid curve and a dashed curve, the cumulative distribution function of an aggressively adjusted $p_i^a(\psi_0)$ (by using a large λ with $\lambda = 0.8$ in (13) for an individual study) and the combined $p_{(c)}^a(\psi_0)$ resulting from combining 10 independent such individual studies. It is easy to see that, even though each individual test is overly adjusted, the proposed combining approach has such a smoothing effect that the difference $\{R_{(c)}(s; \psi_0) - s\}$ does not stray far from 0. Finally, we also note that all the theoretical results in Section 3 remain valid as long as the properties of the adjusted p-value function $p_i^a(\psi)$ are known. In particular, the formula for determining the overall type I error rate in Theorem 1 still holds, and it can enable us to: i) estimate the overall type I error rate from the observed data, and ii) monitor the adjustment effect to avoid over-adjustment by calibrating the tuning parameter λ. The numerical studies in Section 5 indicate that the choice of $\lambda = 0.4$ works well in general among the cases we have considered.

5 Numerical studies

We now proceed to examine the performance of the proposed exact meta-analysis approach through simulated and real data sets. Specifically, we apply our approach to two data sets, one involves

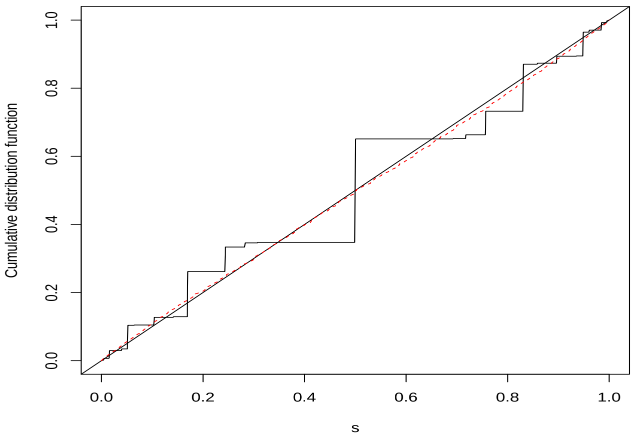

Figure 4: An illustration for the smoothing effect of the proposed combining approach. The solid curve depicts the cumulative distribution function $\Pr\{p_i^a(\psi_0) \leq s\}$ after an "aggressive" adjustment ($\lambda = 0.8$) for an individual study where $\psi_0 = 1$ and $\pi_{0i} = 0.01$ with the sample sizes $n_i = m_i = 100$. The dashed curve depicts the cumulative distribution function $\Pr\{p_{(c)}^a(\psi_0) \leq s\}$ for the overall inference after combining 10 such individual studies with over-adjustment.

the diabetes drug Avandia and the other the promotion rates for white and black employees. The Avandia data (see Nissen and Wolski, 2007, Table 3) are used to examine whether Avandia is associated with myocardial infarction or cardiovascular death. The promotion data (see Gastwirth, 1984, Table 8) are used to compare the promotion rates of white and black employees in a certain workplace. The Avandia data set consists of a large number of clinical trials of moderate to large sizes with very low adverse event rates ($K = 48$, median$\{n_i\} = 222$, median$\{m_i\} = 142$). It consists of many zero events, among them there are 10 zero total event studies. The promotion data set consists of a small number of small survey studies with moderately low event rates ($K = 10$, median$\{n_i\} = 25$, median$\{m_i\} = 9$). One particular feature of this data set is that zero event is observed throughout one arm. To be consist with the original analyses of the two data sets and also to facilitate a direct comparison later, we use the odds ratio as the risk measure. For this measure, we compare our combining p-value functions method with the two most common methods, by Mantel–Haenszel and Peto, as well as with an existing exact method by (Gart, 1970) (which is based on conditional likelihood inference). In our approach, the individual p-value functions are obtained based on the mid-p adaptation of Fisher exact test, which is one of the most popular exact tests in practice. Finally, when the risk difference is used as the effect measure, we provide a comparison study between our method and the exact method by Tian et al. (2009). Note that the conditional likelihood inference method in Gart (1970) is not applicable for risk difference.

5.1 Simulation results

In the first simulation study, we generate $K = 48$ independent studies with the sample sizes corresponding to those in the Avandia data. For the i-th study, the event rate π_{0i} in the control arm is generated from a uniform distribution U(0,ξ), where ξ is set to be a small number to ensure a certain low event rate. The event rate in the other arm is determined by $\text{logit}(\pi_{1i}) = \log(\psi) + \text{logit}(\pi_{0i})$ for a fixed odds ratio ψ ranging from 1 to 10. The data (x_i, y_i) are generated using the binomial model described in Section 2.1, now with a non-negligible probability of generating a sizable zero total event studies. This simulation setting is similar to those in Bradburn et al. (2007) and Tian et al. (2009), and can thus facilitate more direct comparisons between their findings and ours.

Figure 5(a) presents the empirical coverage probability of 95% confidence intervals when $\pi_{0i} \sim \text{U}(0, 0.01)$. We can see that the coverage probabilities of Mantel–Haenszel method with 0.5 correction to zero events (denoted by MH-0.5), Peto method without and with 0.5 correction to zero events (denoted respectively by Peto-0 and Peto-0.5) all decrease quickly as the true odds ratio increases above one. Only the proposed method of combining original p-value functions or adjusted p-value functions, Mantel–Haenszel method without correction to zero events (denoted by MH-0) and Gart's exact method can yield confidence intervals with adequate coverage probability. Among these four valid methods, our method of combining adjusted p-value functions yields the highest power for testing the hypothesis $H_0 : \psi = 1$ versus $H_1 : \psi \neq 1$, as shown in Figure 5(b). Our findings on the existing methods here are in line with Finkelstein and Levin (2012).

