Original Article

Research Synthesis Methods

Received 15 June 2011, Revised 10 January 2012, Accepted 10 January 2012 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/jrsm.54

On the moments of Cochran's *Q* statistic under the null hypothesis, with application to the meta-analysis of risk difference

Elena Kulinskaya,[a*,†] Michael B. Dollinger[b] and Kirsten Bjørkestøl[c]

W. G. Cochran's *Q* statistic was introduced in 1937 to test for equality of means under heteroscedasticity. Today, the use of *Q* is widespread in tests for homogeneity of effects in meta-analysis, but often these effects (such as risk differences and odds ratios) are not normally distributed. It is common to assume that *Q* follows a chi-square distribution, but it has long been known that this asymptotic distribution for *Q* is not accurate for moderate sample sizes. In this paper, the effect and weight for an individual study may depend on two parameters: the effect and a nuisance parameter. We present expansions for the first two moments of *Q* without any normality assumptions. Our expansions will have wide applicability in testing for homogeneity in meta-analysis. As an important example, we present a homogeneity test when the effects are the differences of risks between treatment and control arms of the several studies—a test which is substantially more accurate than that currently used. In this situation, we approximate the distribution of *Q* with a gamma distribution. We provide the results of simulations to verify the accuracy of our proposal and an example of a meta-analysis of medical data. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: gamma distribution; heterogeneity test; nuisance parameter; weighted analysis of variance; weighted sum of squares

1. Introduction

In a meta-analysis, it is usual to conduct a homogeneity test to determine whether the effects measured by the included studies are sufficiently similar to justify their combination. If the ith study produces an estimate of an underlying effect θ_i, the null hypothesis is that these underlying effects are all equal. The effects can be measured by a variety of statistics, such as sample means, correlation coefficients, and for studies that involve treatment and control arms, differences of sample means, standardized mean differences, odds ratios, and differences or ratios of binomial proportions, known as risk differences and relative risks.

The most commonly used test statistic is Cochran's Q (Cochran, 1937; Cochran, 1954), which calculates a weighted sum of the squared distances of the observed effects from the overall mean effect (see Section 2 for details). The weight for each study is typically chosen to be the inverse of the variance of the estimator of the effect θ_i so that larger or more accurate studies are weighted more heavily. There is no exact analytic expression for the distribution of Q when the weights are treated as unknown, so an approximation must be used. Further, the distribution of Q depends on the effect measure. For example, for sample means, Q will have a different distribution than for risk differences. Nevertheless, most meta-analyses refer Q to the chi-square distribution with $I-1$ degrees of freedom (χ^2_{I-1}), where I is the number of studies. This chi-square distribution appears to be

[a]School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, UK
[b]Department of Mathematics, Pacific Lutheran University, Tacoma, WA, USA
[c]Faculty of Engineering and Science, University of Agder, Kristiansand, Norway
*Correspondence to: Elena Kulinskaya, School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, UK.
†E-mail: e.kulinskaya@uea.ac.uk

asymptotically valid (as the sizes n_i of the studies become large) for many types of effects. But, the approximation is not accurate for small and medium sample sizes—sizes that frequently arise in applications. See the simulation studies by Hedges and Olkin (1985), Viechtbauer (2007), and the references therein for information about the accuracy of the chi-square approximation for the distribution of Q.

A number of results have been published about the distribution of Q for the case in which the effects are normally distributed sample means and the weights are estimated by the inverses of sample variances. Under these normality assumptions, the chi-square distribution is exact if the variances are *known*, resulting in known weights. Randomness of the weights is often ignored in meta-analysis; see Biggerstaff and Tweedie (1997), Jackson (2006), and Biggerstaff and Jackson (2008). In contrast, Cochran, as early as his 1937 paper, which dealt with the normally distributed case, proposed a correction to the chi-square distribution for moderate sample sizes. James (1951) and Welch (1951) separately proposed order $1/n$ corrections to the distribution of Q for the normal case. Welch's proposal (more commonly used and known as the Welch test) refers Q to a rescaled F-distribution with $I-1$ and ν degrees of freedom ($cF_{I-1,\nu}$) where ν and c are estimated from the data. Kulinskaya *et al.* (2004) and Kulinskaya and Dollinger (2007) extended the Welch test to the case of independent estimators of effects and weights to include the differences of sample means and their robust versions.

Kulinskaya *et al*. (2011) dealt with the non-normally distributed case in which both the effect and the weight from an individual study depend on a single parameter, with principal application to the standardized mean difference between treatment and control arms of a study. In the current paper, the effect and weight for an individual study may depend on two parameters: the effect θ_i and a nuisance parameter ζ_i. For example, if effects are means from normally distributed data, then the variances of the distributions are nuisance parameters that must be separately estimated. We present fairly general expansions for the first two moments of Q in terms of the weights and the moments of the estimators of the two parameters. These expansions, which contain both the results of Welch (1951) and the one-parameter results of Kulinskaya *et al.* (2011) as special cases, should have wide applicability.

We apply our expansions to an important special case: the risk difference. In this context, we recommend the use of a gamma distribution, with first and second moments estimated using our general expansions to approximate the null distribution of Q. The resulting homogeneity test is substantially more accurate than current tests, especially when the sample sizes are small or moderate.

The subsequent sections are as follows. Section 2 contains the definitions and assumptions. Section 3 and Appendix A contain the approximations to the moments of the null distribution of Q. Section 4 and Web Appendix B show that Welch's results follow from our expansions. (Sections 3 and 4 are fairly technical and may be skipped by the reader who is primarily interested in the material of Section 5.) Section 5 contains the application to risk differences, including both a real data example and the results of simulations to verify the accuracy of our proposal. Section 6 contains a summary and discussion. The four Web Appendices are described in the Supporting Information.

2. Notation and assumptions

We assume that there are I studies, with effects θ_i. The distribution in the ith study that generates the data involves an additional 'nuisance' parameter, denoted by ζ_i. For example, in our application to risk difference, the effect θ_i of the ith study is the difference of two binomial parameters from treatment and control arms of a study (in this context, we will denote the effect by Δ_i). But, the single parameter Δ_i is not sufficient to describe the distribution of its estimator; a variety of choices are possible for the nuisance parameter ζ_i, and this subject is discussed in Section 5.

The homogeneity test is a test of the hypothesis that $\theta_1 = \cdots = \theta_I$; we denote the common effect under the null hypothesis by θ. The effects are estimated by random variables $\hat{\theta}_i$, and the nuisance parameters are estimated by random variables $\hat{\zeta}_i$. Each study is assigned a weight w_i, which in turn is estimated by a random variable $\hat{w}_i$. In most applications, we will have $w_i = 1/\mathrm{Var}[\hat{\theta}_i]$, thus weighting more accurate studies more heavily, but in this and the next section, we merely assume that the weight estimators are twice-differentiable functions f_i of the estimators of the effects and the nuisance parameters; that is, $\hat{w}_i = f_i(\hat{\theta}_i, \hat{\zeta}_i)$, where the functions f_i will generally depend on additional constants such as the sample sizes. The theoretical weights under the null hypothesis are $w_i = f_i(\theta, \zeta_i)$ (the functions f_i will be different for different effects; examples of such functions can be found in Section 4 for the case of normally distributed sample means and in Section 5.1, Equation (2), for the case in which effects are risk differences). The estimator of the common effect under the null hypothesis of homogeneity is given as a weighted average of the individual effect estimators, $\hat{\theta}_w = \sum_i \hat{w}_i \hat{\theta}_i / \sum_i \hat{w}_i$.

The sample size of the ith study will be denoted by n_i. However, if the studies have two arms (as in the application to the risk difference in Section 5), let n_i be the minimum of the sample sizes of the two arms. Let $n = \min\{n_i\}$; we will sometimes express approximations in terms of orders of n. Throughout this paper, we use the 'big O' notation: for a function $a(n)$ and a fixed integer k, we say that $a(n) = O(n^k)$ (or $a(n)$ is of order n^k) if there is a constant C such that $a(n) \leq Cn^k$ for all sufficiently large n.

Cochran's Q statistic is defined by $Q = \sum_i \hat{w}_i \left(\hat{\theta}_i - \hat{\theta}_w\right)^2$. The main theoretical results of this paper (Section 3 and Appendix A) are series expansions to approximate the mean and the second moment of Q under the null hypothesis in terms of the weights w_i and the first six central moments of the parameter estimators $\hat{\theta}_i$ and $\hat{\zeta}_i$. Cochran (Cochran, 1937) first introduced Q for the case that the estimators of the effects are normally distributed and the numbers of degrees of freedom for their estimated variances are the same. In his 1954 paper (Cochran, 1954), he considered additionally the case of unequal numbers of degrees of freedom.

For normally distributed sample means and weights that are inverse variances, Welch (Welch, 1951) produced expansions for the first two moments of Q, expansions which neglected terms of order $1/n^2$. These expansions are order $1/n$ corrections to the first two moments ($I - 1$ and $I^2 - 1$, respectively) of the chi-square distribution (see Section 4 for more details on Welch's moments). The results of our expansions (Section 3) are also order $1/n$ corrections to the chi-square moments, but these expansions are considerably more complicated because of the absence both of normality assumptions and of the assumption of the independence of the effects and their weights.

Readers who are interested primarily in the behavior of Q for risk difference may proceed directly to Section 5.

3. The moments of Cochran's *Q*-statistic

In the following text, we discuss expansions for the first two moments of Q, without the assumption of normality and without the assumption that the estimators of the effect and nuisance parameters of an individual study are independent. Independence of $\hat{\theta}_i$ and $\hat{\zeta}_i$ holds in the important case in which $\hat{\theta}_i$ is a normally distributed sample mean and $\hat{\zeta}_i$ is the sample variance, but such independence is rare. For example, independence does not hold for the risk difference application, which we present in Section 5. Lack of such independence combined with the dependence of the weight estimators $\hat{w}_i$ on the effect estimators $\hat{\theta}_i$ and the nuisance parameter estimators $\hat{\zeta}_i$ greatly complicates the distribution of Q and our expansions for its first two moments. In general, the expressions for these moments are so complicated that we would expect the necessary calculations to be carried out using a computer program.

For these expansions, we make some fairly standard assumptions about the orders (relative to the sample sizes) of the central moments of $\hat{\theta}_i$ and $\hat{\zeta}_i$ and also about the orders of the weights and their derivatives. Define $\Theta_i = (\hat{\theta}_i - \theta_i)$ and $Z_i = (\hat{\zeta}_i - \zeta_i)$. We assume first that $\mathrm{E}[\Theta_i] = O\left(1/n_i^2\right)$ and $\mathrm{E}[Z_i] = O\left(1/n_i^2\right)$. These conditions will certainly be satisfied if the estimators $\hat{\theta}_i$ and $\hat{\zeta}_i$ are unbiased. In regular parametric problems, it is easy to remove the first-order term from the asymptotic bias of maximum likelihood estimators (see Firth, 1993). We also assume the following orders, which generally follow from $\sqrt{n_i}$ asymptotic normality: $\mathrm{E}[\Theta_i^2] = O\left(1/n_i\right)$, $\mathrm{E}[\Theta_i^3] = O\left(1/n_i^2\right)$, $\mathrm{E}[\Theta_i^4] = O\left(1/n_i^2\right)$, $\mathrm{E}[\Theta_i^5] = O\left(1/n_i^3\right)$, and $\mathrm{E}[\Theta_i^6] = O\left(1/n_i^3\right)$ and similar conditions for the moments of Z_i. These order conditions are not restrictive; they are satisfied for a wide range of effect measures, including standardized mean differences, binomial and Poisson rates, and bias corrected correlation coefficients (Kulinskaya *et al.*, 2011). Order conditions that we need for mixed moments (such as for $\mathrm{E}[\Theta_i Z_i]$ and which are needed because independence is not assumed) follow from the given order conditions and the Cauchy–Schwarz inequality. We further assume that the weight estimators $\hat{w}_i$ and their first two derivatives with respect to $\hat{\theta}_i$ and $\hat{\zeta}_i$ will be $O(n_i)$, as will happen whenever the weights are inverses of the variances. With these assumptions, our expansions for $\mathrm{E}[Q]$ and $\mathrm{E}[Q^2]$ should be accurate to order $1/n$.

