Biometrics 74, 785–794
September 2018

DOI: 10.1111/biom.12817

Quantifying Publication Bias in Meta-Analysis

Lifeng Lin* **and Haitao Chu****

Division of Biostatistics, University of Minnesota, Minneapolis 55455, Minnesota, U.S.A.
**email:* linl@umn.edu
***email:* chux0051@umn.edu

Summary. Publication bias is a serious problem in systematic reviews and meta-analyses, which can affect the validity and generalization of conclusions. Currently, approaches to dealing with publication bias can be distinguished into two classes: selection models and funnel-plot-based methods. Selection models use weight functions to adjust the overall effect size estimate and are usually employed as sensitivity analyses to assess the potential impact of publication bias. Funnel-plot-based methods include visual examination of a funnel plot, regression and rank tests, and the nonparametric trim and fill method. Although these approaches have been widely used in applications, measures for quantifying publication bias are seldom studied in the literature. Such measures can be used as a characteristic of a meta-analysis; also, they permit comparisons of publication biases between different meta-analyses. Egger's regression intercept may be considered as a candidate measure, but it lacks an intuitive interpretation. This article introduces a new measure, the skewness of the standardized deviates, to quantify publication bias. This measure describes the asymmetry of the collected studies' distribution. In addition, a new test for publication bias is derived based on the skewness. Large sample properties of the new measure are studied, and its performance is illustrated using simulations and three case studies.

Key words: Heterogeneity; Meta-analysis; Publication bias; Skewness; Standardized deviate; Statistical power.

1. Introduction

Meta-analysis has become a powerful and widely used tool to integrate findings from different studies and inform decision making in evidence-based medicine (Sutton and Higgins, 2008). However, the chance of a study being published by a scientific journal is frequently associated with the statistical significance of its results: more significant findings are more likely to be published, causing publication bias in meta-analysis of published studies (Begg and Berlin, 1988; Stern and Simes, 1997; Kicinski et al., 2015). Detecting publication bias is a critical problem because such bias may lead to incorrect conclusions of systematic reviews (Sutton et al., 2000).

One class of approaches to detecting publication bias is based on selection models. These approaches typically use the weighted distribution theory to model the selection (i.e., publication) process and develop estimation procedures that account for the selection process; see, for example, Dear and Begg (1992), Hedges (1992), and Silliman (1997a,b). Sutton et al. (2000) provide a comprehensive review. The selection models are usually complicated, limiting their applicability. Moreover, they incorporate weight functions in an effort to correct publication bias, but strong and largely untestable assumptions are often made (Sutton et al., 2000). Therefore, the validity of their adjusted results may be doubtful, and these methods are usually employed as sensitivity analyses.

Another class of methods for publication bias is based on a funnel plot, which usually presents effect sizes plotted against their standard errors or precisions (the inverse of standard errors) (Light and Pillemer, 1984; Sterne and Egger, 2001). In the presence of publication bias, the funnel plot is expected to be skewed; see the illustrative example in Figure 1. One may intuitively assess publication bias by examining the asymmetry of the funnel plot; however, the visual examination is usually subjective. Various statistical tests have been proposed for publication bias in the funnel plot, such as Begg's rank test (Begg and Mazumdar, 1994) and Egger's regression test (Egger et al., 1997) and its extensions (e.g., Macaskill et al., 2001; Rothstein et al., 2005; Harbord et al., 2006; Peters et al., 2006). The rank test examines the correlation between the effect sizes and their corresponding sampling variances; a strong correlation implies publication bias. Egger's test regresses the standardized effect sizes on their precisions; in the absence of publication bias, the regression intercept is expected to be zero. Note that this regression is equivalent to a weighted regression of the effect sizes on their standard errors, weighted by the inverse of their variances; the weighted regression's slope, instead of the intercept, is expected to be zero in the absence of publication bias (Rothstein et al., 2005). The weighted regression version of the test is popular among meta-analysts, probably because it directly links the effect sizes to their standard errors without the standardization process. However, this article considers only the original version of regression as in Egger et al. (1997), since it is closely related to commonly used meta-analysis models; see Section 2 for details. In addition, another attractive method is the trim and fill method, which not only tests for publication bias but also adjusts the estimated overall effect size (Duval and Tweedie, 2000a,b). Although these publication bias tests have been widely used in meta-analysis applications, they may suffer from inflated type I error rate or poor power in certain simulation settings (Sterne et al., 2000; Terrin et al., 2003; Peters

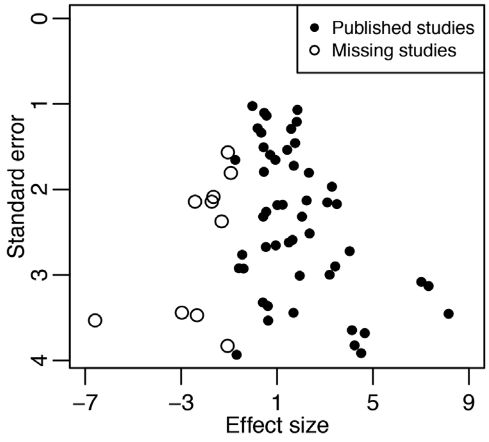

Figure 1. The funnel plot of a simulated meta-analysis containing 60 studies. The 10 studies with the most negative effect sizes were suppressed due to publication bias, and the remaining 50 studies were "published."

et al., 2006, 2007; Rücker et al., 2008). Besides detecting publication bias using selection models and funnel-plot-based methods, it is also important to *quantify* publication bias using measures that permit comparisons between different meta-analyses. A candidate measure is the intercept of the regression test (Egger et al., 1997). However, as a measure of asymmetry of the collected study results, the regression intercept lacks a clear interpretation; for example, it is difficult to provide a range guideline to determine mild, moderate, or substantial publication bias based on the regression intercept. Due to this limitation, meta-analysts usually report the p-value of Egger's regression test, but not the magnitude of the intercept. We will show that the regression intercept basically estimates the average of study-specific standardized deviates; it does not account for the shape of the deviates, which is skewed in the presence of publication bias. This may limit the statistical power of Egger's regression test.