In the second simulation study, we repeat the same simulation procedure but use the data structure of the promotion data. The analysis results are shown in Figure 5(c)–(d) for $\pi_{0i} \sim \text{U}(0,0.05)$. In this situation, there is a non-negligible chance that zero events are observed in one arm for all the simulated studies, just as what is seen in the real promotion data. For such a case, implementing Mantel–Haenszel method requires corrections to zero events. Figure 5(c) shows that MH-0.5 method and Peto-0.5 method have a severe coverage problem with very low coverage probability. For example, their 95% confidence intervals have coverage probabilities below 80% and 70% when the true odds ratio $\psi = 2$ and 3, respectively. For Peto-0 method, the coverage probability is adequate when $\psi \leq 4$, but falls quickly as ψ increases further. These observations are consistent with the findings in Bradburn et al. (2007), where Peto-0 method is recommended for its best confidence interval coverage and most powerful test result when the true odds ratio is not too large. On the other hand, Figure 5(c) shows that our proposed exact method and Gart's exact method maintain adequate coverage probability consistently for all the odds ratios throughout the range of the plot. We therefore compare in Figure 5(d) their testing power, together with Peto-0 method. We see that the combining of adjusted p-value functions achieves the highest power, Peto-0 method and Gart's exact method have comparable power, and the combining of the original p-value functions has power increment not as rapidly as other methods, most likely because the individual tests here are overly conservative.

The proposed approach enables us to evaluate $R_{(c)}(\alpha;\psi_0)$ and thus the actual coverage probability of the level $100(1-\alpha)\%$ confidence interval $(p_{(c)}^{-1}(\alpha/2), p_{(c)}^{-1}(1-\alpha/2))$

$$\Pr\left\{\psi_0 \in (p_{(c)}^{-1}(\alpha/2), p_{(c)}^{-1}(1-\alpha/2))\right\} = R_{(c)}(1-\alpha/2;\psi_0)^{-} - R_{(c)}(\alpha/2;\psi_0).$$

To evaluate the accuracy of such estimation for our method, we compare the estimated with the actual coverage probability for the 95% confidence intervals, all in the setting of our first simulation study. The results are shown in Table 1. Part I of Table 1 shows that, for combining the original p-value functions, the absolute difference between the estimated and actual coverage probability is never greater than 0.5% for the listed odds ratios, ranging from 1 to 10. Part II of Table 1 shows that, for combining the adjusted p-value functions, the difference is slightly higher, but still no greater than 1.5%. Both cases show that our approaches, with or without adjustments, achieve high accuracy.

We also examine the impact of zero total event studies on the proposed combining method. To achieve this, we repeat the same simulation procedure as presented before but artificially remove all zero total event studies from our analysis. We compare in Table 2 such analysis with the full analysis of all the available data. Table 2 shows that, overall, the full analysis results in slightly higher coverage probability, and correspondingly slightly lower testing power. This observation indicates that zero total event studies yield a slightly more conservative inference result, if the mid-p adaptation of Fisher exact test is used in the proposed combining method. It is worth noting that this observed phenomenon is test specific and does not necessarily hold for other exact tests.

Finally, if the risk difference RD is the risk measure of interest, our proposed method and the exact method by Tian et al. (2009) can be readily implemented, but not Gart (1970). We report in Table 3 the simulation result obtained from the two methods with π_{0i} being generated from U(0,0.01%) under the setting of the first simulation study. Comparing to the method by Tian et al. (2009), the results in Table 3 show that our method yields slightly better though comparable results when the original p-value functions are combined, and it leads to substantial improvement in efficiency when the adjusted p-value functions are combined.

5.2 Real data analysis results

Nissen and Wolski (2007) used Peto method (Peto-0 in Table 4) to analyze the Avandia data. For the endpoint of myocardial infarction, they obtained a 95% confidence interval of (1.031, 1.979) and a p-value of 0.032 for testing that the odds ratio is 1, and thus concluded that Avandia is significantly associated with myocardial infarction. Table 4 shows that Mantel–Haenszel method