The derivation of the expansions is a straightforward application of the delta method in which Q (respectively Q^2) is first expanded in a multivariate Taylor series centered at the null hypothesis, and then expectations are taken of the resulting expansion, keeping only those terms of order $O(1)$ and $O(1/n)$. For example, for the Taylor expansion of Q, we consider Q as a function of the estimators of the effect and nuisance parameters as follows:

$$Q = \sum_i \hat{w}_i \left(\hat{\theta}_i - \hat{\theta}_w\right)^2 = Q\left[\hat{\theta}_1, \ldots, \hat{\theta}_I, \hat{w}_1, \ldots, \hat{w}_I\right] = Q\left[\hat{\theta}_1, \ldots, \hat{\theta}_I, f_1\left(\hat{\theta}_1, \hat{\zeta}_1\right), \ldots, f_I\left(\hat{\theta}_I, \hat{\zeta}_I\right)\right].$$

Under the null hypothesis, all the effect parameter values are equal, that is, $\theta_1 = \cdots = \theta_I$, and we denote this common value by θ. The desired Taylor expansion of Q is centered at $\vec{\theta} := (\theta, \ldots, \theta, \zeta_1, \ldots, \zeta_I)$. Details and additional explanation of the derivation are included in Web Appendix A (see Supporting Information). The result for $\mathrm{E}[Q]$ is given in the following text, and the result for $\mathrm{E}[Q^2]$ is given in Appendix A.1, Equation (3).

$$
\begin{aligned}
\mathrm{E}[Q] = {} & \frac{1}{2}\frac{\sum_i \partial^2 Q(\vec{\theta})}{\partial\hat{\theta}_i^2}\,\mathrm{E}[\Theta_i^2] + \frac{1}{6}\sum_i \frac{\partial^3 Q\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^3}\,\mathrm{E}[\Theta_i^3] + \frac{1}{2}\sum_i \frac{\partial^3 Q\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^2\partial\hat{\zeta}_i}\mathrm{E}[\Theta_i^2 Z_i] \\
& + \frac{1}{24}\frac{\sum_i \partial^4 Q(\vec{\theta})}{\partial\hat{\theta}_i^4}\,\mathrm{E}[\Theta_i^4] + \frac{1}{6}\sum_i \frac{\partial^4 Q\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^3\partial\hat{\zeta}_i}\,\mathrm{E}[\Theta_i^3 Z_i] + \frac{1}{4}\sum_i \frac{\partial^4 Q\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^2\partial\hat{\zeta}_i^2}\mathrm{E}[\Theta_i^2 Z_i^2] \\
& + \frac{1}{8}\sum_{i\neq j}\sum \frac{\partial^4 Q\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^2\partial\hat{\theta}_j^2}\mathrm{E}[\Theta_i^2]\mathrm{E}\left[\Theta_j^2\right] + \frac{1}{2}\sum_{i\neq j}\sum \frac{\partial^4 Q\left(\vec{\theta}\right)}{\partial\hat{\theta}_i\partial\hat{\zeta}_i\partial\hat{\theta}_j\partial\hat{\zeta}_j}\mathrm{E}[\Theta_i Z_i]\mathrm{E}\left[\Theta_j Z_j\right] \\
& + \frac{1}{2}\sum_{i\neq j}\sum \frac{\partial^4 Q\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^2\partial\hat{\theta}_j\partial\hat{\zeta}_j}\mathrm{E}[\Theta_i^2]\mathrm{E}[\Theta_j Z_j] + \frac{1}{4}\sum_{i\neq j}\sum \frac{\partial^4 Q\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^2\partial\hat{\zeta}_j^2}\mathrm{E}[\Theta_i^2]\mathrm{E}[Z_j^2] + O\left(\frac{1}{n^2}\right).
\end{aligned} \tag{1}
$$

The derivatives needed for the aforementioned expansions can be quite complicated. A list of these, using the notation $W = \sum w_i$ and $U_i = 1 - w_i/W$, appears in Appendix A.2.

For a better understanding of our expansions, it is useful to compare the results of these expansions with the chi-square moments, which are $I-1$ for the first moment and I^2-1 for the second moment. We begin with the first moment. The first term in Equation (1) (the quadratic term) is $O(1)$; the remaining terms are $O(1/n)$. If weights are taken (as usual) to be inverse variances, then $\mathrm{E}[\Theta_i^2] = 1/w_i$, and using $\partial^2 Q(\vec{\theta})/\partial\hat{\theta}_i^2 = 2w_i U_i$ from Appendix A.2, the first term becomes $\sum U_i = \sum(1 - w_i/W) = I - 1$, which is the chi-square first moment. Thus, we see that the additional terms of Equation (1) give an order $1/n$ correction to the chi-square first moment.

To compare the leading terms of the second moment of Q (Equation (3), from Appendix A.1) to the second moment of the chi-square distribution, we define the coefficient of kurtosis as $\gamma_i = \mathrm{E}[\Theta_i^4]/\left(\mathrm{E}[\Theta_i^2]\right)^2 - 3$. The first two terms of Equation (3) (i.e., those of fourth degree) are the only terms that are $O(1)$; the remaining terms are $O(1/n)$. If weights are taken to be inverse variances, then (using the relevant derivatives from Appendix A.3) the first two terms simplify as follows:

$$
\begin{aligned}
& \frac{1}{24}\sum_i \frac{\partial^4 Q^2\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^4}\;\mathrm{E}[\Theta_i^4] + \frac{1}{8}\sum_{i\neq j}\sum \frac{\partial^4 Q^2\left(\vec{\theta}\right)}{\partial\widehat{\theta}_i^2\partial\widehat{\theta}_j^2}\;\mathrm{E}[\Theta_i^2]\mathrm{E}[\Theta_j^2] \\
& \quad = \sum_i (3+\gamma_i)U_i^2 + \sum_{i\neq j}\sum\left(U_i U_j + 2w_i w_j/W^2\right).
\end{aligned}
$$

Expanding the sums and collecting terms yield $\mathrm{E}[Q^2] = I^2 - 1 + \sum_i \gamma_i U_i^2 + O(1/n)$. Because U_i are $O(1)$, the leading terms of our expansion agree with the chi-square second moment of I^2-1 to order $O(1/n)$ if and only if the kurtosis coefficients γ_i are $O(1/n)$. In the examples that we have considered $\gamma_i = O(1/n)$, but we do not know whether this fact will generally follow from the moment assumptions of Section 2.

4. The Welch moments of Q

As a first application of Equations (1) and (3), we derive Welch's moments when the effects are means from normal populations, estimated by sample means, and the weights are inverses of the variances of the sample means, estimated by the sample variances. In this situation, Welch (Welch, 1951) obtained the following moments of Q under the null hypothesis:

$$
\mathrm{E}[Q] = I - 1 + 2\sum_i\left[U_i^2/(n_i-1)\right] + O\left(1/n^2\right)
$$

and

$$
\mathrm{E}[Q^2] = I^2 - 1 + (4I+10)\;\sum_i\left[U_i^2/(n_i-1)\right] + O\left(1/n^2\right).
$$

We take $X_i \sim N(\theta_i, \sigma_i^2)$, so in our notation, we have $\widehat{\theta}_i = \bar{X}_i$ and $\Theta_i = \bar{X}_i - \theta_i$. For the nuisance random variables, we may take the sample variances; thus, $\hat{\zeta}_i = S_i^2$. We have $\mathrm{E}[\hat{\zeta}_i] = \sigma_i^2$, so $Z_i = S_i^2 - \sigma_i^2$. Our formula for $\mathrm{E}[Q]$ requires moments for Θ_i and Z_i. The moments are $\mathrm{E}[\Theta_i] = 0$ and $\mathrm{E}[\Theta_i^2] = \mathrm{Var}[\bar{X}_i] = \sigma_i^2/n_i$; also, $\mathrm{E}[Z_i] = 0$, and $\mathrm{E}[Z_i^2] = \mathrm{Var}[S_i^2] = 2\sigma_i^4/(n_i-1)$. Further, because Θ_i and Z_i are independent, product moments such as $\mathrm{E}[\Theta_i Z_i]$ are readily calculated.

The inverse variance weights are given by $w_i = n_i/\sigma_i^2$, so we have $\hat{w}_i = f_i(\hat{\theta}_i, \hat{\zeta}_i) = n_i/\hat{\zeta}_i$. Because the weight estimators do not depend on the $\hat{\theta}$s, all derivatives of the weight functions f_i with respect to the $\hat{\theta}$s are zero. For derivatives with respect to the $\hat{\zeta}$s (evaluated at the null hypothesis), we have $\partial f_i/\partial\hat{\zeta}_i = -n_i/\zeta_i^2 = -w_i^2/n_i$ and $\partial^2 f_i/\partial\hat{\zeta}_i 2 = 2n_i/\zeta_i^3 = 2w_i^3/n_i^2$.

Examination of Equation (1) together with the list of derivatives in Appendix A.2 shows that seven of the ten terms in Equation (1) are zero because either the expectations are zero or the weight derivatives are zero. The non-zero terms are the quadratic term involving $\mathrm{E}[\Theta_i^2]$ and the fourth-degree terms involving $\mathrm{E}[\Theta_i^2 Z_i^2]$ and $\mathrm{E}[\Theta_i^2]\mathrm{E}[Z_j^2]$. Thus, we have

$$\mathrm{E}[Q] = \frac{1}{2}\sum_i \frac{\partial^2 Q(\vec{\theta})}{\partial \hat{\theta}_i^2}\,\mathrm{E}[\Theta_i^2] + \frac{1}{4}\sum_i \frac{\partial^4 Q(\vec{\theta})}{\partial \hat{\theta}_i^2 \partial \hat{\zeta}_i^2}\mathrm{E}[\Theta_i^2 Z_i^2] + \frac{1}{4}\sum_{i \neq j} \frac{\sum \partial^4 Q(\vec{\theta})}{\partial \hat{\theta}_i^2 \partial \hat{\zeta}_j^2}\mathrm{E}[\Theta_i^2]\mathrm{E}[Z_j^2] + O\left(\frac{1}{n^2}\right).$$

The remaining algebra to show that this expression reduces to the Welch first moment is carried out in Web Appendix B.1.

Examination of the second moment approximation given by Equation (3) and the derivatives in Appendix A.3 shows that, as for the first moment, many of the terms of Equation (3) are zero because of either zero moments or zero weight derivatives. The six non-zero terms are those involving the following expectations: $\mathrm{E}[\Theta_i^4]$, $\mathrm{E}[\Theta_i^2]\mathrm{E}[\Theta_j^2]$, $\mathrm{E}\left[\Theta_i^4 Z_i^2\right]$, $\mathrm{E}\left[\Theta_i^4\right]\mathrm{E}\left[Z_j^2\right]$, $\mathrm{E}\left[\Theta_i^2 Z_i^2\right]\mathrm{E}\left[\Theta_j^2\right]$, and $\mathrm{E}\left[Z_i^2\right]\mathrm{E}\left[\Theta_j^2\right]\mathrm{E}\left[\Theta_k^2\right]$. As explained in Section 3, the two fourth-degree terms simplify to the chi-square second moment $I^2 - 1$, and the four sixth-degree terms yield Welch's $O(1/n)$ correction. Web Appendix B.2 gives the details of the calculations needed to verify this claim. Thus we actually need the sixth-degree Taylor expansion for $\mathrm{E}[Q^2]$ to obtain the necessary accuracy, even in the case of normally distributed sample means.