This article introduces an alternative measure to quantify publication bias, the skewness of the standardized deviates. The new measure not only has an intuitive interpretation as the asymmetry of the collected study results but also can serve as a test statistic. The large sample properties of the new measure are studied. We also evaluate its performance using simulations and three actual meta-analyses published in the *Cochrane Database of Systematic Reviews*.

2. Notation and the Regression Test

Suppose a meta-analysis collects n studies; each study reports an effect size y_i (e.g., log odds ratio for binary outcomes) and its within-study variance s_i^2, due to sampling error ($i = 1, \ldots, n$). If the collected studies are deemed homogeneous, sharing a common underlying true effect size μ, then the fixed-effect model is customarily used, specified by $y_i \sim N(\mu, s_i^2)$. The studies are heterogeneous if they have different underlying effect sizes μ_i; the corresponding random-effects model assumes $y_i \sim N(\mu_i, s_i^2)$ and $\mu_i \sim N(\mu, \tau^2)$, where τ^2 is the between-study variance and μ is interpreted as the overall mean effect size (Borenstein et al., 2010). The random-effects model reduces to the fixed-effect model by setting $\tau^2 = 0$.

To detect publication bias, Egger et al. (1997) proposed a regression test, regressing the standardized effect sizes (y_i/s_i) on the corresponding precisions ($1/s_i$); that is,

$$y_i/s_i = \alpha + \mu \cdot 1/s_i + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2).$$

Egger's regression test transforms the original null hypothesis, H_0: no publication bias, to testing H_0': the regression intercept is zero. Alternatively, in the presence of noticeable heterogeneity between studies, we may slightly modify Egger's test by using the marginal standard deviations to produce the regression predictors and responses under the random-effects model. Note that the random-effects model can be written marginally as $y_i = \mu + \delta_i + \xi_i$, where $\delta_i \overset{\text{iid}}{\sim} N(0, \tau^2)$ is the random effect and $\xi_i \sim N(0, s_i^2)$ is the sampling error in study i. Dividing by the marginal standard deviation $(s_i^2 + \tau^2)^{1/2}$, we have the following modified regression test:

$$y_i(s_i^2 + \tau^2)^{-1/2} = \alpha + \mu(s_i^2 + \tau^2)^{-1/2} + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2). \tag{1}$$

Like Egger's test, the intercept α is zero under the true model; in the presence of publication bias, it departs from zero. The overall mean effect size μ becomes the regression slope. Also, σ^2 allows potential under- or over-dispersion of the errors. In practice, heterogeneity is routinely assessed using the Q or I^2 statistic (Whitehead and Whitehead, 1991; Higgins and Thompson, 2002; Higgins et al., 2003; Borenstein et al., 2010), and the between-study variance can be estimated as $\hat{\tau}^2$ using the method of moments or the maximum restricted likelihood method (DerSimonian and Laird, 1986; Normand, 1999). If heterogeneity is not significant, then setting $\tau^2 = 0$ reduces equation (1) to Egger's original test. Since the heterogeneity frequently appears in meta-analyses (Higgins, 2008), this article will introduce publication bias measures based on the modified regression test.

Let the least squares estimates of the regression coefficients in model (1) be $\hat{\alpha}$ and $\hat{\mu}$. The estimated regression intercept is essential in the regression test; we denote this statistic as

$$T_I = \hat{\alpha}.$$

Under the null hypothesis, T_I divided by its standard error follows the t-distribution with degrees of freedom $n-2$, which gives the p-value of the regression test, denoted as P_I. Since the standardized effect sizes are unit-free, the estimated regression intercept T_I is also unit-free. Therefore, T_I can serve as a measure for quantifying publication bias (Egger et al., 1997). However, the regression intercept T_I lacks an intuitive interpretation for the asymmetry of the collected study results. Meta-analysts usually report only the p-value of the regression test, not the magnitude of T_I, to describe the severity of publication bias.

3. Skewness and Skewness-Based Test

The regression test does not fully describe the asymmetry of the collected study results. By linear regression theory, the estimated intercept can be expressed as $T_I = n^{-1} \sum_{i=1}^{n} \hat{d}_i$, where

$$\hat{d}_i = \frac{y_i - \hat{\mu}}{\sqrt{s_i^2 + \hat{\tau}^2}}$$

is an estimate of the study-specific standardized deviate $d_i = (y_i - \mu)(s_i^2 + \tau^2)^{-1/2}$. Therefore, the regression intercept T_I only reflects the *average* of the standardized deviates. To better test and quantify publication bias, we further consider the *shape* of the d_i's.

Note that $d_i = \alpha + \epsilon_i$, so the standardized deviates d_i are distributed with the same shape as the errors ϵ_i. To test the original H_0, we may alternatively test H_0'': $\alpha = 0$ and $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ versus H_1'': $\alpha \neq 0$ or ϵ_i's are iid from a skewed distribution with mean zero. Clearly, H_0'' is stronger than the null hypothesis H_0' of Egger's test, but it is still a necessary condition if the original null hypothesis H_0 holds. Hence, the statistical power should be enhanced by testing H_0'' compared to testing H_0'.

In the statistical literature, skewness has long been used as a descriptive quantity for the asymmetry of a distribution (MacGillivray, 1986), but it is fairly novel in the literature of meta-analysis. To assess publication bias in meta-analysis, we may quantify the asymmetry of $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ by the skewness, calculated as $\text{Skew}(\boldsymbol{\epsilon}) = m_3/s^3$, where $s = \left\{(n-1)^{-1} \sum_{i=1}^{n} (\epsilon_i - \bar{\epsilon})^2\right\}^{1/2}$ is the sample standard deviation, $m_3 = n^{-1} \sum_{i=1}^{n} (\epsilon_i - \bar{\epsilon})^3$ is the sample third central moment, and $\bar{\epsilon} = n^{-1} \sum_{i=1}^{n} \epsilon_i$. In practice, we may replace the unknown errors $\boldsymbol{\epsilon}$ with the regression residuals $\hat{\boldsymbol{\epsilon}} = (\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n)^T$, where $\hat{\epsilon}_i = \hat{d}_i - T_I$. Denote the sample skewness of the errors as

$$T_S = \text{Skew}(\hat{\boldsymbol{\epsilon}}),$$

which we propose as an alternative measure of publication bias. We will show that T_S is a consistent estimate of the true skewness.