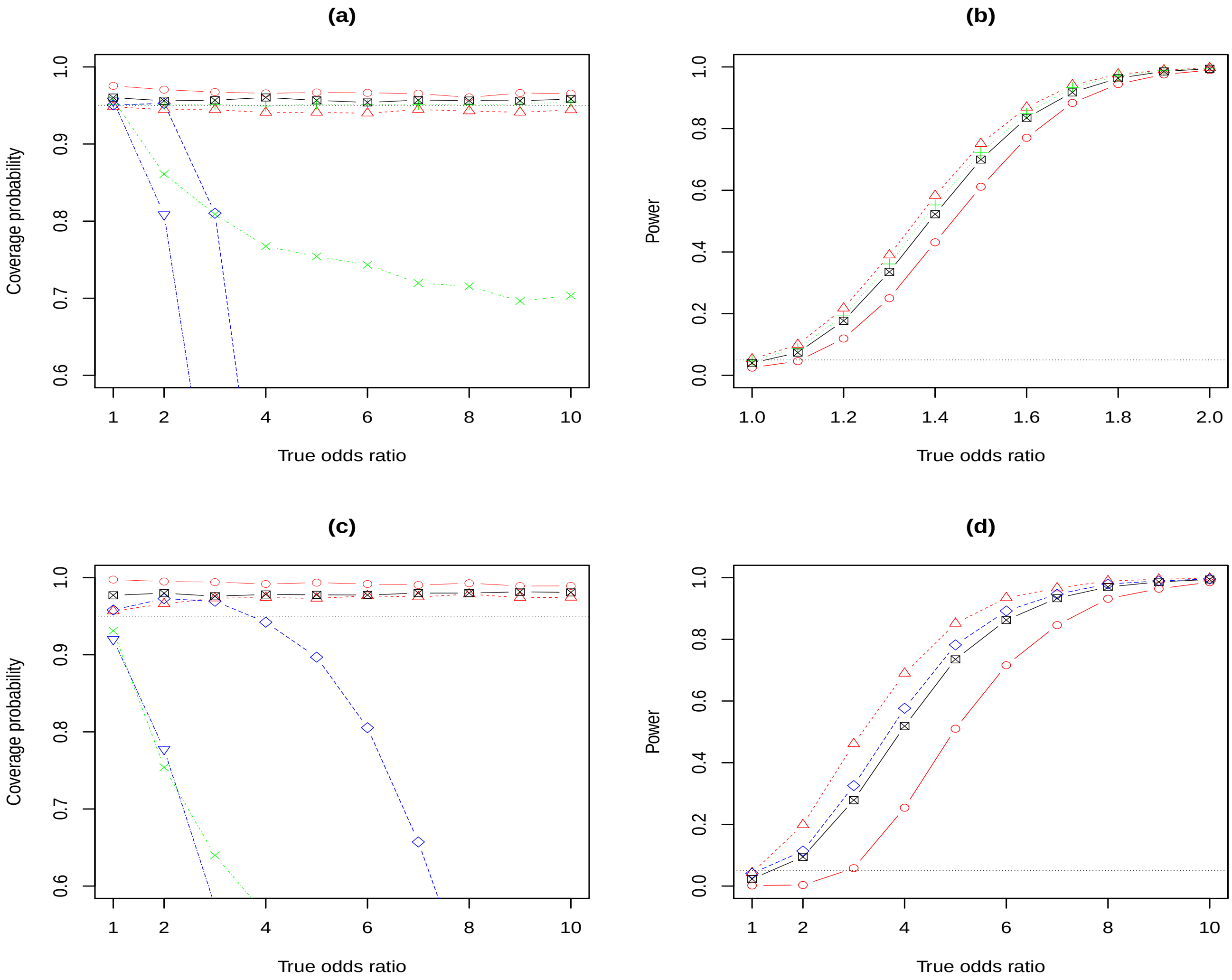

Figure 5: Empirical coverage probability of 95% confidence intervals and empirical power of testing $H_0 : \psi = 1$ versus $H_1 : \psi \neq 1$, for the odds ratios between 1 and 10. The empirical results are calculated based on 10000 data sets simulated from the structures of the Avandia data (a)-(b), and from the promotion data (c)-(d). The baseline event rate π_{0i}, $i = 1, \ldots, K$, are generated from U(0, 0.01) and U(0, 0.05) for illustrations (a)-(b) and (c)-(d), respectively. The methods illustrated are: the proposed method of combining p-value functions ($\circ$) and combining the adjusted p-value functions with tuning parameter $\lambda = 0.4$ ($\triangle$); Mantel–Haenszel method without ($+$) and with ($\times$) 0.5 corrections for every cell of the 2×2 table with zero event; Peto method without ($\diamond$) and with ($\triangledown$) 0.5 corrections for every cell of the 2×2 table with zero event; and Gart's exact method ($\boxtimes$).

Table 1: Estimated coverage probability of a 95% confidence interval

Part I. Combining the original p-value functions

True odds ratio	1	2	3	4	5	6	7	8	9	10
Actual coverage (%)	97.5	97.0	96.7	96.6	96.7	96.6	96.5	96.0	96.6	96.5
Estimated coverage (%)	97.8	97.1	96.8	96.7	96.6	96.6	96.5	96.5	96.5	96.4
Absolute difference (%)	0.3	0.1	0.1	0.1	0.1	0.0	0.0	0.5	0.1	0.1

Part II. Combining the adjusted p-value functions

True odds ratio	1	2	3	4	5	6	7	8	9	10
Actual coverage (%)	94.8	94.5	94.4	94.1	94.1	94.0	94.5	94.3	94.1	94.4
Estimated coverage (%)	96.0	95.7	95.6	95.5	95.4	95.4	95.4	95.3	95.3	95.3
Absolute difference (%)	1.2	1.2	1.2	1.4	1.3	1.4	0.9	1.0	1.2	0.9

(MH-0 in Table 4) yields similar significant result. However, after applying 0.5 corrections to zero events, the two methods, Peto-0.5 and MH-0.5, yield p-values of 0.158 and 0.163, respectively. Neither of the results is significant even at $\alpha = 0.1$ significance level. This observation implies that, for Peto and Mantel–Haenszel methods, the use of corrections to zero events may result in contradictory conclusions. Our finding here is consistent with that in Sweeting et al. (2004) which reports that the imputation to zero events can result in very different conclusions, depending on the numbers imputed. Table 4 shows that the results from different exact analysis are more consistent with each other. For example, Gart's exact method yields a 95% confidence interval of (1.016, 2.005) and a p-value of 0.040. Combining the original p-value functions, our method yields a 95% confidence interval of (0.972, 2.001) and a p-value of 0.071. We calculate the estimate of coverage probability, which is 97.3% indicating that the result here may be conservative. Combining the adjusted p-value functions, our method yields a 95% confidence interval of (1.037, 2.004) and a p-value of 0.029, and the estimated coverage probability is 96.1%. When all the zero total event studies are removed, combining the original and the adjusted p-value functions yield 95% confidence intervals of (0.978,1.994) and (1.040,1.996), respectively. Both intervals are slightly narrower than those obtained by analyzing the entire data. Similar results and discussion can be made for the other endpoint of cardiovascular death studied in Table 4.