We remark that knowledge of the first two moments of Q under the null hypothesis is not sufficient to conduct a homogeneity test. An approximate distribution must be fit to Q by using these moments. Welch fit a rescaled F-distribution.

5. Risk difference

As a second application of our moment expansions, we consider the situation in which the effects are the differences of risks between two arms of the studies. (See the following text for definitions and notation for the risk difference.) Testing for the homogeneity of risk differences has many applications to medical data, including the meta-analysis of studies with binomial outcomes and the analysis of multi-center clinical trials. The use of the Q statistic with the χ^2_{I-1} distribution for testing for homogeneity in this situation has been suggested, for example, by Fleiss (1981), DerSimonian and Laird (1986), and Deeks and Higgins (2005) in spite of the fact that it is too liberal (i.e., it rejects the null hypothesis too often). In contrast, Lipsitz *et al*. (1998), Lui and Kelly (2000), Böhning and Sarol (2000), and Kelly *et al*. (2005), primarily focusing on small samples, have suggested a number of modifications including transformations of Q and the use of weights different from inverse variance weights. The approach suggested here is to use the more accurate moments of Q from Section 3 and to refer Q to a gamma distribution with shape parameter α and scale parameter β; in terms of the mean μ and variance σ^2 of the gamma distribution, the parameters are $\beta = \sigma^2/\mu$ and $\alpha = \mu/\beta$. Thus, estimates of these parameters are calculated from the estimated moments of Q by the relations

$$\hat{\beta} = \left(\mathrm{E}\left[Q^2\right] - \mathrm{E}[Q]^2\right)/\mathrm{E}[Q] \text{ and } \hat{\alpha} = \mathrm{E}[Q]/\hat{\beta}.$$

The gamma distribution was selected after extensive simulations and the determination that it provides an excellent fit for the distribution for Q; moreover, the gamma family includes the chi-square distributions as special cases. The details of the application of our methods to the risk difference are presented below, but the actual calculations are best performed with the use of our program in *R* (see Supporting Information).

5.1. Notation and formulas for risk difference

In this section, we simplify notation by omitting the subscript i for the ith study. Each study has two arms: treatment and control (denoted by the subscripts T and C). The sizes of the arms are n_T and n_C. Let $N = n_T + n_C$, and let $q = n_C/N$ be the proportion of the study in the control arm. The probabilities of success in the arms are given by p_T and p_C and will be estimated by the proportions of successes $\hat{p}_T$ and $\hat{p}_C$. The effect of interest is the risk difference $\Delta = p_T - p_C$, which is estimated by $\hat{\Delta} = \hat{p}_T - \hat{p}_C$. Because a given value of Δ may arise from various values of the parameters p_T and p_C, a nuisance parameter must be introduced to uniquely specify a distribution for $\hat{\Delta}$. We have adopted the average of the underlying parameters p_T and p_C weighted by the corresponding sample sizes: $\zeta = (1-q)p_T + qp_C$ with the corresponding estimator $\hat{\zeta} = (1-q)\widehat{p}_T + q\hat{p}_C$. We will call ζ the overall risk of the study. The estimators $\hat{\Delta}$ and $\hat{\zeta}$ are not independent, in contrast to situations involving sample means and variances of Gaussian data. We note that other choices of a nuisance parameter are possible. We have not studied the extent to which a different choice of nuisance parameter will affect the estimated moments of Q; this issue is one for future investigation.

We take the weight function to be the inverse variance of $\hat{\Delta}$. It is given by

$$w = \left[\frac{p_T(1-p_T)}{(1-q)N} + \frac{p_C(1-p_C)}{qN}\right]^{-1} = \frac{q(1-q)N}{\zeta - \zeta^2 - (1-2q)\Delta(1-2\zeta) - (1-3q+3q^2)\Delta^2} := f(\Delta, \zeta). \tag{2}$$

The weight is estimated by replacing Δ and ζ by $\hat{\Delta}$ and $\hat{\zeta}$, respectively, in the aforementioned expression. The derivatives of $\hat{w}$ are given in Web Appendix C.2. When zeros are observed in both arms of a study (an event occurring with a probability less than $O(1/n)$), 0.5 is added to all cells in the 2×2 array for that study.

The moments of $\Theta = \hat{\Delta} - \Delta$ and $Z = \hat{\zeta} - \zeta$ necessary to apply Equations (1) and (3) can be found in Web Appendix C.1. It is readily seen from an examination of these moments that the order conditions of Section 3 are valid. Also, a straightforward calculation using the second and fourth moments shows that the kurtosis coefficient for Θ is of order $O(1/n)$. Thus (cf. end of Section 3), the leading terms of our expansion in Equation (3) for the second moment of Q for the risk difference agree with the chi-square second moment $I^2 - 1$.

5.2. Example: a meta-analysis of Crowley

This section illustrates the theory of Sections 3 and 5.1 and gives an indication of the improvement in accuracy of the homogeneity test. The calculations can be performed using our computer program (see Supporting Information).

We use the data from the review by Crowley (Crowley, 2000) of clinical trials on the use of corticosteroids prior to pre-term delivery to prevent respiratory distress syndrome, a serious complication of prematurity that can cause immediate and long-term morbidity and even mortality. Comparison number 01.01.03 of Crowley (2000) considered the subset of babies born at less than 30 weeks gestation. We excluded two studies with three or fewer babies in the treatment arm. Summary data and the results from the standard analysis are found in Figure 1, produced by the R package `meta` (Schwarzer, 2010).

The weights in the last two columns of the figure are given as percentages for ease of comparison. The actual weights needed for the Q statistic can be recovered by using the weight total, $W = 458.07$. The weighted average of the risk differences is $\hat{\Delta}_w = -0.130$. The value of Cochran's Q statistic is 9.53. The standard χ^2_4 approximation yields the p-value of 0.049 for the test for homogeneity.

To use our expansions, the weights need to be recalculated under the null hypothesis of equal risk differences. We take the null value (as found earlier) to be $\hat{\Delta}_w = -0.130$ for each of the five studies, keep the individual overall risks $\zeta_i = \hat{\zeta}_i$ fixed, and recalculate the weights by using Equation (2). Then, the estimated moments of the null distribution of Q are $E[Q] = 4.39$ and $E[Q^2] = 29.78$. The parameters of the approximating gamma distribution are $\alpha = 1.84$ and $\beta = 2.39$. The p-value corresponding to the original observed value of $Q = 9.53$ for this gamma distribution is $p = 0.076$.

To assess the accuracy of the two approximations (χ^2_4 and $\Gamma_{1.84,\,2.39}$) to the null distribution of Q, we conducted a simulation of 100,000 random samples with five studies having the same sizes and the same overall risks ζ_i as in this example but with all studies having the null value of the risk difference $\Delta = -0.130$. A comparison of the simulations with the two approximations is contained in Table 1. As can be seen, the gamma approximation is clearly more accurate.

5.3. Simulations for risk difference

Our simulations were designed to investigate the accuracy of the moment formulas of Section 3 and the accuracy of the proposed hypothesis test under a variety of conditions. Each simulation pattern to be described was repeated 10,000 times.

The values of the various parameters in the simulations are as follows. First, for equal studies (i.e., each study has the same size N, the same proportion q in the control arm, and the same overall risk ζ), we take $\Delta = 0.1$ with

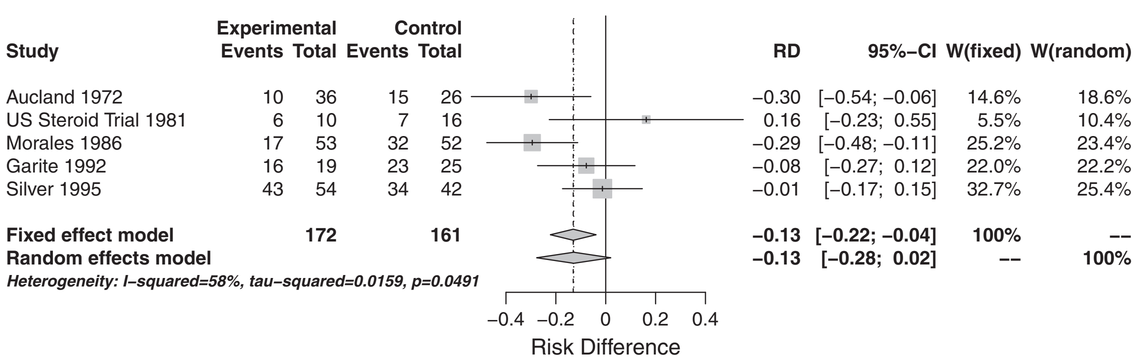

Figure 1. Data and standard meta-analysis for the Crowley (2000) systematic review of the use of corticosteroids prior to pre-term delivery.

Table 1. Comparison of two approximations used in the analysis of the Crowley data with the simulations based on that data.

	χ^2_4	$\Gamma_{1.84,2.39}$	Simulation
First moment	4	4.39	4.49
Second moment	24	29.8	32.7
p-value	0.049	0.076	0.081

corresponding $\zeta=\{0.15, 0.35, 0.5, 0.65, 0.85\}$ and $\Delta=0.3$ with corresponding $\zeta=\{0.25, 0.35, 0.5, 0.65, 0.75\}$, the number of studies $I=\{5, 10, 20, 40\}$, the total size of the studies $N=\{60, 120, 180\}$, and the proportion of the observations in the control arm $q=\{1/3, 1/2\}$. Second, we consider equal-sized studies with parameters as earlier but with unequal values of ζ. Here, each simulation uses all the values of ζ as earlier with each value appearing once when $I=5$, twice when $I=10$, and so on. In this set of simulations, we include both values of Δ and both values of q.

Finally, for unequal study sizes, we follow a suggestion of Sánchez-Meca and Marýn-Martýnez (2000), who selected study sizes having the skewness of 1.464, which they consider typical for meta-analyses in behavioral and health sciences. For $I=5$ studies, the average total size $\bar{N}$ and the individual study sizes (in parentheses) are the following: 60 (24, 32, 36, 40, 168), 100 (64, 72, 76, 80, 208), and 160 (124, 132, 136, 140, 268). For $I=10$, 20, and 40, the vector of individual sizes is repeated 2, 4, and 8 times, respectively. In these simulations, we also vary the values of ζ, by repeating each of the aforementioned values $I/5$ times; for example, for $\Delta=0.1$ and $I=10$, the values of ζ would be $(0.15, 0.15, 0.35, 0.35, \ldots)$. Again, we include both values of Δ and both values of q.

In our report of the simulation results in the following text, we include only $\Delta=0.1$ and test level 0.05. The corresponding results for $\Delta=0.3$ can be found in Web Appendix D. Table 2 provides a quick summary (without details) of the simulation results, with focus on small study sizes—sizes for which the gamma approximation is most superior to the chi-square approximation. Figures 2–5 provide more detailed information about the simulations in a graphical format.

