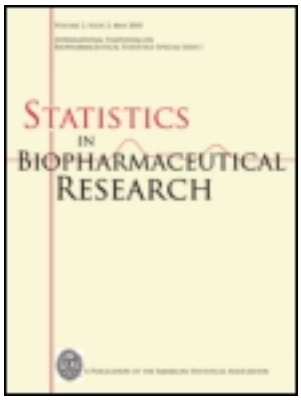

Statistics in Biopharmaceutical Research
Publication details, including instructions for authors and subscription information:
http://amstat.tandfonline.com/loi/usbr20

The Significance Level of the Standard Test for a Treatment Effect in Meta-analysis

Dan Jackson

Version of record first published: 01 Jan 2012.

To cite this article: Dan Jackson (2009): The Significance Level of the Standard Test for a Treatment Effect in Meta-analysis, Statistics in Biopharmaceutical Research, 1:1, 92-100

To link to this article: http://dx.doi.org/10.1198/sbr.2009.0009

The Significance Level of the Standard Test for a Treatment Effect in Meta-analysis

DAN JACKSON

The random effects model is routinely used in meta-analysis. Although a variety of approaches have been proposed for implementing this in practice, the most popular method is that suggested by DerSimonian and Laird. The resulting inferences are not exact, however, and modifications to the methodology have been suggested. By deriving the exact distribution of the resulting test statistic, under the null hypothesis and in the simplified scenario where all studies are the same size, it is possible to obtain the actual significance level of the standard hypothesis test for a treatment effect. In particular it is found that, if there is a small number of studies and considerable between-study variation, the actual significance level can be much larger than the nominal level. This and other findings are illustrated using examples concerning the effectiveness of treatments to reduce serum cholesterol for preventing death in patients with no history of heart attacks, the use of glycerol for patients who have suffered an acute stroke, and the use of amisulpride to treat schizophrenia.

Key Words: Hypothesis test; Modified test; Random effects model.

1. Introduction

Meta-analysis, the statistical process of pooling the results from separate studies concerned with the same treatment or issue, is used in a wide variety of applications. In particular, it is routinely used in medical contexts and provides the means to synthesize the efficacy of medical interventions from all relevant studies that have been carried out. Meta-analysis therefore plays a fundamental role in the evidence-based medicine program.

One difficulty, however, is presented by between-study variation, which refers to the possibility that the studies' results may be too disparate to assume that they measure the same underlying effect. In order to model this heterogeneity but still obtain a meaningful estimate of treatment effect, the random effects model—discussed in Section 2—is frequently used. Modeling the studies' estimates as being homogenous is a rather strong assumption, so the random effects model will be adopted here as the standard model, although debate continues concerning fixed versus random effects modeling. For a full account of this important and unresolved issue, see Sutton et al. (2002, p. 83).

However it is implemented in practice, the random effects model tacitly invokes the central limit and Slutsky's theorems in order to justify the use of normal approximations for the studies' estimated treatment effects, conditional on their true underlying values, with effectively fixed and known within-study variances. It is generally well understood where such an approximation is suitable and, by adopting a measure of treatment effect where the normality assumption is most appropriate, meta-analysts can improve the accuracy of their probability statements which give rise to inferences. It will be assumed throughout that studies are sufficiently large to justify the use of such normal approximations.

The random effects model further assumes that the true underlying treatment effects are themselves normally distributed, centered at the overall treatment effect with a between-study variance. These two parameters are therefore required by the random effects model

Statistics in Biopharmaceutical Research
February 2009, Vol. 1, No. 1
DOI: 10.1198/sbr.2009.0009

and the overall treatment effect, providing a measure of the effectiveness of the treatment for a typical or average study, is the parameter of central interest, although the importance of inferences concerning the degree of between-study variation has also been emphasized (Biggerstaff and Tweedie 1997; Jackson 2006a). The random effects model can be used to provide confidence intervals, results from hypothesis tests and, with suitable prior distributions, posterior distributions for the overall treatment effect. As this is one of the first articles to investigate analytically an exact sampling distribution resulting from the random effects model, the focus here is on hypothesis testing, perhaps the simplest of these types of inferences. Despite this, the ideas presented could be developed to provide similar insights into other methods of statistical inference.