The promotion data has a special feature that no promotion is observed in the arm of black employees across all the studies (i.e., observing all zeros in one arm). For this case, MH-0 method is clearly not applicable, and MH-0.5, Peto-0 and Peto-0.5 in Table 4 all yield confidence intervals

Table 2: Simulation result with and without zero total event studies

Part I. The Avandia data structure

True odds ratio	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9
Combining the original p-value functions										
CP (%)	97.5	97.7	97.6	97.4	97.2	97.3	97.3	97.0	97.3	97.3
CP-w/o (%)	97.2	97.1	97.0	97.3	96.9	96.7	97.2	97.2	96.9	96.6
Power (%)	2.5	4.6	11.9	25.0	43.1	61.1	77.0	88.3	94.4	97.6
Power-w/o (%)	2.8	4.9	11.9	26.8	45.2	62.9	79.4	88.4	94.7	97.4
Combining the adjusted p-value functions										
CP (%)	94.8	95.1	94.8	94.8	94.7	94.7	94.9	94.6	94.7	94.5
CP-w/o (%)	94.8	94.3	94.3	94.7	94.4	94.1	94.6	94.5	94.6	94.1
Power (%)	5.3	10.0	21.7	39.0	58.2	75.1	86.9	94.1	97.6	99.0
Power-w/o (%)	5.6	10.1	21.3	40.4	59.0	75.2	87.6	93.9	97.7	98.9

Part II. The promotion data structure

True odds ratio	1	2	3	4	5	6	7	8	9	10
Combining the original p-value functions										
CP (%)	99.8	99.5	99.4	99.2	99.4	99.2	99.0	99.3	98.9	98.9
CP-w/o (%)	99.3	98.5	99.0	98.9	99.1	99.4	99.2	98.8	98.7	98.8
Power (%)	0.2	0.3	5.8	25.4	51.0	71.6	84.6	93.2	96.4	98.5
Power-w/o (%)	0.6	1.3	9.4	27.3	50.9	76.0	88.5	92.8	97.0	98.4
Combining the adjusted p-value functions										
CP (%)	95.7	96.6	97.3	97.4	97.3	97.6	97.5	97.9	97.4	97.4
CP-w/o (%)	94.0	95.6	96.4	96.6	97.4	98.1	97.4	97.3	97.1	97.7
Power (%)	4.3	19.9	46.1	69.0	85.2	93.5	96.6	98.9	99.5	99.8
Power-w/o (%)	5.9	23.9	46.7	72.9	86.7	95.2	98.3	98.8	99.9	99.7

Remark: CP=Coverage Probability; Power=Power for rejecting the hypothesis $H_0 : \psi = 1$ versus; w/o=without zero total event studies.

Table 3: Empirical coverage probability and average length of 95% confidence intervals when risk difference is used as the effect measure

	Tian et al.		Proposed exact		Proposed exact (adj)	
Risk difference(%)	CP(%)	Avg. length	CP(%)	Avg. length	CP(%)	Avg. length
0	100	1.47	100	1.45	100	0.98
0.01	100	1.59	100	1.54	100	1.07
0.05	100	1.88	100	1.81	99.6	1.34

Remark: CP = Coverage Probability; Avg. length= Average length ×1000.

Table 4: Analysis result of the Avandia data and the promotion data

	Avandia data				Promotion data	
	Myocardial infarction		Cardiovascular death			
	95% CI	P	95% CI	P	95% CI	P
Peto-0	(1.031, 1.979)	0.032	(0.980, 2.744)	0.060	(1.522, 12.86)	0.006
Peto-0.5	(0.921, 1.659)	0.158	(0.761, 1.690)	0.538	(0.776, 4.270)	0.168
MH-0	(1.029, 1.978)	0.033	(0.984, 2.930)	0.057	-	-
MH-0.5	(0.919, 1.647)	0.163	(0.760, 1.689)	0.541	(0.738, 5.396)	0.174
Gart's exact	(1.016, 2.005)	0.040	(0.949, 2.981)	0.078	$(2.298, \infty)$	0.004
Proposed exact	(0.972, 2.001)	0.071	(0.765, 2.965)	0.252	$(0.842, \infty)$	0.080
Proposed exact(adj)	(1.037, 2.004)	0.029	(0.956, 2.981)	0.073	$(1.054, \infty)$	0.042

Remark: CI = Confidence interval; P = p-value for hypothesis testing $H_0 : \psi = 1$ versus $H_1 : \psi \neq 1$.

with finite upper bounds. This is because Mantel–Haenszel and Peto methods obtain Wald-type intervals by computing point estimates plus/minus a constant times estimated standard errors. It is important to note that, however, the promotion data set, with no events in one whole arm, does not provide any evidence for rejecting the hypothesis $H_0 : \psi = \psi^*$ and favoring $H_1 : \psi < \psi^*$, for any value of ψ^*. Thus, any finite upper bound placed on the odds ratio may be misleading. In contrast, Table 4 shows that all the exact methods, Gart's and ours, do not have such a problem. They yield infinity as the upper end of the confidence interval, due to the use of the exact distributions of test statistics. Finally, we remark that, although Gart's method performs comparably to ours here, it is tailored specifically for the inference of odds ratios with no clear extensions beyond this setting.

6 Discussion

In this paper, we have proposed a general exact meta-analysis approach for discrete data settings by combining p-value functions associated with the exact tests from individual studies. This approach encompasses a broad class of exact meta-analysis methods, as it permits a wide range of choices for the combining elements, such as tests, weights, transformation functions, and adjustments. The combining formula used in the approach has a simple structure that allows us to explicitly derive the theoretical statements about the combined inference. Guided by those statements, we have been able to devise empirical methods to further improve the inference efficiency of our approach. We have demonstrated, through numerical studies in the rare events setting, that our exact approach is efficient and, generally, outperforms existing commonly used meta-analysis methods. Although Gart's method performs comparably to ours for the inference of odds ratios, it does not apply outside the realm of odds ratios while ours applies readily in general settings and parameters.

Throughout the paper, we have emphasized the great generality of our approach, especially in terms of its flexibility to accommodate different choices of weights, tests, transformation functions, and even adjustments to the original p-value functions. However, the emphasis on generality should not be misconstrued as an advocacy for "arbitrary choice". Instead it calls for the understanding of the impact of any particular choice towards the pre-set desired properties if there are any. For example, we made recommendations on different choices in Section 4 based on considerations of efficiency, accuracy, or other practical concerns. Some remarks on the choices of combining elements are also given below.