5.3.1. Results for equal studies. The first conclusion from the simulations is that the moments of Q are always substantially larger than the corresponding moments of χ^2_{I-1}. For a given simulation pattern, denote by $\bar{Q}$ and $\overline{Q^2}$, the average of the values of Q and Q^2, respectively, from the 10,000 repetitions. We take these values to be the 'true' values of the moments of Q. Denote by E1 and E2 the values of the approximate moments given by Equations (1) and (3) calculated for known values of the parameters. Define the percent error in our first-moment approximation by $100(\mathrm{E1}-\bar{Q})/\bar{Q}$ and similarly for the second moment. Similarly, the percent error in the chi-square approximation is given by replacing E1 with $I-1$ and E2 with I^2-1 in those expressions. Use of relative error, rather than absolute error, allows us to compare the errors in the approximations over different values of I.

Figure 2 displays the relative error in the approximation to the first moment for the case in which all studies are of equal size and have equal overall risk ζ. Figure 3 displays the relative error in the approximations to the second moment.

The approximation E1 is remarkably accurate, especially when the arms are balanced. Even for the smallest study size, the error (in absolute value) is never more than 1.6%, independent of the number of studies, and it improves as

Table 2. Summary comparison of the χ^2_I and the $\Gamma_{\alpha,\beta}$ approximations to the null distribution of Q. The comparisons in the table focus on small study sizes (total in the two arms of 60) and on moderate overall risk ($0.35\le\zeta\le0.65$). Test levels are based on a nominal level of 0.05. For larger study sizes, both approximations improve, with $\Gamma_{\alpha,\beta}$ continuing to be more accurate than χ^2_I.

Summary of typical simulation results
First moment
χ^2_I moment is always very low with relative error in the range (−25%, −7%).
$\Gamma_{\alpha,\beta}$ moment is very accurate with relative error in the range (−2%, +2%).
Second moment
χ^2_I moment is always very low with relative error in the range (−30%, −15%).
$\Gamma_{\alpha,\beta}$ moment is a bit low with relative error in the range (−7%, 0%).
Level (proportion of simulations which reject the null hypothesis)
χ^2_I level is always too high with levels typically in the range (0.07, 0.15).
$\Gamma_{\alpha,\beta}$ level is always more accurate, typically in the range (0.045, 0.07).
Warning
When overall risk ζ is low ($\le$0.15) or high ($\ge$0.85) and one of the corresponding study arms is small ($\le$20), the test based on $\Gamma_{\alpha,\beta}$ is usually superior to that based on χ^2_I, but neither test provides accurate levels.

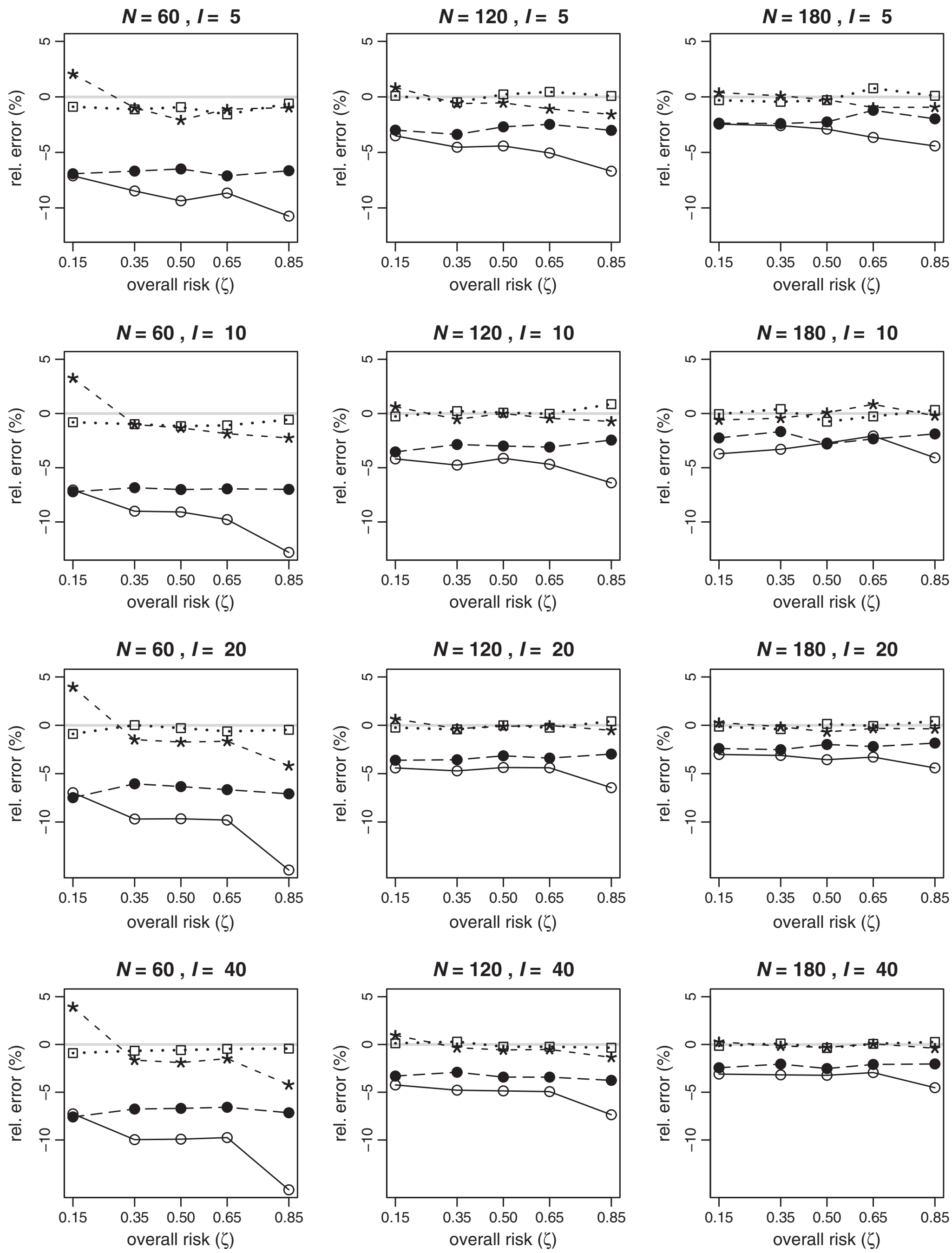

Figure 2. Relative error in two approximations of the mean of Q, for equal studies and risk difference $\Delta = 0.1$. The data are represented as follows: squares (Equation (1), balanced arms), asterisks (Equation (1), unbalanced arms), solid circles (chi-square, balanced arms), and open circles (chi-square, unbalanced arms). Here, I is the number of studies, and N is the total size of the two arms of each study.

the study size increases. When the arms are unbalanced, the error in E1 is less than 2% except when ζ is very small (0.15) or very large (0.85), where the error reaches a maximum of about 4.2%. In contrast, for balanced arms and for the smallest study size, the error in the chi-square approximation is always more negative than about -7%. For unbalanced arms, the error worsens to more than -15%. Even in the optimal situation (balanced arms and large study sizes), the error in the chi-square approximation remains more negative than -2%.

The approximation E2 for the second moment is also quite good but not nearly as accurate as the corresponding approximation for the first moment (Figure 3). In contrast, the chi-square second moment always substantially underestimates the empirical second moment. For balanced arms, the error in E2 for all of the study sizes lies in the interval $[-3.5\%, +2.2\%]$. For unbalanced arms, the error range is $[-10\%, +12\%]$ for the smallest study size but improves substantially for larger study sizes. In contrast, for the smallest study sizes and balanced arms, the error in the chi-square approximation lies in the interval $[-15\%, -12\%]$, whereas for unbalanced arms, the interval is $[-30\%, -15\%]$.

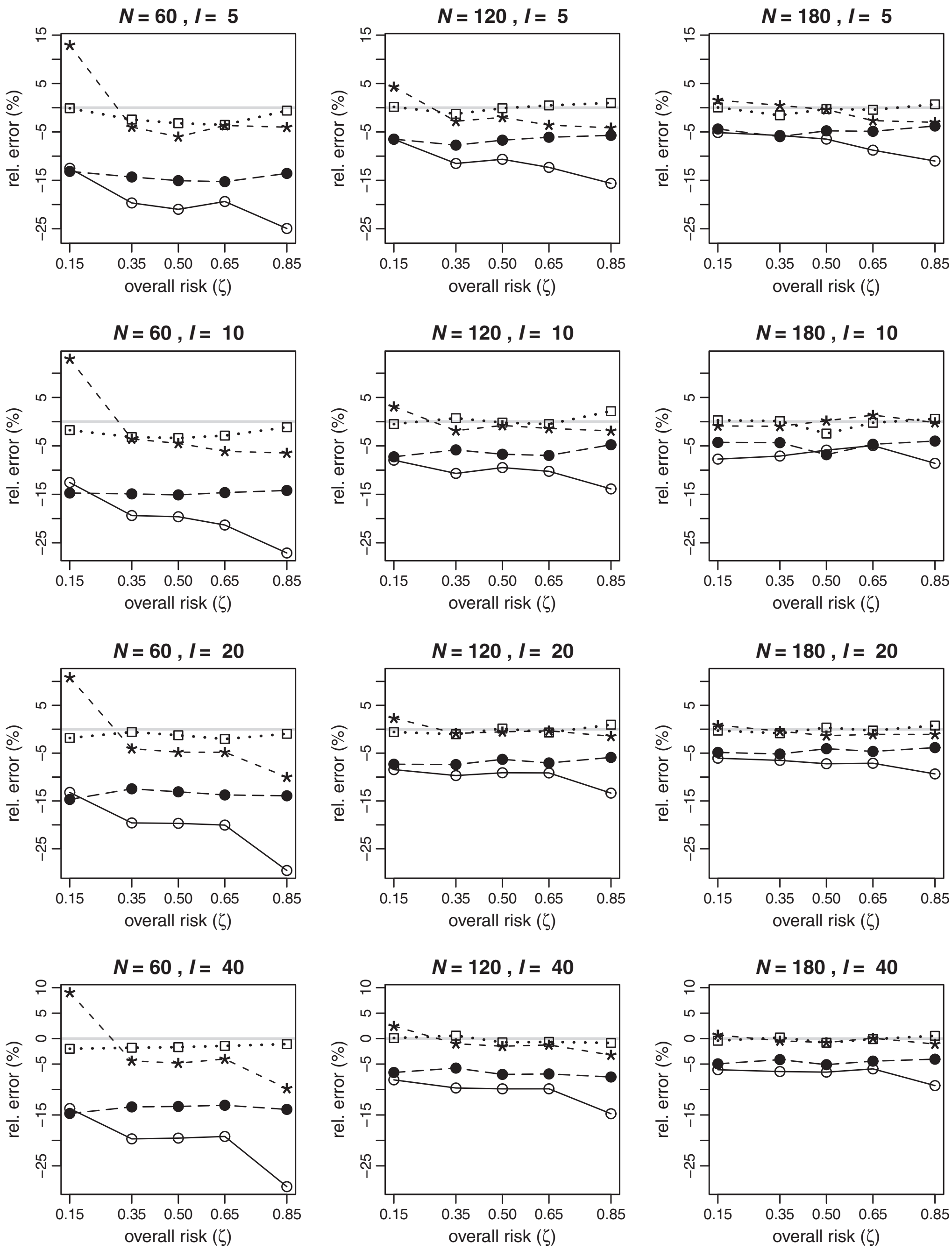

Figure 3. Relative error in two approximations of the second moment of Q, for equal studies and risk difference $\Delta=0.1$. The data are represented as follows: squares (Equation (3), balanced arms), asterisks (Equation (3), unbalanced arms), solid circles (chi-square, balanced arms), and open circles (chi-square, unbalanced arms). Here, I is the number of studies and, N is the total size of the two arms of each study.