The sample skewness T_S can take any real value. A symmetric distribution (i.e., publication bias is not present) has zero skewness. A noticeably large positive skewness indicates that the right tail of standardized deviates' distribution is longer than its left tail. Therefore, some studies on the left side in the funnel plot (i.e., those with negative effect sizes) might be missing due to publication bias. In this situation, the regression intercept T_I is also expected to be positive. On the other hand, a large negative skewness implies that some studies may be missing on the right side. A common but rough rule of interpreting skewness is as follows. If the skewness is less than 0.5 in absolute magnitude, the distribution of the standardized deviates is approximately symmetric; the skewness is deemed considerable if it is between 0.5 and 1 in absolute magnitude, and it may be substantial if its absolute value is greater than 1. To interpret the skewness more rigorously, we study its large sample properties.

Denote $\beta_k = \text{E}(\epsilon_1 - \beta)^k$ as the kth central moment of the errors ϵ_i, where $\beta = \text{E}(\epsilon_1) = 0$, and the sample kth central moment is $m_k = n^{-1} \sum_{i=1}^{n} (\epsilon_i - \bar{\epsilon})^k$. Then the true skewness of the errors is $\gamma = \beta_3/\beta_2^{3/2}$. In addition, let $\hat{m}_k = n^{-1} \sum_{i=1}^{n} (\hat{\epsilon}_i - \bar{\hat{\epsilon}})^k$ be the sample kth central moment after plugging in the known residuals $\hat{\epsilon}_i$; note that $\bar{\hat{\epsilon}} = n^{-1} \sum_{i=1}^{n} \hat{\epsilon}_i = 0$. Denote $\xrightarrow{D}$ as the convergence in distribution. We have the following proposition regarding the asymptotic distribution of the sample skewness T_S.

Proposition 1. *Assume that the study-specific errors ϵ_i have finite sixth central moment (i.e., $\beta_6 < \infty$) and the marginal precisions $(s_i^2 + \tau^2)^{-1/2}$ have finite third moment. Then, $\sqrt{n}(T_S - \gamma)/\sqrt{\hat{v}} \xrightarrow{D} N(0, 1)$ as $n \to \infty$, where*

$$\hat{v} = 9 + \frac{35}{4}\hat{m}_2^{-3}\hat{m}_3^2 - 6\hat{m}_2^{-2}\hat{m}_4 + \hat{m}_2^{-3}\hat{m}_6 + \frac{9}{4}\hat{m}_2^{-5}\hat{m}_3^2\hat{m}_4 - 3\hat{m}_2^{-4}\hat{m}_3\hat{m}_5.$$

Proposition 1 provides an approximate 95% confidence interval (CI) of the sample skewness T_S. Consequently, T_S not only quantifies publication bias but also serves as a test statistic. Under H_0'', we can simplify the asymptotic distribution of T_S as follows.

Corollary 1. *Under the null hypothesis H_0'', $\sqrt{n/6}T_S \xrightarrow{D} N(0, 1)$ as $n \to \infty$.*

The Web Appendix provides the proofs. The p-value of the skewness-based test is calculated using Corollary 1:

$$P_S = 2\left(1 - \Phi\left(\sqrt{n/6}|T_S|\right)\right).$$

The regression intercept T_I quantifies the departure of the average standardized deviate from zero; the skewness T_S quantifies the departure of the standardized deviates' distribution from symmetry. The regression test and the skewness-based test may differ in power in different situations. Therefore, we may combine the test results of T_I and T_S so that the combined test maintains high power across various settings. Under H_0'', note that T_I is the least squares estimate of the intercept and T_S depends only on the residuals $\hat{\epsilon}_i$. Because the least squares estimates of regression coefficients are independent of the residuals if the errors ϵ_i are normally distributed, we immediately have the following proposition.

Proposition 2. *Under the null hypothesis H_0'', T_I and T_S are independent.*

Due to the independence of T_I and T_S, the adjusted p-value for combining T_I and T_S can be calculated as $P_C = 1 - (1 - P_{\min})^2$, where $P_{\min} = \min\{P_I, P_S\}$ (Wright, 1992). The performance of the skewness-based test and the combined test will be studied using simulations and actual meta-analyses.

In practice, many meta-analyses only collect a small number of studies, and the large sample properties may apply poorly for them. Alternatively, a nonparametric bootstrap can be used to derive the 95% CI of the skewness: take samples of size n with replacement from the original data $\{(y_i, s_i^2)\}_{i=1}^n$ for B (say 1000) iterations and calculate 2.5% and 97.5% quantiles of the skewness over the B bootstrap samples. A parametric resampling method can also be used to produce a p-value for the skewness-based test. Specifically, first, estimate the overall mean effect size $\bar{\mu}$ under the null hypothesis that there is no publication bias. Second, draw n samples under the null hypothesis, that is, $y_i^\star \sim N(\bar{\mu}, s_i^2 + \hat{\tau}^2)$, and repeat this for B iterations. Third, based on the B sets of bootstrap samples, calculate the skewness as $T_S^{(b)}$ for $b = 1, \ldots, B$. Finally, the p-value of the skewness-based test is $P_S = \left[\sum_{b=1}^{B} \mathbb{I}(|T_S^{(b)}| \geq |T_S|) + 1\right] / (B + 1)$, where $\mathbb{I}(\cdot)$ is the indicator function. Similar procedures can also be used for the regression intercept T_I.

The proposed methods can be implemented by the functions in the Supplementary Materials, which will be included in our R (R Core Team, 2016) package "altmeta", available on the Comprehensive R Archive Network (CRAN).