The idea of combining functions in meta-analysis was proposed in Xie et al. (2011) for continuous settings and in the context of combining *confidence distribution functions*. The combining formula (2) was introduced and investigated in this context. In this paper, we extend and justify the use of this approach to combining p-value functions in discrete settings. With p-value functions as the specific elements for combining, we are able to establish several more explicit formulas, which make our combined inference results also explicitly tractable. In addition, unlike the approach in the continuous case, our approach still applies even if the p-value functions under the true parameter value deviates from the desired distribution U(0,1), as discussed in Sections 3.1 and 3.2.

To a certain extent, our approach may be viewed as a generalization of the classical approach of combining p-values (see, e.g., Fisher, 1932; Stouffer et al., 1949). However, unlike the classical approach which is to combine the observed p-values, our approach is to combine the entire p-value *functions*. Moreover, the classical approach uses only equal weights in the combination, which is known to be inefficient in terms of preserving Fisher information (e.g., Xie et al., 2011). In contrast, our approach can afford flexible weights in the combination. In fact, we show in Section 4.1 that with suitably chosen weights our approach can achieve substantial gain of efficiency in analyzing rare events data, and in the case where a large sample theory applies, our combined estimator is asymptotically efficient.

Our approach allows $p(\psi_0)$ (ψ_0 is the true value of ψ) to be non-uniformly distributed, even asymptotically, which is inevitably in discrete data analysis. Clearly, such non-uniform distribution results in loss of inference accuracy in the sense that the type I error rate strays from the nominal level. We have proposed in Section 4.2 some proper empirical adjustments to p-value functions, and shown through numerical studies that combining adjusted p-value functions can significantly improve the testing power in the rare events setting. Thus, we recommend combining adjusted p-value functions when the sample size is small, where the the test is overly conservative and the power of the original p-value may be curtailed. On the other hand, if the sample size is not small, it may be unnecessary to implement additional adjustments, since the proposed adjustment function $G(\cdot)$ is almost an identity function. Naturally, the approach of combining original p-value functions should be preferred if the investigation requires strict control of the type I error and cannot tolerate any over-adjustment. Indeed there is then no theoretical guarantee that the overall type I error is always below the nominal value after applying the proposed adjustment.

The combining formula (2) can accommodate any transformation function $h(\cdot)$ as long as it is monotonically increasing. The associated theoretical results for different choices of h may vary in form but can be established following similar procedures in this paper. We choose to use the inverse function of the standard normal distribution function $\Phi^{-1}(\cdot)$ as the default transformation function in (2), because this choice approximates the most efficient inference in the large-sample setting (cf. Theorem 4) and yields good numerical performance in the small-sample setting as well. Different transformation functions may be employed to obtain certain desired properties. For example, to achieve Bahadur efficiency for combining inferences in the continuous data setting, Singh et al. (2005) recommended using the double exponential distribution function ($F(x) = \{1 + \mathrm{sgn}(x)(1-e^{-x})\}/2$) as the transformation function. However, such a choice may be inferior to using the normal distribution in terms of Fisher efficiency (as observed in Xie et al. (2011)). We have conducted a numerical comparison study for these two transformation functions for the rare events setting considered in this paper. Our findings (not reported here) agree with those in Xie et al. (2011). Note that our approach is intrinsically designed to achieve proper probability coverages for confidence intervals, and the choice of transformation functions obviously would affect the length of the resulting confidence interval which is of importance in most meta-analysis settings.

Although we have used the mid-p adaptation of Fisher exact test throughout the paper to illustrate our approach, we stress that our approach applies to any valid exact test, and, better still, it even allows individual studies to use different tests. There exist many exact tests for testing the association in a 2×2 table. The mid-p adaptation of Fisher exact test, Fisher-Boschloo test

(Boschloo, 1970) and Suissa and Shuster test (Suissa and Shuster, 1985) have been recommended in the review paper Lydersen, Fagerland, and Laake (2009). There are also exact tests for the odds ratio and risk ratio proposed by Agresti and Min (2002) and Reiczigel, Abonyi-Tóth, and Singer (2008) respectively. Among the existing ones, the mid-p Fisher exact test is perhaps the most commonly used one. Although it does not guarantee its type I error rate to be no higher than the nominal level, it is considered as "a good compromise" by many (cf. Agresti and Min, 2002). But, if the investigation requires a strict control of the type-I error rate, other tests such as the point-probability method discussed in (Fleiss et al., 2003; Section 2.7.1) may be more appropriate. In fact, Fleiss et al. (2003) remarked that the point-probability method tends to produce results less conservative than the equal-tailed method. In the context of Avandia data, the point-probability method yields a 95% confidence interval of (1.024, 2.002) and a p-value of 0.037 for the endpoint of myocardial infarction, and appears to approximate the mid-p correction in efficiency without violating the type I error constraint.

Despite the broad range of choices of the combining elements, our approach can be summed up with a simple guideline:

I) **Assume that there are required or preferred tests for individual studies** –

Step 1. Use the recommended weights in (12), following the reasoning in Section 4.1.

Step 2. The inverse function of the standard normal distribution function can be considered the default transformation function in the combining formula (2), following the justifications given in this paper. If specific properties are desired for the combined inference, they may be achieved by using certain transformation functions. For example, one may consider using the double exponential distribution function as the transformation function if Bahadur efficiency for combined inference is desired.

Step 3. Apply the formulas in Section 3.1 to monitor the observed type I error rate (or accuracy) of the combined test. If it deviates too much from the nominal one and an improvement is desired, one can apply the adjustment formula (13) in the procedure described in Section 4.2 to adjust p-value functions. The accuracy can be adjusted by calibrating the tuning parameter λ in (13).

II) **Assume that there are no preferred tests for individual studies** –

One can look into a few commonly recommended tests for individual studies. For each particular set of tests one can follow the steps in I) to assess their combined inference and choose the best performing one(s).

Finally, we make a few remarks on meta-analysis of rare events data in discrete settings.