Figure 4 displays the level of the homogeneity test on the basis of a gamma distribution whose parameters are estimated via the moment expansions. The values of the achieved levels give the proportion of the simulations for which the null hypothesis was rejected. We see that the levels for the gamma approximation are always superior to the levels by using the chi-square approximation, which are always too high—meaning that the chi-square test is always too liberal.

For the balanced arm simulations, the gamma levels are quite close to the nominal level of 0.05. Even for the smallest study size, the range of achieved levels is in the interval [0.046, 0.062]. Further, the departure from the nominal level does not seem to increase with the number of studies I, in contrast to such an increase for the chi-square approximation. For the larger study sizes, the gamma approximation yields achieved levels that are quite close to the nominal level (on the basis of 10,000 repetitions, a 90% confidence interval for the achieved level is approximately ± 0.004).

A similar pattern obtains for unbalanced arms, but there are some differences. Here, we need to describe the behavior of the approximations separately for moderate values of ζ (0.35, 0.5, and 0.65) and for the more

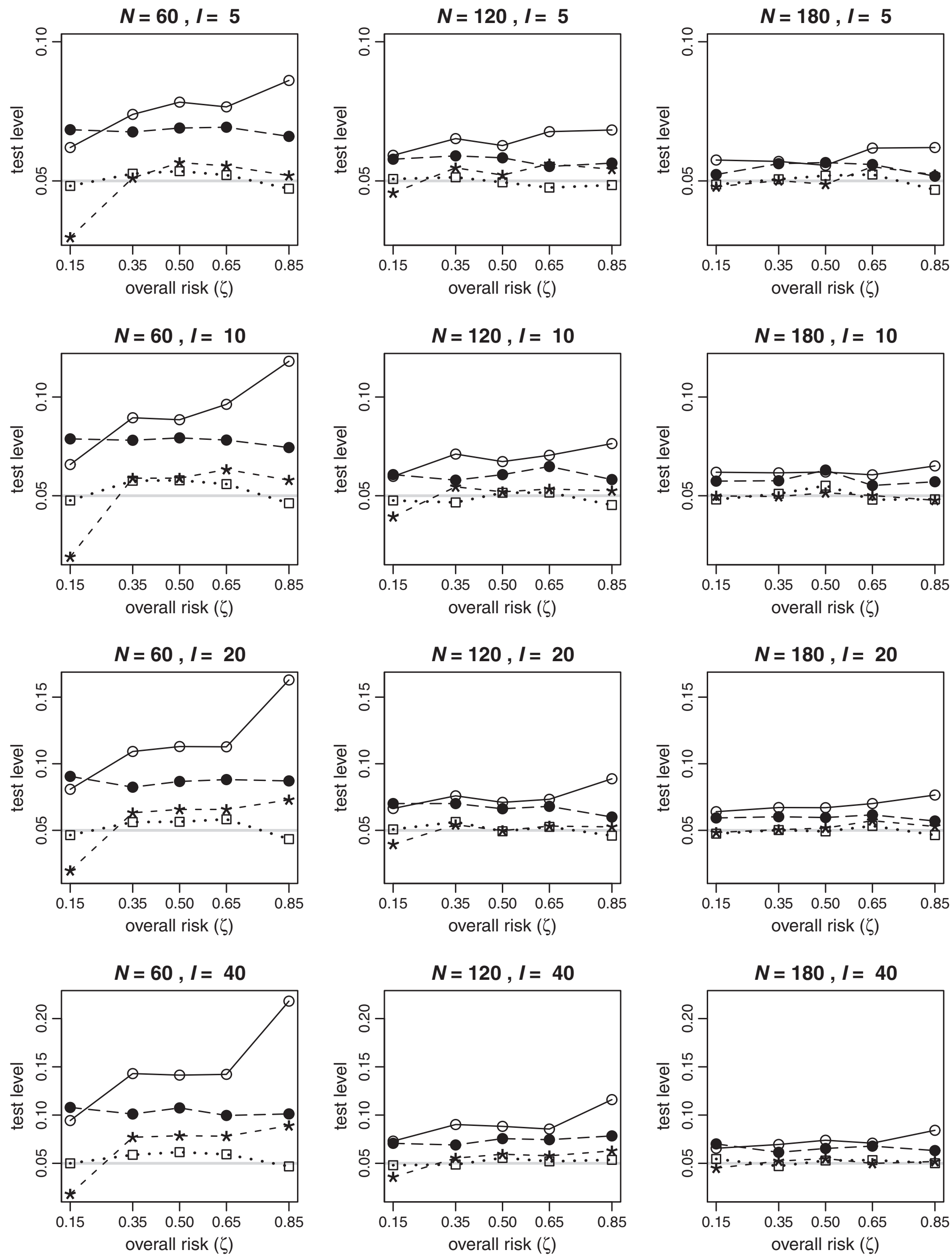

Figure 4. Achieved levels at the nominal level of 0.05 of the Q test for equal studies and risk difference $\Delta = 0.1$ using two approximations to the distribution of Q. The data are represented as follows: squares (gamma, balanced arms), asterisks (gamma, unbalanced arms), solid circles (chi-square, balanced arms), and open circles (chi-square, unbalanced arms). Here, I is the number of studies, and N is the total sample size of the two arms of each study.

extreme values of ζ (0.15 and 0.85), especially for $N = 60$. For the moderate values of ζ, accuracy worsens as the number of studies increases; for example, for $\zeta = 0.5$ and I increasing from 5 to 40, the gamma levels are 0.057, 0.059, 0.066, and 0.077, whereas the corresponding chi-square levels are 0.078, 0.088, 0.113, and 0.141. For $\zeta = 0.15$, $\Delta = 0.1$, and $q = 1/3$, the treatment and control probabilities are $p_T = 0.183$ and $p_C = 0.083$; the achieved levels of the gamma approximation in this situation are too small: 0.030, 0.019, 0.020, and 0.018 as I increases from 5 to 40. At the other extreme value of $\zeta = 0.85$, we have $p_T = 0.883$ and $p_C = 0.783$; here, the gamma levels are greater than the nominal level of 0.05, rising from 0.052 to 0.089 (as I increases from 5 to 40). However, in this latter situation, the chi-square levels are much too high, rising from 0.086 (for $I = 5$) to 0.218 (for $I = 40$). Thus, we may conclude that for unbalanced arms, the gamma approximation is typically superior to the chi-square approximation. However, caution is needed in applying the homogeneity test when both sample sizes and probabilities are very small because in these cases the test based on either of the approximations is inaccurate.

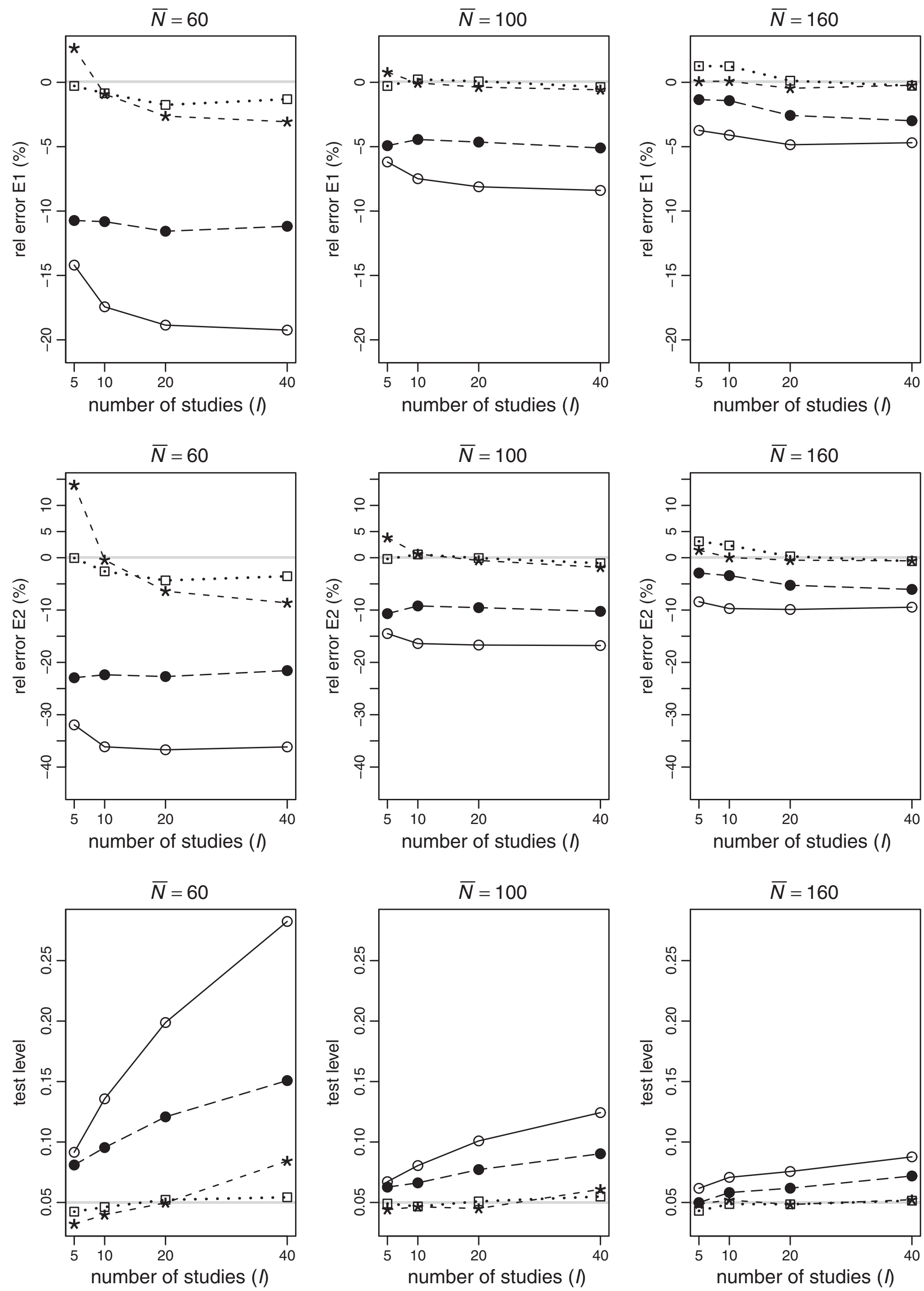

Figure 5. Comparison, for studies of unequal sizes and overall risks described in Section 5.3 , of moment approximations and test levels on the basis of two approximations of the distribution of Q. Here, $\bar{N}$ is the average, across studies, of the total sample sizes of the two arms of each study, the risk difference is $\Delta = 0.1$, and the nominal test level is 0.05. The data are represented as follows: squares (our approximations, balanced arms), asterisks (our approximations, unbalanced arms), solid circles (chi-square, balanced arms), and open circles (chi-square, unbalanced arms).

5.3.2. Results for unequal studies. Because of the exceedingly large number of variables, our simulations for the unequal study size situation (as described earlier) must be regarded as illustrative rather than systematic. Because each simulation pattern involves multiple values of ζ, Figure 5 presents the results as functions of the number of studies I, and each indicated study size is the average of the individual study sizes for the given simulation pattern. We have not included a graphical display of the simulations in which the studies all had equal size but the values of ζ varied among the studies; the results from those simulations are similar to the results displayed in Figure 5 but are somewhat better.