4. Simulations

We performed simulations to evaluate the type I error rate and power of the modified regression test T_I, the proposed skewness-based test T_S, and the combined test based on the adjusted p-value P_C. The commonly used Egger's regression test, Begg's rank test, and the trim and fill method (T & F) were also considered. In addition, we calculated the p-values of T_I and T_S using both their theoretical null distributions and the resampling methods. As suggested by many other authors (e.g., Macaskill et al., 2001), the nominal significance level was set to 10% for publication bias tests because the tests usually have low power. For each simulated meta-analysis, the true overall effect size was $\mu = 1$, the within-study standard errors were drawn from $s_i \sim U(1, 4)$, and the between-study standard deviation was set to $\tau = 0$ ($I^2 = 0\%$), 1 ($6\% \leq I^2 \leq 50\%$), and 4 ($50\% \leq I^2 \leq 94\%$). The study-specific effect sizes were then generated as $y_i \sim N(\mu_i, s_i^2)$ and $\mu_i \sim N(\mu, \tau^2)$. The number of studies collected in each meta-analysis was set to $n = 10$, 30, and 50. We considered the following three scenarios to induce publication bias:

I. (Suppressing non-significant findings) We used the above parameters to generate artificial studies, and suppose that they aimed at testing $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$. We assumed that studies with significant findings (i.e., p-value < 0.05 for treatment effect size) were published with probability 1. Also, studies with

Table 1

Type I error rates ($\pi = 1$) and powers ($\pi < 1$) expressed as percentage, for various tests for publication bias due to suppressing non-significant findings (Scenario I). The nominal significance level is 10%.

	$\tau = 0$ ($I^2 = 0\%$)				$\tau = 1$ ($6\% \leq I^2 \leq 50\%$)				$\tau = 4$ ($50\% \leq I^2 \leq 94\%$)			
Test	$\pi = 1$	$\pi = 0.05$	$\pi = 0.02$	$\pi = 0$	$\pi = 1$	$\pi = 0.05$	$\pi = 0.02$	$\pi = 0$	$\pi = 1$	$\pi = 0.05$	$\pi = 0.02$	$\pi = 0$
$n = 10$												
Egger	10	15	23	35	11	14	20	31	13	10	10	11
Begg	7	13	28	57	5	12	23	44	5	4	4	4
T & F	11	8	12	30	11	7	10	21	5	8	9	9
T_I	10	17	26	40	10	17	25	39	10	14	15	16
T_I*	[9]	[21]	[29]	[41]	[11]	[19]	[27]	[39]	[9]	[15]	[17]	[17]
T_S	1	7	20	37	1	8	18	32	1	3	3	4
T_S*	[10]	[27]	[48]	[59]	[10]	[29]	[46]	[58]	[10]	[15]	[17]	[19]
Combined	6	14	29	61	6	14	27	52	5	9	10	11
Combined*	[10]	[26]	[50]	[75]	[10]	[27]	[47]	[68]	[8]	[15]	[17]	[18]
$n = 30$												
Egger	10	17	27	45	10	14	23	35	14	11	12	12
Begg	7	28	64	97	7	24	55	89	5	4	5	6
T & F	12	16	18	17	13	19	20	18	9	21	21	20
T_I	10	18	27	42	10	17	25	36	10	15	16	18
T_I*	[9]	[22]	[33]	[49]	[11]	[21]	[31]	[43]	[10]	[18]	[20]	[22]
T_S	6	50	83	94	6	59	83	92	5	16	20	24
T_S*	[10]	[61]	[88]	[96]	[10]	[70]	[88]	[94]	[10]	[26]	[30]	[34]
Combined	8	42	77	93	8	48	76	90	8	16	19	23
Combined*	[10]	[53]	[85]	[96]	[11]	[61]	[84]	[94]	[9]	[23]	[28]	[32]
$n = 50$												
Egger	9	20	35	58	11	17	28	46	14	12	13	14
Begg	7	38	83	100	7	33	75	98	5	5	7	9
T & F	12	20	17	10	12	23	19	13	9	18	18	18
T_I	9	19	31	49	10	18	28	43	10	16	18	20
T_I*	[9]	[24]	[38]	[57]	[11]	[23]	[34]	[51]	[10]	[19]	[21]	[24]
T_S	7	77	96	99	7	84	96	98	7	30	36	41
T_S*	[10]	[82]	[97]	[99]	[10]	[87]	[97]	[99]	[10]	[37]	[44]	[49]
Combined	8	67	94	99	9	75	93	98	8	25	30	35
Combined*	[9]	[74]	[96]	[100]	[11]	[81]	[96]	[99]	[9]	[31]	[36]	[42]

*The results in square brackets are based on the parametric resampling method.

non-significant findings were published with probability π; the publication rate was set to $\pi = 0$, 0.02, 0.05, and 1. Note that $\pi = 1$ implies no publication bias. Studies were generated iteratively until we obtained n published studies to form a simulated meta-analysis.

II. (Suppressing small studies with non-significant findings) In many cases, small studies with non-significant findings are more likely to be suppressed than large studies; hence, some authors prefer to treat the funnel-plot-based methods as approaches to checking for "small-study effects" (Harbord et al., 2006). We also simulated meta-analyses following this scenario. Studies with significant findings were published with probability 1. Large studies with non-significant findings and standard errors $s_i < 1.5$ were also published with probability 1; however, small studies with non-significant findings and standard errors $s_i \geq 1.5$ were published with probability π, where $\pi = 0$, 0.1, 0.2, and 1. Again, $\pi = 1$ implies no publication bias. The studies were generated iteratively until we obtained n published studies to form a simulated meta-analysis.

III. (Suppressing negative effect sizes) Publication bias can be also induced on the basis of study effect size (Duval and Tweedie, 2000a,b; Peters et al., 2006). For each simulated meta-analysis, $n + m$ studies were generated, and the m studies with the most negative effect sizes were suppressed. We set $m = 0$, $\lfloor n/3 \rfloor$, and $\lfloor 2n/3 \rfloor$, where $\lfloor x \rfloor$ denotes the largest integer not greater than x. Note that $m = 0$ implies no publication bias.

For each setting, 10,000 meta-analyses were simulated. The Monte Carlo standard errors of all type I error rates and powers reported below were less than 1%.