Remark A. The widely used 0.5 correction to zero events should be avoided, because such a correction can lead to severe bias in the inference, as seen in this paper as well as (Bradburn et al., 2007; Finkelstein and Levin, 2012). Other corrections to zero events may be acceptable, but it is imperative that a pursuit of any type of correction be calibrated by an accompanying sensitivity analysis, as asserted by (Sweeting et al., 2004). Our proposed approach has the advantage of requiring no corrections at all.

Remark B. Asymptotic methods that rely on large-sample approximations should be used with the understanding that the associated inference may be invalid, as shown in this paper, Bradburn et al. (2007), Tian et al. (2009) and others.

Remark C. So far, we have observed that the so-called zero total event study (i.e., observing zero events in both arms) has no impact on the inference outcome with Mantel–Haenszel, Peto and Gart exact methods, but it makes the overall inference slightly more conservative (with widened confidence intervals) in our proposed approach when the mid-p adaptation of Fisher exact test is used. It appears that a definitive conclusion or consensus is yet to be reached on the effect of a zero total event study on meta-analysis, and that further research on this subject is clearly needed. Meanwhile, our proposed approach has the advantage of automatically including zero total events studies in the analysis without having to first evaluate their possible impact. Some discussion on zero total event studies and the related debates concerning the litigation aspects of Avandia cases can be found in Finkelstein and Levin (2012).

Appendix

Part I. The empirical method for estimating π_{1i} and π_{0i}

We make the working assumption that π_{0i} is a realization from a beta distribution beta(β_1, β_2), noting that the beta distribution family is broad enough for capturing or approximating distributions of different shapes. The estimates of the parameters (β_1, β_2, ψ) are then obtained using the maximum likelihood method as follows:

$$(\hat{\beta}_1, \hat{\beta}_2, \hat{\psi}) = \arg \max_{(\beta_1,\beta_2,\psi)} \sum_{i=1}^{K} \log \int_0^1 f_\psi(x_i, y_i \mid \pi_{0i}) f_{\beta_1,\beta_2}(\pi_{0i}) \mathrm{d}\pi_{0i}, \qquad (14)$$

where $f_{\beta_1,\beta_2}(\pi_{0i}) = \pi_{0i}^{\beta_1-1}(1-\pi_{0i})^{\beta_2-1}/\int_0^1 \pi_{0i}^{\beta_1-1}(1-\pi_{0i})^{\beta_2-1}\mathrm{d}\pi_{0i}$, $f_\psi(x_i, y_i \mid \pi_{0i}) = c(x_i, y_i)\pi_{1i}^{x_i}(1-\pi_{1i})^{n_i-x_i}\ \pi_{0i}^{y_i}(1-\pi_{0i})^{m_i-y_i}$, and $\pi_{1i} = (\psi\pi_{0i})/(1-\pi_{0i}+\psi\pi_{0i})$ in the situation of a common odds ratio. We obtain the empirical conditional density of π_{0i}, namely $f_{\hat{\beta}_1,\hat{\beta}_2,\hat{\psi}}(\pi_{0i} \mid x_i, y_i) \propto f_{\hat{\psi}}(x_i, y_i \mid \pi_{0i}) f_{\hat{\beta}_1,\hat{\beta}_2}(\pi_{0i})$, by substituting the parameters (β_1, β_2, ψ) with their estimates $(\hat{\beta}_1, \hat{\beta}_2, \hat{\psi})$. We then use the mean of this distribution, denoted by $\hat{\pi}_{0i}$, to estimate π_{0i} and the estimate of π_{1i} is $\hat{\pi}_{1i} = (\hat{\psi}\hat{\pi}_{0i})/(1-\hat{\pi}_{0i}+\hat{\psi}\hat{\pi}_{0i})$. This estimation method can apply to the situations of other common parameters, such as the risk ratio and others, with straightforward modification.

The working beta distribution assumption is simply a catalyst for borrowing information from the other studies. For example, in our simulation studies presented in Section 5.1, we generate the event rate π_{0i} from the uniform distributions U(0, ξ). Clearly, such uniform distributions do not belong to the beta distribution family. Nevertheless, our simulation results show that the empirical estimation method still performs well.

Part II. Proofs

Proof of Theorem 1. Define random variables B_{ij} $(i = 1, \ldots, K; j = 1, \ldots, K)$ as seen in Theorem 1. By sequentially conditioning on $p_i(\psi_0)$ $(i = 1, \ldots, K)$, we can establish that

$$\Pr\Big\{p_{(c)}(\psi_0) \leq s\Big\} = E\Bigg\{\Pr\Bigg(\Phi\Bigg[\bigg(\sum_{i=1}^{K} w_i^2\bigg)^{-1/2} \sum_{i=1}^{K} w_i \Phi^{-1}\{p_i(\psi_0)\}\Bigg] \leq s \;\Bigg|\; p_2(\psi_0), \ldots, p_k(\psi_0)\Bigg)\Bigg\}$$

$$\begin{aligned}
&= E\left(\Pr\left[p_1(\psi_0) \leq \Phi\left\{\left(1 + \sum_{j\neq 1}\frac{w_j^2}{w_1^2}\right)^{1/2}\Phi^{-1}(s) - \sum_{j\neq 1}\frac{w_j}{w_1}\Phi^{-1}(p_j(\psi_0))\right\} \;\middle|\; p_2(\psi_0), \ldots, p_k(\psi_0)\right]\right)\\
&= E\left(\Pr\left[U_1 \leq \Phi\left\{\left(1 + \sum_{j\neq 1}\frac{w_j^2}{w_1^2}\right)^{1/2}\Phi^{-1}(s) - \sum_{j\neq 1}\frac{w_j}{w_1}\Phi^{-1}(p_j(\psi_0))\right\} \;\middle|\; p_2(\psi_0), \ldots, p_k(\psi_0)\right]\right)\\
&\qquad + E\left(D_1\left[\Phi\left\{\left(1 + \sum_{j\neq 1}\frac{w_j^2}{w_1^2}\right)^{1/2}\Phi^{-1}(s) - \sum_{j\neq 1}\frac{w_j}{w_1}\Phi^{-1}(B_{1j})\right\}\right]\right)\\
&= \Pr\left[\Phi\left\{\left(\sum_{i=1}^{K} w_i^2\right)^{-1/2}\sum_{i=1}^{K} w_i \Phi^{-1}(B_{1i})\right\} \leq s\right] + d_1(s)\\
&= \Pr\left[\Phi\left\{\left(\sum_{i=1}^{K} w_i^2\right)^{-1/2}\sum_{i=1}^{K} w_i \Phi^{-1}(B_{Ki})\right\} \leq s\right] + \sum_{i=1}^{K} d_i(s)\\
&= s + \sum_{i=1}^{K} d_i(s)
\end{aligned}$$