For the unequal studies, we again see that the achieved levels based on the gamma approximation and the moment approximations themselves are always substantially superior to the corresponding quantities given by the chi-square approximation. For the simulations in which the arms are balanced, the achieved levels are excellent with values always in the interval [0.042, 0.055], whereas the corresponding chi-square levels range from 0.08 to 0.15 for the smaller study sizes ($\bar{N} = 60$) and from 0.06 to 0.09 for the intermediate study sizes ($\bar{N} = 100$).

For the simulations with unbalanced arms, the results are not as good as the results for the balanced arms (as is the case for the equal study size simulations), with (as described in the previous section) the greatest difficulty occurring when small studies are combined with small probabilities. Nevertheless, even for the small sample sizes ($\bar{N} = 60$), the achieved levels lie between 0.03 and 0.085, in comparison with the corresponding chi-square levels, which lie between 0.09 and 0.28.

6. Summary and discussion

The simplicity of the standard chi-square approximation for the null distribution of Cochran's Q statistic has made its use common in meta-analytic homogeneity tests. However, as Cochran (Cochran, 1937; Cochran, 1954) pointed out, this approximation is not accurate for small or moderate study sizes, so it requires a 'correction'. The situation is complicated by the fact that different corrections are needed for different types of effects. When the effects are normally distributed sample means, Welch's correction (Welch, 1951) is known to be accurate and should be used. A correction for use when the effects are standardized mean differences is presented in Kulinskaya *et al*. (2011).That correction uses a chi-square distribution with fractional degrees of freedom (a special case of a gamma distribution) as the approximate null distribution of Q, estimating the degrees of freedom from the data. In this paper, we present a correction based on the gamma distribution for use when the effects are the differences of binomial probabilities.

We present expansions, without distributional assumptions, for the first two moments of Q, accurate to order $1/n$, for the situation in which the variance of the estimator of the effect is a function of the effect and additionally, of a nuisance parameter. We apply these expansions to the case when the effects are risk differences, and in extensive simulations, we have found that the gamma distribution whose first two moments match those given by the expansions is an excellent fit for the distribution of Q—substantially superior to the fit of the standard chi-square distribution. We also tried the fractional chi-square distribution based on the first moment of Q and an appropriate adaptation of the Welch test, but the results were inferior to the gamma distribution, and these other approximations are not recommended. All our simulations of the distribution of Q for risk differences have yielded values for the first two moments of Q that are larger than standard chi-square moments. Calculations using our expansions for the first two moments of Q agree well with the simulations.

Our simulations show that use of the gamma distribution for the null distribution of Q when the effects are risk differences improves the accuracy of the significance levels, especially for small to moderate study sizes. The improved test works well in a variety of circumstances, such as when the individual studies have unbalanced sizes between the two arms or when the studies' total sizes differ substantially from one another.

In previous empirical studies, Engels *et al*. (2000) and Deeks (2002) noted that meta-analyses based on risk difference were, on average, more heterogeneous than meta-analyses based on relative risk or odds ratio. Our simulations clearly show that the standard chi-square approximation to the distribution of Q results in a very liberal test. The use of our improved approximation should overcome this difficulty.

It is well known that the standard Q test based on the chi-square distribution has low power (see e.g., Viechtbauer, 2007); the use of the more accurate test will result in a decrease or an increase in power depending on whether the standard Q test is liberal or conservative, respectively. In the case of risk difference, the use of the more accurate Q test will result in a decrease in power, but the power is increased in the case of the standardized mean difference, see Kulinskaya *et al*. (2011). Needless to say, higher power is not an advantage for a test with wrong levels.

In future work, we plan to apply our expansions to such important effect measures as the relative risk and the odds ratio. Another direction for future work would be to investigate the non-null distribution of Q, with an important application to finding confidence intervals for the between-studies variance component in a random effects model of meta-analysis.

Appendix A:

A.1. Formula for the second moment of Q

The expansion for $E[Q^2]$ is given in the following text. The list of the derivatives of Q^2 needed for insertion into this expansion is given in Appendix A.3.

$$
\begin{aligned}
\mathrm{E}[Q^2] = {} & \frac{1}{24}\sum_i \frac{\partial^4 Q^2\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^4}\ \mathrm{E}\left[\Theta_i^4\right] + \frac{1}{8}\sum_{i\neq j}\sum \frac{\partial^4 Q^2\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^2\partial\hat{\theta}_j^2}\ \mathrm{E}\left[\Theta_i^2\right]\mathrm{E}\left[\Theta_j^2\right] \\
& + \frac{1}{120}\sum_i \frac{\partial^5 Q^2\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^5}\ \mathrm{E}\left[\Theta_i^5\right] + \frac{1}{24}\sum_i \frac{\partial^5 Q^2\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^4\partial\hat{\zeta}_i}\ \mathrm{E}\left[\Theta_i^4 Z_i\right] + \frac{1}{12}\sum_{i\neq j}\sum \frac{\partial^5 Q^2\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^3\partial\hat{\theta}_j^2}\ \mathrm{E}\left[\Theta_i^3\right]\mathrm{E}\left[\Theta_j^2\right] \\
& + \frac{1}{6}\sum_{i\neq j}\sum \frac{\partial^5 Q^2\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^3\partial\hat{\theta}_j\partial\hat{\zeta}_j}\ \mathrm{E}\left[\Theta_i^3\right]\mathrm{E}\left[\Theta_j Z_j\right] + \frac{1}{4}\sum_{i\neq j}\sum \frac{\partial^5 Q^2\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^2\partial\hat{\zeta}_i\partial\hat{\theta}_j^2}\mathrm{E}\left[\Theta_i^2 Z_i\right]\mathrm{E}\left[\Theta_j^2\right] \\
& + \frac{1}{720}\sum_i \frac{\partial^6 Q^2\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^6}\ \mathrm{E}\left[\Theta_i^6\right] + \frac{1}{120}\sum_i \frac{\partial^6 Q^2\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^5\partial\hat{\zeta}_i}\ \mathrm{E}\left[\Theta_i^5 Z_i\right] + \frac{1}{48}\sum_i \frac{\partial^6 Q^2\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^4\partial\hat{\zeta}_i^2}\ \mathrm{E}\left[\Theta_i^4 Z_i^2\right] \\
& + \frac{1}{48}\sum_{i\neq j}\sum \frac{\partial^6 Q^2\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^4\partial\hat{\theta}_j^2}\ \mathrm{E}\left[\Theta_i^4\right]\mathrm{E}\left[\Theta_j^2\right] + \frac{1}{24}\sum_{i\neq j}\sum \frac{\partial^6 Q^2\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^4\left(\partial\hat{\theta}_j\partial\hat{\zeta}_j\right)}\ \mathrm{E}\left[\Theta_i^4\right]\mathrm{E}\left[\Theta_j Z_j\right] \\
& + \frac{1}{48}\sum_{i\neq j}\sum \frac{\partial^6 Q^2\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^4\partial\hat{\zeta}_j^2}\ \mathrm{E}\left[\Theta_i^4\right]\mathrm{E}\left[Z_j^2\right] + \frac{1}{12}\sum_{i\neq j}\sum \frac{\partial^6 Q^2\left(\vec{\theta}\right)}{\left(\partial\hat{\theta}_i^3\partial\hat{\zeta}_i\right)\partial\hat{\theta}_j^2}\ \mathrm{E}\left[\Theta_i^3 Z_i\right]\mathrm{E}\left[\Theta_j^2\right] \\
& + \frac{1}{6}\sum_{i\neq j}\sum \frac{\partial^6 Q^2\left(\vec{\theta}\right)}{\left(\partial\hat{\theta}_i^3\partial\hat{\zeta}_i\right)\left(\partial\hat{\theta}_j\partial\hat{\zeta}_j\right)}\ \mathrm{E}\left[\Theta_i^3 Z_i\right]\mathrm{E}\left[\Theta_j Z_j\right] + \frac{1}{8}\sum_{i\neq j}\sum \frac{\partial^6 Q^2\left(\vec{\theta}\right)}{\left(\partial\hat{\theta}_i^2\partial\hat{\zeta}_i^2\right)\partial\hat{\theta}_j^2}\ \mathrm{E}\left[\Theta_i^2 Z_i^2\right]\mathrm{E}\left[\Theta_j^2\right] \\
& + \frac{1}{48}\sum_{i\neq j}\sum_{\neq k}\sum \frac{\partial^6 Q^2\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^2\partial\hat{\theta}_j^2\partial\hat{\theta}_k^2}\ \mathrm{E}\left[\Theta_i^2\right]\mathrm{E}\left[\Theta_j^2\right]\mathrm{E}\left[\Theta_k^2\right] + \frac{1}{8}\sum_{i\neq j}\sum_{\neq k}\sum \frac{\partial^6 Q^2\left(\vec{\theta}\right)}{\left(\partial\hat{\theta}_i\partial\hat{\zeta}_i\right)\partial\hat{\theta}_j^2\partial\hat{\theta}_k^2}\ \mathrm{E}\left[\Theta_i Z_i\right]\mathrm{E}\left[\Theta_j^2\right]\mathrm{E}\left[\Theta_k^2\right] \\
& + \frac{1}{4}\sum_{i\neq j}\sum_{\neq k}\sum \frac{\partial^6 Q^2\left(\vec{\theta}\right)}{\left(\partial\hat{\theta}_i\partial\hat{\zeta}_i\right)\left(\partial\hat{\theta}_j\partial\hat{\zeta}_j\right)\partial\hat{\theta}_k^2}\ \mathrm{E}\left[\Theta_i Z_i\right]\mathrm{E}\left[\Theta_j Z_j\right]\mathrm{E}\left[\Theta_k^2\right] \\
& + \frac{1}{16}\sum_{i\neq j}\sum_{\neq k}\sum \frac{\partial^6 Q^2\left(\vec{\theta}\right)}{\partial\hat{\zeta}_i^2\partial\hat{\theta}_j^2\partial\hat{\theta}_k^2}\ \mathrm{E}\left[Z_i^2\right]\mathrm{E}\left[\Theta_j^2\right]\mathrm{E}\left[\Theta_k^2\right] + O\left(\frac{1}{n^2}\right)
\end{aligned}
\tag{3}
$$

A.2. Derivatives for the first moment of Q

Here, we list the derivatives of Q needed for insertion into Equation (1). We first did the calculations by hand and then checked them by using the computer algebra package `Mathematica` (Wolfram Research, Inc., 2008). We use the notation $W = \Sigma_i w_i$ and $U_i = 1 - w_i/W$ and evaluate all derivatives at the null hypothesis. All multi-index derivatives assume inequality of the indices i and j.