Table 1 presents the type I error rates and powers for Scenario I. Type I error rates of most tests are controlled well, while that of Egger's test is a little inflated when the heterogeneity is substantial ($\tau = 4$). For weak or moderate heterogeneity ($\tau = 0$ or 1), Egger's regression test and the modified regression test T_I have similar power, and Begg's rank test seems to be more powerful than the regression test. Also, the trim and fill method performs poorly. Note that its power drops as π decreases from 0.05 to 0 when $n = 50$

Table 2

Type I error rates ($\pi = 1$) and powers ($\pi < 1$) expressed as percentage, for various tests for publication bias due to suppressing small studies with non-significant findings (Scenario II). The nominal significance level is 10%.

	$\tau = 0$ ($I^2 = 0\%$)				$\tau = 1$ ($6\% \leq I^2 \leq 50\%$)				$\tau = 4$ ($50\% \leq I^2 \leq 94\%$)			
Test	$\pi = 1$	$\pi = 0.2$	$\pi = 0.1$	$\pi = 0$	$\pi = 1$	$\pi = 0.2$	$\pi = 0.1$	$\pi = 0$	$\pi = 1$	$\pi = 0.2$	$\pi = 0.1$	$\pi = 0$
$n = 10$												
Egger	10	14	22	51	11	13	19	43	13	9	10	12
Begg	7	8	13	30	5	7	12	30	5	4	5	7
T & F	11	10	11	15	11	9	10	13	5	4	5	5
T_I	10	15	23	56	10	14	23	54	10	13	16	21
T_I*	[9]	[19]	[29]	[61]	[11]	[19]	[28]	[59]	[9]	[15]	[18]	[25]
T_S	1	1	1	5	1	1	1	5	1	1	1	3
T_S*	[10]	[10]	[11]	[19]	[10]	[9]	[11]	[22]	[10]	[9]	[11]	[17]
Combined	6	9	16	48	6	8	15	46	5	7	9	14
Combined*	[10]	[15]	[23]	[58]	[10]	[14]	[22]	[55]	[8]	[10]	[14]	[21]
$n = 30$												
Egger	10	20	34	69	10	18	30	62	14	10	12	16
Begg	7	16	30	68	7	14	28	66	5	5	7	13
T & F	12	18	23	32	13	15	17	21	9	13	14	13
T_I	10	21	36	70	10	20	33	66	10	14	18	25
T_I*	[9]	[24]	[40]	[74]	[11]	[23]	[37]	[71]	[10]	[17]	[22]	[32]
T_S	6	5	12	54	6	6	14	58	5	6	10	21
T_S*	[10]	[10]	[18]	[59]	[10]	[10]	[21]	[64]	[10]	[11]	[17]	[31]
Combined	8	16	30	80	8	14	28	75	8	10	14	24
Combined*	[10]	[20]	[36]	[83]	[11]	[18]	[33]	[81]	[9]	[13]	[20]	[33]
$n = 50$												
Egger	9	26	46	82	11	24	41	78	14	12	14	20
Begg	7	21	43	85	7	19	41	84	5	5	9	19
T & F	12	17	19	21	12	14	15	13	9	12	12	10
T_I	9	26	46	82	10	25	42	79	10	15	19	29
T_I*	[9]	[29]	[50]	[85]	[11]	[27]	[46]	[82]	[10]	[19]	[24]	[36]
T_S	7	7	20	79	7	9	24	83	7	10	18	36
T_S*	[10]	[10]	[25]	[81]	[10]	[11]	[30]	[85]	[10]	[14]	[24]	[43]
Combined	8	20	41	92	9	19	39	89	8	12	18	34
Combined*	[9]	[23]	[46]	[93]	[11]	[22]	[44]	[91]	[9]	[16]	[24]	[41]

*The results in square brackets are based on the parametric resampling method.

Table 3

Type I error rates ($m = 0$) and powers ($m > 0$) expressed as percentage, for various tests for publication bias due to suppressing the m most negative effect sizes out of a total of $n + m$ studies (Scenario III). The nominal significance level is 10%.

	$\tau = 0$ ($I^2 = 0\%$)			$\tau = 1$ ($20\% \leq I^2 \leq 50\%$)			$\tau = 3$ ($70\% \leq I^2 \leq 90\%$)		
Test	$m = 0$	$\lfloor n/3 \rfloor$	$\lfloor 2n/3 \rfloor$	$m = 0$	$\lfloor n/3 \rfloor$	$\lfloor 2n/3 \rfloor$	$m = 0$	$\lfloor n/3 \rfloor$	$\lfloor 2n/3 \rfloor$
$n = 10$									
Egger	10	21	31	10	19	25	13	15	14
Begg	6	12	18	6	10	14	4	5	6
T & F	11	27	38	11	25	33	5	13	18
T_I	10	21	31	10	18	25	10	11	13
T_I*	[9]	[12]	[13]	[11]	[13]	[12]	[9]	[12]	[13]
T_S	2	2	4	2	2	3	1	2	3
T_S*	[10]	[13]	[17]	[10]	[13]	[16]	[10]	[14]	[16]
Combined	6	14	20	6	12	16	6	7	8
Combined*	[9]	[13]	[15]	[11]	[13]	[15]	[8]	[13]	[16]
$n = 30$									
Egger	10	57	77	11	44	60	14	18	20
Begg	8	46	67	7	35	54	5	12	17
T & F	13	87	97	13	81	92	9	51	63
T_I	10	57	77	10	44	60	10	14	16
T_I*	[10]	[38]	[46]	[12]	[33]	[39]	[10]	[13]	[17]
T_S	6	25	40	6	25	39	6	26	40
T_S*	[10]	[34]	[51]	[11]	[34]	[51]	[10]	[37]	[52]
Combined	8	54	76	8	43	64	8	23	35
Combined*	[10]	[42]	[56]	[12]	[39]	[53]	[9]	[30]	[44]
$n = 50$									
Egger	10	77	93	11	61	80	14	19	22
Begg	8	69	89	8	56	76	5	18	26
T & F	12	98	100	13	95	99	9	69	75
T_I	10	77	93	11	61	80	10	16	20
T_I*	[10]	[59]	[74]	[12]	[52]	[61]	[10]	[15]	[19]
T_S	8	46	69	7	47	68	7	51	69
T_S*	[10]	[53]	[75]	[10]	[54]	[75]	[10]	[58]	[76]
Combined	9	77	95	10	67	87	8	44	62
Combined*	[10]	[65]	[85]	[12]	[62]	[79]	[9]	[49]	[67]