This completes the proof. □

Proof of Theorem 2. For simplicity, we prove the result when $K = 2$.

$$\begin{aligned}
\Pr\{p_{(c)}(\psi^*) \leq s\} &= E\left[\Pr\left\{p_1(\psi^*) \leq \Phi\left(\frac{\sqrt{w_1^2 + w_2^2}}{w_1}\Phi^{-1}(s) - \frac{w_2}{w_1}\Phi^{-1}(p_2(\psi^*))\right) \;\middle|\; p_2(\psi^*)\right\}\right]\\
&\geq 1 - \left[1 - E\left\{\Phi\left(\frac{\sqrt{w_1^2 + w_2^2}}{w_1}\Phi^{-1}(s) - \frac{w_2}{w_1}\Phi^{-1}(p_2(\psi^*))\right)\right\}\right]\Bigg/ f_1(\sqrt{N_1}).
\end{aligned}$$

Let U_1 be a random variable following the U(0,1) distribution, independent of $p_2(\psi^*)$. We can show that

$$
\begin{aligned}
& E\left\{\Phi\left(\frac{\sqrt{w_1^2+w_2^2}}{w_1}\Phi^{-1}(s)-\frac{w_2}{w_1}\Phi^{-1}(p_2(\psi^*))\right)\right\} \\
= \;& E\left[\Pr\left\{U_1 \le \Phi\left(\frac{\sqrt{w_1^2+w_2^2}}{w_1}\Phi^{-1}(s)-\frac{w_2}{w_1}\Phi^{-1}(p_2(\psi^*))\right) \,\middle|\, p_2(\psi^*)\right\}\right] \\
= \;& \Pr\left\{U_1 \le \Phi\left(\frac{\sqrt{w_1^2+w_2^2}}{w_1}\Phi^{-1}(s)-\frac{w_2}{w_1}\Phi^{-1}(p_2(\psi^*))\right)\right\} \\
= \;& E\left[\Pr\left\{p_2(\psi^*) \le \Phi\left(\frac{\sqrt{w_1^2+w_2^2}}{w_2}\Phi^{-1}(s)-\frac{w_1}{w_2}\Phi^{-1}(U_1)\right) \,\middle|\, U_1\right\}\right] \\
\ge \;& 1-\left[1-E\left\{\Phi\left(\frac{\sqrt{w_1^2+w_2^2}}{w_2}\Phi^{-1}(s)-\frac{w_1}{w_2}\Phi^{-1}(U_1)\right)\right\}\right] \Big/ f_2(\sqrt{N_2}) \\
= \;& 1-\frac{1-s}{f_2(\sqrt{N_2})}.
\end{aligned}
$$

Therefore,

$$
\Pr\{p_{(c)}(\psi^*) \le s\} \ge 1-\left\{1-\left(1-\frac{1-s}{f_2(\sqrt{N_2})}\right)\right\} \Big/ f_1(\sqrt{N_1}) = 1-\frac{1-s}{f_1(\sqrt{N_1})f_2(\sqrt{N_2})}.
$$

This completes the proof. □

Proof of Theorem 3. The proof is similar to the proof of Theorem 1 and thus is omitted. □

Proof of Theorem 4. It is easy to show that

$$
p_{(c)}(\psi) = \Phi\left[\frac{1}{\left\{\sum_{i=1}^K w_i^2\right\}^{1/2}}\sum_{i=1}^K w_i\Phi^{-1}\{p_i(\psi)\}\right] = \Phi\left[\left\{\sum_{i=1}^K \frac{1}{\widehat{\mathrm{aVar}(\hat{\psi}_{i,MLE})}}\right\}^{1/2}(\psi-\hat{\psi}_c)\right]+o(1),
$$

where

$$
\hat{\psi}_c = \left\{\sum_{i=1}^K \frac{\hat{\psi}_{i,MLE}}{\widehat{\mathrm{aVar}(\hat{\psi}_{i,MLE})}}\right\} \Big/ \left\{\sum_{i=1}^K \frac{1}{\widehat{\mathrm{aVar}(\hat{\psi}_{i,MLE})}}\right\}.
$$

The result of Theorem 4 then follows. □

Proof of Corollary 3. Denote by $\gamma = \eta_{\hat{w}} - \eta_w$, we have

$$
\begin{aligned}
& \Pr\{p_{(c)}(\psi_0) \le s\} = \Pr\{\eta_w + \gamma \le \Phi^{-1}(s)\} \\
= \;& \Pr\{\eta_w \le \Phi^{-1}(s) - \gamma \mid \gamma > 0\}\Pr\{\gamma > 0\} + \Pr\{\eta_w \le \Phi^{-1}(s) - \gamma \mid \gamma \le 0\}\Pr\{\gamma \le 0\} \\
= \;& \big[\Pr\{\eta_w \le \Phi^{-1}(s) \mid \gamma > 0\} - \Pr\{\Phi^{-1}(s) - \gamma < \eta_w \le \Phi^{-1}(s) \mid \gamma > 0\}\big]\Pr\{\gamma > 0\} + \\
& \big[\Pr\{\eta_w \le \Phi^{-1}(s) \mid \gamma \le 0\} + \Pr\{\Phi^{-1}(s) < \eta_w \le \Phi^{-1}(s) - \gamma \mid \gamma \le 0\}\big]\Pr\{\gamma \le 0\} \\
= \;& \Pr\{\eta_w \le \Phi^{-1}(s)\} + A = s + \sum_{i=1}^K d_i(s) + A
\end{aligned}
$$

where $A = \Pr\{0 < \eta_w - \Phi^{-1}(s) \leq -\gamma \mid \gamma \leq 0\} \Pr\{\gamma \leq 0\} - \Pr\{-\gamma < \eta_w - \Phi^{-1}(s) \leq 0 \mid \gamma > 0\} \Pr\{\gamma > 0\}$. In the following, we show $A \to 0$ and study its convergence rate.