$$
\begin{aligned}
\frac{\partial^2 Q(\vec{\theta})}{\partial\hat{\theta}_i^2} &= 2w_i U_i \\
\frac{\partial^3 Q(\vec{\theta})}{\partial\hat{\theta}_i^3} &= 6U_i^2\frac{\partial f_i(\vec{\theta})}{\partial\hat{\theta}_i} \\
\frac{\partial^3 Q(\vec{\theta})}{\partial\hat{\theta}_i^2\partial\hat{\zeta}_i} &= 2U_i^2\frac{\partial f_i(\vec{\theta})}{\partial\hat{\zeta}_i} \\
\frac{\partial^4 Q(\vec{\theta})}{\partial\hat{\theta}_i^4} &= 12U_i^2\left[\frac{\partial^2 f_i(\vec{\theta})}{\partial\hat{\theta}_i^2} - \frac{2}{W}\left(\frac{\partial f_i(\vec{\theta})}{\partial\hat{\theta}_i}\right)^2\right] \\
\frac{\partial^4 Q(\vec{\theta})}{\partial\hat{\theta}_i^3\partial\hat{\zeta}_i} &= 6U_i^2\left[\frac{\partial^2 f_i(\vec{\theta})}{\partial\hat{\theta}_i\partial\hat{\zeta}_i} - \frac{2}{W}\left(\frac{\partial f_i(\vec{\theta})}{\partial\hat{\theta}_i}\right)\left(\frac{\partial f_i(\vec{\theta})}{\partial\hat{\zeta}_i}\right)\right] \\
\frac{\partial^4 Q(\vec{\theta})}{\partial\hat{\theta}_i^2\partial\hat{\zeta}_i^2} &= 2U_i^2\left[\frac{\partial^2 f_i(\vec{\theta})}{\partial\hat{\zeta}_i^2} - \frac{2}{W}\left(\frac{\partial f_i(\vec{\theta})}{\partial\hat{\zeta}_i}\right)^2\right] \\
\frac{\partial^4 Q(\vec{\theta})}{\partial\theta_i^2\partial\theta_j^2} &= \frac{-4}{W^3}\left[w_j^2\left(\frac{\partial f_i(\vec{\theta})}{\partial\hat{\theta}_i}\right)^2 + w_i^2\left(\frac{\partial f_j(\vec{\theta})}{\partial\hat{\theta}_j}\right)^2\right] + \frac{2}{W^2}\left[w_j^2\frac{\partial^2 f_i(\vec{\theta})}{\partial\hat{\theta}_i^2} + w_i^2\frac{\partial^2 f_j(\vec{\theta})}{\partial\hat{\theta}_j^2}\right] \\
&\quad + \frac{8}{W^2}\left[w_i U_j + w_j U_i - W\right]\left(\frac{\partial f_i(\vec{\theta})}{\partial\hat{\theta}_i}\right)\left(\frac{\partial f_j(\vec{\theta})}{\partial\hat{\theta}_j}\right) \\
\frac{\partial^4 Q(\vec{\theta})}{\partial\hat{\theta}_i\partial\hat{\zeta}_i\partial\hat{\theta}_j\partial\hat{\zeta}_j} &= \frac{2}{W^2}\left[w_i U_j + w_j U_i - W\right]\left(\frac{\partial f_i(\vec{\theta})}{\partial\hat{\zeta}_i}\right)\left(\frac{\partial f_j(\vec{\theta})}{\partial\hat{\zeta}_j}\right) \\
\frac{\partial^4 Q(\vec{\theta})}{\partial\hat{\theta}_i^2\partial\hat{\theta}_j\partial\hat{\zeta}_j} &= \frac{4}{W^2}\left[w_i U_j + w_j U_i - W\right]\left(\frac{\partial f_i(\vec{\theta})}{\partial\hat{\theta}_i}\right)\left(\frac{\partial f_j(\vec{\theta})}{\partial\hat{\zeta}_j}\right) \\
&\quad + \frac{2w_i^2}{W^2}\left(\frac{\partial^2 f_j(\vec{\theta})}{\partial\hat{\theta}_j\partial\hat{\zeta}_j}\right) - \frac{4w_i^2}{W^3}\left(\frac{\partial f_j(\vec{\theta})}{\partial\hat{\theta}_j}\right)\left(\frac{\partial f_j(\vec{\theta})}{\partial\hat{\zeta}_j}\right) \\
\frac{\partial^4 Q(\vec{\theta})}{\partial\hat{\theta}_i^2\partial\hat{\zeta}_j^2} &= \frac{2w_i^2}{W^2}\left(\frac{\partial^2 f_j(\vec{\theta})}{\partial\hat{\zeta}_j^2}\right) - \frac{4w_i^2}{W^3}\left(\frac{\partial f_j(\vec{\theta})}{\partial\hat{\zeta}_j}\right)^2
\end{aligned}
$$

A.3. Derivatives for the second moment of Q

Here, we list the derivatives of Q^2 needed for insertion into Equation (3). From the order assumptions of Section 2, we know which derivatives of Q^2 will yield a contribution of $O(1/n)$ or greater to $\mathrm{E}[Q^2]$. The calculations were carried out with the aid of the computer algebra package `Mathematica` (Wolfram Research, Inc., 2008).

$$\frac{\partial^4 Q^2\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^4} = 24 w_i^2 U_i^2$$

$$\frac{\partial^4 Q^2\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^2\partial\hat{\theta}_j^2} = 8 w_i w_j\left(U_i U_j + \frac{2 w_i w_j}{W^2}\right)$$

$$\frac{\partial^5 Q^2\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^5} = 240 w_i U_i^3 \frac{\partial f_i\left(\vec{\theta}\right)}{\partial\hat{\theta}_i}$$

$$\frac{\partial^5 Q^2\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^4\partial\hat{\zeta}_i} = 48 w_i U_i^3 \frac{\partial f_i\left(\vec{\theta}\right)}{\partial\hat{\zeta}_i}$$

$$\frac{\partial^5 Q^2\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^3\partial\hat{\theta}_j^2} = 24 U_i w_j\left[U_i U_j + \frac{5 w_i w_j}{W^2}\right]\frac{\partial f_i\left(\vec{\theta}\right)}{\partial\hat{\theta}_i} - 48\frac{w_i^2}{W}\left[U_i U_j + \frac{w_i w_j}{W^2}\right]\frac{\partial f_j\left(\vec{\theta}\right)}{\partial\hat{\theta}_j}$$

$$\frac{\partial^5 Q^2\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^3\left(\partial\hat{\theta}_j\partial\hat{\zeta}_j\right)} = -24\frac{w_i^2}{W}\left[U_i U_j + \frac{w_i w_j}{W^2}\right]\frac{\partial f_j\left(\vec{\theta}\right)}{\partial\hat{\zeta}_j}$$

$$\frac{\partial^5 Q^2\left(\vec{\theta}\right)}{\left(\partial\hat{\theta}_i^2\partial\hat{\zeta}_i\right)\partial\hat{\theta}_j^2} = 8 U_i w_j\left[U_i U_j + \frac{5 w_i w_j}{W^2}\right]\frac{\partial f_i\left(\vec{\theta}\right)}{\partial\hat{\zeta}_i}$$

$$\frac{\partial^6 Q^2\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^6} = 720 U_i^3\left[\left(U_i - \frac{2 w_i}{W}\right)\left(\frac{\partial f_i\left(\vec{\theta}\right)}{\partial\hat{\theta}_i}\right)^2 + w_i\frac{\partial^2 f_i\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^2}\right]$$

$$\frac{\partial^6 Q^2\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^5\partial\hat{\zeta}_i} = 240 U_i^3\left[\left(U_i - \frac{2 w_i}{W}\right)\left(\frac{\partial f_i\left(\vec{\theta}\right)}{\partial\hat{\theta}_i}\right)\left(\frac{\partial f_i\left(\vec{\theta}\right)}{\partial\hat{\zeta}_i}\right) + w_i\frac{\partial^2 f_i\left(\vec{\theta}\right)}{\partial\hat{\theta}_i\partial\hat{\zeta}_i}\right]$$

$$\frac{\partial^6 Q^2\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^4\partial\hat{\zeta}_i^2} = 48 U_i^3\left[\left(U_i - \frac{2 w_i}{W}\right)\left(\frac{\partial f_i\left(\vec{\theta}\right)}{\partial\hat{\zeta}_i}\right)^2 + w_i\frac{\partial^2 f_i\left(\vec{\theta}\right)}{\partial\hat{\zeta}_i^2}\right]$$

$$\begin{aligned}\frac{\partial^6 Q^2\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^4\partial\hat{\theta}_j^2} = \frac{48}{W^4}\Bigg\{ & -2 W U_i w_j\left(W^2 U_i - 4 W w_j + 9 w_i w_j\right)\left(\frac{\partial f_i\left(\vec{\theta}\right)}{\partial\hat{\theta}_i}\right)^2 \\ & + W^2 U_i w_j\left(W^2 - W w_i - W w_j + 6 w_i w_j\right)\frac{\partial^2 f_i\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^2} \\ & - 8 W U_i w_i\left(W^2 U_i - W w_j + 3 w_i w_j\right)\left(\frac{\partial f_i\left(\vec{\theta}\right)}{\partial\hat{\theta}_i}\right)\left(\frac{\partial f_j\left(\vec{\theta}\right)}{\partial\hat{\theta}_j}\right) \\ & + w_i^3\left(-2W + 3 w_i\right)\left(\frac{\partial f_j\left(\vec{\theta}\right)}{\partial\hat{\theta}_j}\right)^2 + W^2 U_i w_i^3\frac{\partial^2 f_j\left(\vec{\theta}\right)}{\partial\hat{\theta}_j^2}\Bigg\}\end{aligned}$$

$$\begin{aligned}\frac{\partial^6 Q^2\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^4\left(\partial\hat{\theta}_j\partial\hat{\zeta}_j\right)} = & \frac{-192}{W^3} U_i w_i\left(W^2 - W w_i - W w_j + 3 w_i w_j\right)\left(\frac{\partial f_i\left(\vec{\theta}\right)}{\partial\hat{\theta}_i}\right)\left(\frac{\partial f_j\left(\vec{\theta}\right)}{\partial\hat{\zeta}_j}\right) \\ & - \frac{48}{W^4} w_i^3\left(2W - 3 w_i\right)\left(\frac{\partial f_j\left(\vec{\theta}\right)}{\partial\hat{\theta}_j}\right)\left(\frac{\partial f_j\left(\vec{\theta}\right)}{\partial\hat{\zeta}_j}\right) \\ & + \frac{48}{W^2} U_i w_i^3\frac{\partial^2 f_j\left(\vec{\theta}\right)}{\partial\hat{\theta}_j\partial\hat{\zeta}_j}\end{aligned}$$

$$\frac{\partial^6 Q^2\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^4\partial\hat{\zeta}_j^2} = -\frac{48 w_i^3}{W^4}\left(2W - 3 w_i\right)\left(\frac{\partial f_j\left(\vec{\theta}\right)}{\partial\hat{\zeta}_j}\right)^2 + \frac{48 U_i w_i^3}{W^2}\frac{\partial^2 f_j\left(\vec{\theta}\right)}{\partial\hat{\zeta}_j^2}$$

$$\begin{aligned}\frac{\partial^6 Q^2\left(\vec{\theta}\right)}{\left(\partial\hat{\theta}_i^3\partial\hat{\zeta}_i\right)\partial\hat{\theta}_j^2} = & -\frac{48 U_i w_j}{W^3}\left(W^2 - W w_i - 4 W w_j + 9 w_i w_j\right)\left(\frac{\partial f_i\left(\vec{\theta}\right)}{\partial\hat{\theta}_i}\right)\left(\frac{\partial f_i\left(\vec{\theta}\right)}{\partial\hat{\zeta}_i}\right) \\ & - \frac{96 U_i w_i}{W^3}\left(W^2 - W w_i - W w_j + 3 w_i w_j\right)\left(\frac{\partial f_i\left(\vec{\theta}\right)}{\partial\hat{\zeta}_i}\right)\left(\frac{\partial f_j\left(\vec{\theta}\right)}{\partial\hat{\theta}_j}\right) \\ & + \frac{24 U_i w_j}{W^2}\left(W^2 - W w_i - W w_j + 6 w_i w_j\right)\frac{\partial^2 f_i\left(\vec{\theta}\right)}{\partial\hat{\theta}_i\partial\hat{\zeta}_i}\end{aligned}$$