*The results in square brackets are based on the parametric resampling method.

and $\tau = 0$ or 1. Indeed, the trim and fill method is based on the assumption in Scenario III; that is, studies are suppressed if they have most negative (or positive) effect sizes, not according to their p-values. In Scenario I, the two-sided hypothesis testing for treatment effects $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$ can produce significant findings with both negative and positive effect sizes, so the simulated meta-analyses can seriously violate the assumption of the trim and fill method.

For small meta-analysis with $n = 10$, using the asymptotic property in Corollary 1, the skewness-based test T_S is less powerful than the regression test and Begg's rank test when $\pi = 0.02$ or 0.05, and its type I error rate is much smaller than the nominal significance level 10%. This is possibly because T_S's asymptotic property is a poor approximation for small n. However, using the resampling method, the power of T_S is dramatically higher than the other tests when $\tau = 0$ and 1. Moreover, as the number of studies n increases to 30 and 50, the skewness-based test using either the asymptotic property or the resampling method still outperforms the other tests, and its power remains high as the heterogeneity becomes substantial ($\tau = 4$).

Table 2 shows the results for Scenario II. The regression test and Begg's rank test are more powerful than T_S when $\tau = 0$ and 1, while they are outperformed by T_S when $\tau = 4$. In this scenario, T_S seems to be less powerful than in Scenario I. For each simulated meta-analysis, because only small studies with non-significant findings were suppressed, large studies are still symmetric in the funnel plot. Consequently, the distribution of the n studies may have two modes: the large studies are centered around the true overall effect size μ, and the small studies have an overestimated mean due to the suppression. Since the interpretation of skewness is obscure for multi-modal distributions, T_S may lose power in this scenario.

Table 3 presents the type I error rates and powers for Scenario III. Since the trim and fill method's assumption is perfectly satisfied in this scenario, this method is generally more powerful than the other tests. In the absence of heterogeneity ($\tau = 0$), both the regression test and Begg's rank test are more powerful than the skewness-based test T_S; as the

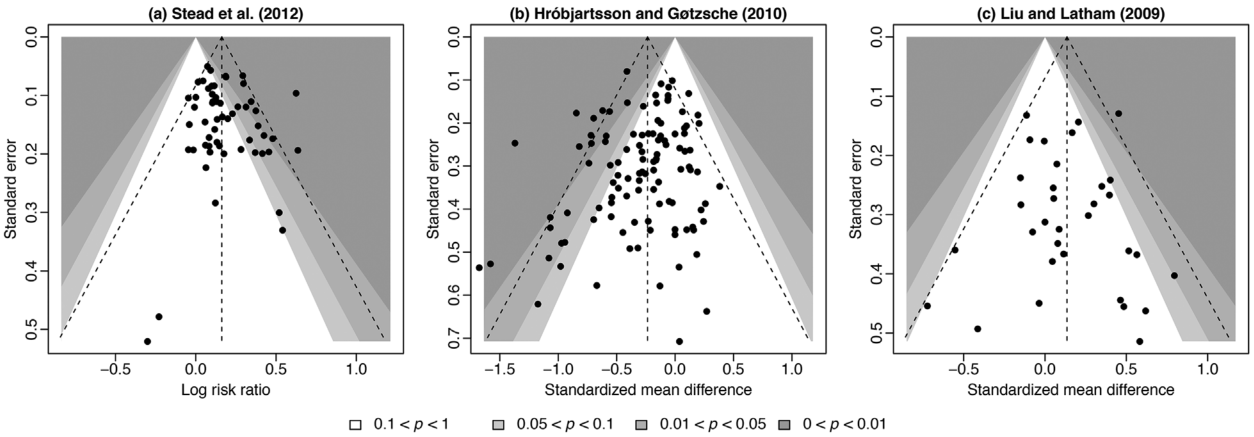

Figure 2. Contour-enhanced funnel plots of the three actual meta-analyses. The vertical and diagonal dashed lines represent the overall estimated effect size and its 95% confidence limits, respectively, based on the fixed-effect model. The shaded regions represent different significance levels for the effect size.

heterogeneity increases, they are outperformed by T_S, especially when n is large.

In summary, the skewness-based test T_S can be much more powerful than the existing tests in some settings, while no test can uniformly outperform the others. Although T_S suffers from low power when the heterogeneity is weak or moderate in Scenarios II and III, the combined test of T_I and T_S maintains high power in most settings by borrowing strengths from each of the separate test.

5. Case Studies

We illustrate the performance of the skewness measure and test by three actual meta-analyses published in the *Cochrane Database of Systematic Reviews.* The first meta-analysis was performed by Stead et al. (2012) to investigate the effect of nicotine gum for smoking cessation; it contains 56 studies and the effect size is the log risk ratio. The second meta-analysis, performed by Hróbjartsson and Gøtzsche (2010), investigates the effect of placebo interventions for all clinical conditions regarding patient-reported outcomes; it contains 109 studies and the effect size is standardized mean difference. The third meta-analysis reported in Liu and Latham (2009) compares the effect of the progressive resistance strength training exercise versus control; it contains 33 studies and the effect size is also standardized mean difference. Figure 2 presents their contour-enhanced funnel plots; the shaded regions represent different significance levels (Peters et al., 2008).

The proposed methods and the commonly used tests were applied to the three meta-analyses, and both the theoretical null distributions and the resampling methods were used to calculate the 95% CIs and p-values for T_I and T_S. We also calculated the p-values for the combined test. Table 4 presents the results. Since the size of each example n is large (for meta-analyses), the 95% CIs and p-values based on the theoretical null distributions are similar to those based on the resampling methods.