Since $\mathrm{E}\{\Phi^{-1}(p_i(\psi_0))\}^2 < C$ and by Chebyshev's inequality, we have $\Phi^{-1}(p_i(\psi_0)) = O_p(1)$, for $i = 1, \ldots, K$. It follows that $\gamma = O_p(1/\sqrt{N_{min}})$. So, for the fixed $0 < s < 1$,

$$|A| \leq \Pr\left\{|\eta_w - \Phi^{-1}(s)| \leq |\gamma|\right\} = O_p\left(1/\sqrt{N_{min}}\right).$$

Thus, the statement in Corollary 3 holds. □

References

Agresti, A. (2007), *An introduction to categorical data analysis*, Hoboken, NJ: Wiley-Interscience, 2nd ed.

Agresti, A. and Min, Y. (2002), "Unconditional small-sample confidence intervals for the odds ratio," *Biostatistics*, 3, 379–386.

Bhaumik, D. K., Amatya, A., Normand, S.-L. T., Greenhouse, J., Kaizar, E., Neelon, B., and Gibbons, R. D. (2012), "Meta-Analysis of Rare Binary Adverse Event Data," *J. Am. Statist. Assoc.*, 107, 555–567.

Boschloo, R. D. (1970), "Raised conditional level of significance for the 2× 2-table when testing the equality of two probabilities," *Statistica Neerlandica*, 24, 1–9.

Bradburn, M. J., Deeks, J. J., Berlin, J. A., and Localio, A. R. (2007), "Much ado about nothing: A comparison of the performance of meta-analytical methods with rare events," *Statist. Med.*, 26, 53–77.

Breslow, N. (1981), "Odds ratio estimators when the data are sparse," *Biometrika*, 68, 73–84.

Cai, T., Parast, L., and Ryan, L. (2010), "Meta-analysis for rare events," *Statist. Med.*, 29, 2078–2089.

Crans, G. G. and Shuster, J. J. (2008), "How conservative is Fisher's exact test? A quantitative evaluation of the two-sample comparative binomial trial," *Statist. Med.*, 27, 3598–3611.

Efron, B. (1996), "Empirical Bayes methods for combining likelihoods," *J. Am. Statist. Assoc.*, 91, 538–550.

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 ed.

Fleiss, J., Levin, B., and Paik, M. (2003), *Statistical Methods for Rates and Proportions*, Hoboken, NJ: Wiley-Interscience, 3rd ed.

Fraser, D. A. S. (1991), "Statistical inference: Likelihood to significance," *J. Am. Statist. Assoc.*, 86, 258–265.

Gart, J. J. (1970), "Point and interval estimation of the common odds ratio in the combination of 2× 2 tables with fixed marginals," *Biometrika*, 57, 471–475.

Gastwirth, J. L. (1984), "Statistical methods for analyzing claims of employment discrimination," *Industr. Labor Relat. Rev.*, 38, 75–86.

Kou, S. G. and Ying, Z. (1996), "Asymptotics for a 2× 2 table with fixed margins," *Statist. Sinica*, 6, 809–829.

Lin, D. Y. and Zeng, D. (2010), "On the relative efficiency of using summary statistics versus individual-level data in meta-analysis," *Biometrika*, 97, 321–332.

Lydersen, S., Fagerland, M. W., and Laake, P. (2009), "Recommended tests for association in 2× 2 tables," *Statist. Med.*, 28, 1159–1175.

Nissen, S. E. and Wolski, K. (2007), "Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes," *N. Engl. J. Med.*, 356, 2457–2471.

Reiczigel, J., Abonyi-Tóth, Z., and Singer, J. (2008), "An exact confidence set for two binomial proportions and exact unconditional confidence intervals for the difference and ratio of proportions," *Computational Statistics & Data Analysis*, 52, 5046–5053.

Robins, J., van der Vaart, A., and Ventura, V. (2000), "Asymptotic distribution of P values in composite null models," *J. Am. Statist. Assoc.*, 95, 1143–1156.

Singh, K., Xie, M., and Strawderman, W. E. (2005), "Combining information from independent sources through confidence distributions," *Ann. Statist.*, 159–183.

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

Suissa, S. and Shuster, J. J. (1985), "Exact unconditional sample sizes for the 2× 2 binomial trial," *J. R. Statist. Soc. A*, 317–327.

Sutton, A. J. and Higgins, J. P. T. (2008), "Recent developments in meta-analysis," *Statist. Med.*, 27, 625–650.

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

Tian, L., Cai, T., Pfeffer, M. A., Piankov, N., Cremieux, P. Y., and Wei, L. J. (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. and Singh, K. (2013), "Confidence distribution, the frequentist distribution estimator of a parameter — a Review (with discussion)," *Int. Statist. Rev.*, 81, 3–39.

Xie, M., Singh, K., and Strawderman, W. E. (2011), "Confidence distributions and a unifying framework for meta-analysis," *J. Am. Statist. Assoc.*, 106, 320–333.

Yang, G., Liu, D., and Xie, M. (2012), "Tian's exact meta-analysis method as a special example under the general framework of combining CDs," *Research note. Available at http://stat.rutgers.edu/home/gyang/researches/gmetaRpackage/SupplementaryNotes2-LTviaCDNote.pdf.*