$$\frac{\partial^6 Q^2\left(\vec{\theta}\right)}{\left(\partial\hat{\theta}_i^3\partial\hat{\zeta}_i\right)\left(\partial\hat{\theta}_j\partial\hat{\zeta}_j\right)} = -\frac{48U_i w_i}{W^3}\left(W^2 - Ww_i - Ww_j + 3w_i w_j\right)\left(\frac{\partial f_i\left(\vec{\theta}\right)}{\partial\hat{\zeta}_i}\right)\left(\frac{\partial f_j\left(\vec{\theta}\right)}{\partial\hat{\zeta}_j}\right)$$

$$\begin{aligned}\frac{\partial^6 Q^2\left(\vec{\theta}\right)}{\left(\partial\hat{\theta}_i^2\partial\hat{\zeta}_i^2\right)\partial\hat{\theta}_j^2} = &-\frac{16U_i w_j}{W^3}\left(W^2 - Ww_i - 4Ww_j + 9w_i w_j\right)\left(\frac{\partial f_i\left(\vec{\theta}\right)}{\partial\hat{\zeta}_i}\right)^2\\ &+\frac{8U_i w_j}{W^2}\left(W^2 - Ww_i - Ww_j + 6w_i w_j\right)\frac{\partial^2 f_i\left(\vec{\theta}\right)}{\partial\hat{\zeta}_i^2}\end{aligned}$$

$$\begin{aligned}\frac{\partial^6 Q^2\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^2\partial\hat{\theta}_j^2\partial\hat{\theta}_k^2} = &-\frac{16w_j w_k}{W^4}\left(Ww_j + Ww_k - 9w_j w_k\right)\left(\frac{\partial f_i\left(\vec{\theta}\right)}{\partial\hat{\theta}_i}\right)^2\\ &-\frac{16w_i w_k}{W^4}\left(Ww_i + Ww_k - 9w_i w_k\right)\left(\frac{\partial f_j\left(\vec{\theta}\right)}{\partial\hat{\theta}_j}\right)^2\\ &-\frac{16w_i w_j}{W^4}\left(Ww_i + Ww_j - 9w_i w_j\right)\left(\frac{\partial f_k\left(\vec{\theta}\right)}{\partial\hat{\theta}_k}\right)^2\\ &+\frac{8w_j w_k}{W^3}\left(Ww_j + Ww_k - 6w_j w_k\right)\frac{\partial^2 f_i\left(\vec{\theta}\right)}{\partial\hat{\theta}_i^2}\\ &+\frac{8w_i w_k}{W^3}\left(Ww_i + Ww_k - 6w_i w_k\right)\frac{\partial^2 f_j\left(\vec{\theta}\right)}{\partial\hat{\theta}_j^2}\\ &+\frac{8w_i w_j}{W^3}\left(Ww_i + Ww_j - 6w_i w_j\right)\frac{\partial^2 f_k\left(\vec{\theta}\right)}{\partial\hat{\theta}_k^2}\\ &-\frac{32w_k}{W^2}\left[U_i U_j(W - 6w_k) + \frac{w_i w_j}{W^2}(W - 12w_k) + 3w_k\right]\left(\frac{\partial f_i\left(\vec{\theta}\right)}{\partial\hat{\theta}_i}\right)\left(\frac{\partial f_j\left(\vec{\theta}\right)}{\partial\hat{\theta}_j}\right)\\ &-\frac{32w_j}{W^2}\left[U_i U_k(W - 6w_j) + \frac{w_i w_k}{W^2}(W - 12w_j) + 3w_j\right]\left(\frac{\partial f_i\left(\vec{\theta}\right)}{\partial\hat{\theta}_i}\right)\left(\frac{\partial f_k\left(\vec{\theta}\right)}{\partial\hat{\theta}_k}\right)\\ &-\frac{32w_i}{W^2}\left[U_j U_k(W - 6w_i) + \frac{w_j w_k}{W^2}(W - 12w_i) + 3w_i\right]\left(\frac{\partial f_j\left(\vec{\theta}\right)}{\partial\hat{\theta}_j}\right)\left(\frac{\partial f_k\left(\vec{\theta}\right)}{\partial\hat{\theta}_k}\right)\end{aligned}$$

$$\begin{aligned}\frac{\partial^6 Q^2\left(\vec{\theta}\right)}{\left(\partial\hat{\theta}_i\partial\hat{\zeta}_i\right)\partial\hat{\theta}_j^2\partial\hat{\theta}_k^2} = &-\frac{16w_j w_k}{W^4}\left(Ww_j + Ww_k - 9w_j w_k\right)\left(\frac{\partial f_i\left(\vec{\theta}\right)}{\partial\hat{\theta}_i}\right)\left(\frac{\partial f_i\left(\vec{\theta}\right)}{\partial\hat{\zeta}_i}\right)\\ &+\frac{8w_j w_k}{W^3}\left(Ww_j + Ww_k - 6w_j w_k\right)\frac{\partial^2 f_i\left(\vec{\theta}\right)}{\partial\hat{\theta}_i\partial\hat{\zeta}_i}\\ &-\frac{16w_k}{W^2}\left[U_i U_j(W - 6w_k) + \frac{w_i w_j}{W^2}(W - 12w_k) + 3w_k\right]\left(\frac{\partial f_i\left(\vec{\theta}\right)}{\partial\hat{\zeta}_i}\right)\left(\frac{\partial f_j\left(\vec{\theta}\right)}{\partial\hat{\theta}_j}\right)\\ &-\frac{16w_j}{W^2}\left[U_i U_k(W - 6w_j) + \frac{w_i w_k}{W^2}(W - 12w_j) + 3w_j\right]\left(\frac{\partial f_i\left(\vec{\theta}\right)}{\partial\hat{\zeta}_i}\right)\left(\frac{\partial f_k\left(\vec{\theta}\right)}{\partial\hat{\theta}_k}\right)\end{aligned}$$

$$\begin{aligned}\frac{\partial^6 Q^2\left(\vec{\theta}\right)}{\partial\hat{\zeta}_i^2\partial\hat{\theta}_j^2\partial\hat{\theta}_k^2} = &-\frac{16w_j w_k}{W^4}\left(Ww_j + Ww_k - 9w_j w_k\right)\left(\frac{\partial f_i\left(\vec{\theta}\right)}{\partial\hat{\zeta}_i}\right)^2\\ &+\frac{8w_j w_k}{W^3}\left(Ww_j + Ww_k - 6w_j w_k\right)\frac{\partial^2 f_i\left(\vec{\theta}\right)}{\partial\hat{\zeta}_i^2}\end{aligned}$$

$$\frac{\partial^6 Q^2\left(\vec{\theta}\right)}{\left(\partial\hat{\theta}_i\partial\hat{\zeta}_i\right)\left(\partial\hat{\theta}_j\partial\hat{\zeta}_j\right)\partial\hat{\theta}_k^2} = -\frac{8w_k}{W^2}\left[U_i U_j(W - 6w_k) + \frac{w_i w_j}{W^2}(W - 12w_k) + 3w_k\right]\left(\frac{\partial f_i\left(\vec{\theta}\right)}{\partial\hat{\zeta}_i}\right)\left(\frac{\partial f_j\left(\vec{\theta}\right)}{\partial\hat{\zeta}_j}\right)$$

Supporting Information

The following supporting information may be found in the online version of this article.

Web Appendix

An appendix with four sections. A. Derivation of the moment expansions for Q; B. Derivation of the Welch moments; C. Formulas needed for risk difference calculations; D. Additional simulation results for risk difference.

R program for homogeneity test of risk differences

R program containing code to perform the homogeneity test described in the article, its description and the data set used as an example.

Acknowledgment

The authors wish to thank two anonymous reviewers, an associate editor, and the editor for their useful suggestions for improving the presentation of the material of this article. We appreciate the especially careful reading of the manuscript by the two reviewers.

References

Biggerstaff BJ, Jackson D. 2008. The exact distribution of Cochran's heterogeneity statistic in one-way random effects meta-analysis. *Statistics in Medicine* **27**: 6093–6110.

Biggerstaff BJ, Tweedie RL. 1997. Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Statistics in Medicine* **16**: 753–768.

Böhning D, Sarol J, Jr. 2000. Estimating risk difference in multicenter studies under baseline-risk heterogeneity (reader reaction). *Biometrics* **56**: 304–308.

Cochran WG. 1937. Problems arising in the analysis of a series of similar experiments. *Supplement to the Journal of the Royal Statistical Society* **40**(1): 102–118.

Cochran WG. 1954. The combination of estimates from different experiments. *Biometrics* **100**(1): 101–129.

Crowley P. 2000. Prophylactic corticosteroids for preterm birth. *Cochrane Database of Systematic Reviews* Issue(2). Art. No.: CD000065.

Deeks JJ. 2002. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine* **21**: 1575–1600.

Deeks J, Higgins J. 2005. Standard statistical algorithms in Cochrane reviews. Version 5. Technical report, The Cochrane Collaboration.

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

Engels EA, Schmid CH, Terrin N, Olkin I, Lau J. 2000. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. *Statistics in Medicine* **19**: 1707–1728.

Firth D. 1993. Bias reduction of maximum likelihood estimates. *Biometrika* **80**: 27–38.

Fleiss JL. 1981. Statistical Methods for Rates and Proportions. 2nd edition. John Wiley: New York.

Hedges LV, Olkin I. 1985. Statistical Methods for Meta-analysis. Academic Press: Orlando, FL.

Jackson D. 2006. The power of the standard test for the presence of heterogeneity in meta-analysis. *Statistics in Medicine* **25**: 2688–2699.

James GS. 1951. The comparison of several groups of observations when the ratios of the population variances are unknown. *Biometrika*, **38**: 324–329.

Kelly C, Rao-Melacini P, Zhao W. 2005. Using James-Stein estimators in homogeneity tests of the risk difference. *Communications in Statistics–Simulation and Computation* **34**: 41–55.

Kulinskaya E, Dollinger MB. 2007. Robust weighted one-way ANOVA: improved approximation and efficiency. *Journal of Statistical Planning and Inference* **137**: 462–472.

Kulinskaya E, Dollinger MB, Knight E, Gao H. 2004. A Welch-type test for homogeneity of contrasts under heteroscedasticity with application to meta-analysis. *Statistics in Medicine* **23**: 3655–3670.

Kulinskaya E, Dollinger MB, Bjørkestøl K. 2011. Testing for homogeneity in meta-analysis I. *The one parameter case: Standardized mean difference. Biometrics* **67**: 203–212.

Lipsitz S, Dear K, Laird N, Molenberghs G. 1998. Tests for homogeneity of the risk difference when data are sparse. *Biometrics* **54**: 148–160.

Lui K-J, Kelly C. 2000. A revisit on tests for homogeneity of the risk difference. *Biometrics* **56**: 309–315.

Sánchez-Meca J, Marýn-Martýnez F. 2000. Testing the significance of a common risk difference in meta-analysis. *Computational Statistics & Data Analysis* **33**: 299–313.

Schwarzer G. 2010. Meta, v.1.5-0. http://CRAN.R-project.org/package=meta. R package. Fixed and random effects meta-analysis. Functions for tests of bias, forest and funnel plot.

Viechtbauer W. 2007. Hypothesis tests for population heterogeneity in meta-analysis. *British Journal of Mathematical and Statistical Psychology*, **60**: 29–60.
Welch BL. 1951. On the comparison of several mean values: an alternative approach. *Biometrika* **38**: 330–336.
Wolfram Research, Inc. 2008. Mathematica, Version 7.0. Wolfram Research, Inc.: Champaign, Illinois.