For the meta-analysis in Stead et al. (2012), the three commonly used tests yield p-values greater than 0.10, indicating non-significant publication bias; the p-value of the modified regression test T_I is also large. However, the proposed skewness T_S is 0.91 with 95% CI (0.14, 1.68) and p-value 0.005 using the resampling methods; it implies substantial publication bias. Since T_S is significantly greater than zero, some studies with negative effect sizes may be missing. Indeed, the funnel plot in Figure 2(a) shows that most studies are massed on the right side, tending to have significant positive results; some studies are potentially missing on the left side. Moreover, benefiting from the high power of the skewness-based test, the combined test also indicates significant publication bias.

For the meta-analysis in Hróbjartsson and Gøtzsche (2010), all tests imply significant publication bias; the p-values of Begg's rank test, the trim and fill method, and the skewness-based test are fairly small (<0.01). Both the regression intercept T_I and the skewness T_S are significantly negative, indicating that some studies are missing on the right side in the funnel plot; Figure 2(b) confirms this. For the meta-analyses in Liu and Latham (2009), Figure 2(c) shows that its funnel plot is approximately symmetric, so there appears to be no publication bias. Indeed, all tests yield p-values much greater than 0.1, and the publication bias measures T_I and T_S are close to zero.

6. Discussion

This article proposed a new measure, the skewness of the standardized deviates, for quantifying potential publication bias in meta-analysis. The intuitive interpretation of the asymmetry of the collected study results makes this measure appealing; its performance was illustrated by three actual meta-analyses. Also, the skewness can serve as a test statistic and its large sample properties have been studied. The simulations showed that the skewness-based test has high power in many cases. The large sample properties of the skewness did not perform well for small n, but this can be remedied by using resampling methods. In addition, we proposed a combined test that depends on the p-values of both the regression and skewness-based tests; it is shown to be powerful in most simulation settings.

The proposed skewness has some limitations. First, for small meta-analyses, the variation of the sample skewness

Table 4

Results for the three actual meta-analyses

Meta-analysis	No. of studies	I^2 (%)	p-value			Intercept T_I			Skewness T_S			p-value of the combined test
			Egger	Begg	T & F	Measure	95% CI	p-value	Measure	95% CI	p-value	
Stead et al. (2012)	56	39	0.173	0.136	0.500	0.47	(−0.47, 1.41)	0.323	0.91	(0.14, 1.68)	0.005	0.011
							[−0.43, 1.42]	[0.317]		[0.06, 1.50]	[0.005]	[0.010]
Hróbjartsson and Gøtzsche (2010)	109	42	0.049	0.009	<0.001	−0.81	(−1.54, −0.09)	0.028	−0.74	(−1.23, −0.24)	0.002	0.003
							[−1.56, −0.10]	[0.030]		[−1.17, −0.25]	[0.002]	[0.004]
Liu and Latham (2009)	33	11	0.905	0.469	0.500	0.06	(−0.91, 1.02)	0.905	0.01	(−0.63, 0.64)	0.989	0.991
							[−1.09, 1.25]	[0.894]	[−0.73, 0.68]	[0.987]	[0.989]	

Note: The results in square brackets are based on the resampling method.

can be large. Researchers should always use skewness along with its 95% confidence interval. Second, although a symmetric distribution has zero skewness, zero skewness does not necessarily imply a symmetric distribution; for example, an asymmetric distribution may have zero skewness if it has a long but thin tail on one side and a short but fat tail on the other side. Also, the skewness generally describes publication bias well when the effect sizes are unimodal, but its interpretation for multi-modal distributions is obscure. Therefore, the regression intercept is preferred when the studies appear to have multiple modes, which may be identified by visual examining the funnel plot. Third, like many other approaches to assessing publication bias, the skewness is based on checking the funnel plot's asymmetry. However, such asymmetry can be caused by sources other than publication bias (Sterne et al., 2001), such as reference bias (Gøtzsche, 1987; Jannot et al., 2013), studies with poor quality in design (Chalmers et al., 1983; Altman, 2002), the existence of multiple subgroups (Sterne et al., 2011), etc. When applying the methods in this article to detect or quantify the asymmetry of study results, researchers may need to examine carefully whether the asymmetry is caused by publication bias or other sources of bias. In addition, in the simulations and actual meta-analyses, different methods for publication bias can lead to fairly different conclusions. Therefore, we are allowed to use a wealth of methods to detect any potential publication bias.

Like the routinely used I^2 statistic for assessing heterogeneity, the skewness may be a good characteristic of meta-analysis for quantifying publication bias. In the statistical literature, the skewness is a conventional descriptive quantity for asymmetry, but it may not be optimal to serve as a test statistic; more sophisticated tests for a continuous distribution have been extensively discussed (e.g., Hill and Rao, 1977; Antille et al., 1982; McWilliams, 1990). Exploring more powerful tests based on the standardized deviates warrants future study.

Supplementary Materials

The Web Appendix referenced in Section 3 and the R code implementing the simulations in Section 4 and the three case studies in Section 5 are available with this article at the *Biometrics* website on Wiley Online Library.

Acknowledgements

We thank the associate editor and an anonymous reviewer for many constructive comments. We also thank Professor James Hodges for helpful discussion that led to a better presentation of this article. This research was supported in part by NIAID R21 AI103012 (HC, LL), NIDCR R03 DE024750 (HC), NLM R21 LM012197 (HC), NIDDK U01 DK106786 (HC), and the Doctoral Dissertation Fellowship from the University of Minnesota Graduate School (LL).

References

Altman, D. G. (2002). Poor-quality medical research: What can journals do? *Journal of the American Medical Association* **287**, 2765–2767.

Antille, A., Kersting, G., and Zucchini, W. (1982). Testing symmetry. *Journal of the American Statistical Association* **77**, 639–646.

Begg, C. B. and Berlin, J. A. (1988). Publication bias: A problem in interpreting medical data. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* **151**, 419–463.

Begg, C. B. and Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics* **50**, 1088–1101.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., and Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods* **1**, 97–111.

Chalmers, T. C., Celano, P., Sacks, H. S., and Smith, H. (1983). Bias in treatment assignment in controlled clinical trials. *New England Journal of Medicine* **309**, 1358–1361.

Dear, K. B. G. and Begg, C. B. (1992). An approach for assessing publication bias prior to performing a meta-analysis. *Statistical Science* **7**, 237–245.

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials* **7**, 177–188.

Duval, S. and Tweedie, R. (2000a). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association* **95**, 89–98.

Duval, S. and Tweedie, R. (2000b). Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics* **56**, 455–463.

Egger, M., Davey Smith, G., Schneider, M., and Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal* **315**, 629–634.

Gøtzsche, P. C. (1987). Reference bias in reports of drug trials. *British Medical Journal* **295**, 654–656.

Harbord, R. M., Egger, M., and Sterne, J. A. C. (2006). A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Statistics in Medicine* **25**, 3443–3457.

Hedges, L. V. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science* **7**, 246–255.

Higgins, J. P. T. (2008). Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. *International Journal of Epidemiology* **37**, 1158–1160.

Higgins, J. P. T. and Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* **21**, 1539–1558.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., and Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal* **327**, 557–560.

Hill, D. L. and Rao, P. V. (1977). Tests of symmetry based on Cramér–von Mises statistics. *Biometrika* **64**, 489–494.

Hróbjartsson, A. and Gøtzsche, P. C. (2010). Placebo interventions for all clinical conditions. *Cochrane Database of Systematic Reviews* **1**, https://dx.doi.org/10.1002/14651858.CD003974.pub3

Jannot, A.-S., Agoritsas, T., Gayet-Ageron, A., and Perneger, T. V. (2013). Citation bias favoring statistically significant studies was present in medical research. *Journal of Clinical Epidemiology* **66**, 296–301.

Kicinski, M., Springate, D. A., and Kontopantelis, E. (2015). Publication bias in meta-analyses from the Cochrane Database of Systematic Reviews. *Statistics in Medicine* **34**, 2781–2793.

Light, R. J. and Pillemer, D. B. (1984). *Summing Up: The Science of Reviewing Research.* Cambridge, MA: Harvard University Press.

Liu, C. J. and Latham, N. K. (2009). Progressive resistance strength training for improving physical function in older adults. *Cochrane Database of Systematic Reviews* **3**, https://dx.doi.org/10.1002/14651858.CD002759.pub2

Macaskill, P., Walter, S. D., and Irwig, L. (2001). A comparison of methods to detect publication bias in meta-analysis. *Statistics in Medicine* **20**, 641–654.

MacGillivray, H. L. (1986). Skewness and asymmetry: Measures and orderings. *The Annals of Statistics* **14**, 994–1011.

McWilliams, T. P. (1990). A distribution-free test for symmetry based on a runs statistic. *Journal of the American Statistical Association* **85**, 1130–1133.

Normand, S.-L. T. (1999). Meta-analysis: Formulating, evaluating, combining, and reporting. *Statistics in Medicine* **18**, 321–359.

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., and Rushton, L. (2006). Comparison of two methods to detect publication bias in meta-analysis. *Journal of the American Medical Association* **295**, 676–680.

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., and Rushton, L. (2007). Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. *Statistics in Medicine* **26**, 4544–4562.

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., and Rushton, L. (2008). Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *Journal of Clinical Epidemiology* **61**, 991–996.

R Core Team (2016). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Rothstein, H. R., Sutton, A. J., and Borenstein, M. (2005). *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments.* Chichester, UK: John Wiley & Sons.

Rücker, G., Schwarzer, G., and Carpenter, J. (2008). Arcsine test for publication bias in meta-analyses with binary outcomes. *Statistics in Medicine* **27**, 746–763.

Silliman, N. P. (1997a). Hierarchical selection models with applications in meta-analysis. *Journal of the American Statistical Association* **92**, 926–936.

Silliman, N. P. (1997b). Nonparametric classes of weight functions to model publication bias. *Biometrika* **84**, 909–918.

Stead, L. F., Perera, R., Bullen, C., Mant, D., Hartmann-Boyce, J., Cahill, K., et al. (2012). Nicotine replacement therapy for smoking cessation. *Cochrane Database of Systematic Reviews* **11**, https://dx.doi.org/10.1002/14651858.CD000146.pub4

Stern, J. M. and Simes, R. J. (1997). Publication bias: Evidence of delayed publication in a cohort study of clinical research projects. *British Medical Journal* **315**, 640–645.

Sterne, J. A. C. and Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal of Clinical Epidemiology* **54**, 1046–1055.

Sterne, J. A. C., Egger, M., and Davey Smith, G. (2001). Investigating and dealing with publication and other biases in meta-analysis. *British Medical Journal* **323**, 101–105.

Sterne, J. A. C., Gavaghan, D., and Egger, M. (2000). Publication and related bias in meta-analysis: Power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology* **53**, 1119–1129.

Sterne, J. A. C., Sutton, A. J., Ioannidis, J. P. A., Terrin, N., Jones, D. R., Lau, J., et al. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *British Medical Journal* **343**, d4002.

Sutton, A. J., Duval, S. J., Tweedie, R. L., Abrams, K. R., and Jones, D. R. (2000). Empirical assessment of effect of publication bias on meta-analyses. *British Medical Journal* **320**, 1574–1577.

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

Sutton, A. J., Song, F., Gilbody, S. M., and Abrams, K. R. (2000). Modelling publication bias in meta-analysis: A review. *Statistical Methods in Medical Research* **9**, 421–445.

Terrin, N., Schmid, C. H., Lau, J., and Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine* **22**, 2113–2126.

Whitehead, A. and Whitehead, J. (1991). A general parametric approach to the meta-analysis of randomized clinical trials. *Statistics in Medicine* **10**, 1665–1677.

Wright, S. P. (1992). Adjusted p-values for simultaneous inference. *Biometrics* **48**, 1005–1013.

Received April 2016. *Revised October* 2016.
Accepted November 2016.