Although many approaches have been suggested when adopting a random effects model for the purposes of carrying out a meta-analysis (e.g., Biggerstaff and Tweedie 1997; Emerson et al. 1997; Hardy and Thompson 1996; Hartung and Knapp 2001; and Van Houwelingen et al. 1993) the most popular approach is that suggested by DerSimonian and Laird (1986). This will therefore be regarded as the standard method, whose popularity is no doubt partly due to its relative simplicity. DerSimonian and Laird provided a moments estimator of the between-study variance and, once this estimate has been obtained, effectively use this as the true value. Although justifiable asymptotically, this is a further approximation and many meta-analyses involve just a handful of studies. Hence there is the concern that this approach may not always be suitable, although it is currently poorly understood when this might be the case. The principal contribution of this article is to investigate this issue analytically, in the simplified context where all studies are the same size, in order to provide an indication of when the hypothesis test for a treatment effect, resulting from using the now standard approach suggested by DerSimonian and Laird, is appropriate. Although this simplification is somewhat artificial, this scenario gives an indication of the appropriateness of the test more generally, as shown in Section 5. In particular, Follmann and Proschan (1999) suggested that the actual significance level can be much larger than the nominal level provided by the test. Although this does indeed turn out to be the biggest issue, the actual significance level can also be too small, and is appropriate under some circumstances. Hence the standard test for a treatment effect is suitable for many meta-analyses, but for others some modification is required.

The rest of the article is set out as follows. In Section 2 the random effects model and the DerSimonian and Laird procedure for implementing this are described in detail. In Section 3 the exact sampling distribution of the resulting test statistic for a treatment effect, under the null hypothesis and in the simplified setting where the studies are the same size, is derived. This distribution is then used in Section 4 to explore the significance level of the standard test, and that of a recently proposed modification, providing valuable insights into the situations where these tests are appropriate. In Section 5 some example datasets provide situations where the study sizes are very disparate, illustrating that the results obtained in Section 4 can be used to give an indication of the significance level for real meta-analyses where the study sizes inevitably differ, and Section 6 provides some concluding remarks.

2. The Random Effects Model

The random effects model assumes that the estimate of treatment effect from the ith study, y_i, is distributed as $y_i|\mu_i \sim N(\mu_i, \sigma_i^2)$, where μ_i is the true underlying treatment effect of the ith study and σ_i^2 is the corresponding within-study variance. This assumption is justified by the central limit theorem, which ensures that $\{(y_i|\mu_i) - \mu_i\}/\sigma_i$ converges in distribution to a standard normal distribution under very general circumstances. The variance σ_i^2 is unknown in practice but a consistent estimate, $\hat{\sigma}_i^2$, provides $\sigma_i/\hat{\sigma}_i \to 1$ in probability. We can therefore effectively replace σ_i^2 with $\hat{\sigma}_i^2$ as Slutsky's theorem implies that $(\sigma_i/\hat{\sigma}_i)\{(y_i|\mu_i) - \mu_i\}/\sigma_i = \{(y_i|\mu_i) - \mu_i\}/\hat{\sigma}_i$ converges in distribution to a standard normal. This justifies the usual approximation $y_i|\mu_i \sim N(\mu_i, \hat{\sigma}_i^2)$ and hence the symbols σ_i^2 and $\hat{\sigma}_i^2$ are used interchangeably to denote the within-study variances, effectively assumed fixed and known but estimated in practice. Note that it is this part of the modeling process that requires sufficiently large studies to justify the approximation and that this is required for both the usual fixed and random effects models. Estimates of within-study variance for a wide range of measures of treatment effect used in meta-analysis can be obtained as described by Sutton et al. (2002).

The random effects model further assumes that $\mu_i \sim N(\mu, \tau^2)$, where μ and τ^2 denote the overall treatment effect and between-study variance, respectively, and that the studies are independent. This provides the approximate marginal distributions $y_i \sim N(\mu, \hat{\sigma}_i^2 + \tau^2)$.

In order to make inferences about the overall treatment effect, a consistent estimate of τ^2 is conventionally evaluated. There is, however, no universally agreed method for estimating this parameter; for a detailed account of the most frequently used procedures, see Biggerstaff and Tweedie (1997). Once a consistent estimate of τ^2 has been evaluated, this is effectively used as the true value. This is because a consistent estimate ensures that $(\sqrt{\hat{\sigma}_i^2 + \tau^2}/\sqrt{\hat{\sigma}_i^2 + \hat{\tau}^2}) \to 1$ in probability (as the *number of studies* tends towards infinity).

Hence $(y_i - \mu)/\sqrt{\hat{\sigma}_i^2 + \hat{\tau}^2}$ tends in distribution to a standard normal, providing the approximation that $y_i \sim N(\mu, \hat{\sigma}_i^2 + \hat{\tau}^2)$. Note that this argument makes considerable use of asymptotic results so it is not surprising that situations can be found where the conventional approximations used in meta-analysis are not appropriate.

2.1 Applying the Random Effects Model in Practice

Once $\hat{\tau}^2$ has been evaluated, inference for μ is straightforward. The estimate of the overall treatment effect is given by $\hat{\mu} = \sum_{i=1}^n w_i^* y_i / \sum_{i=1}^n w_i^*$, where $w_i^* = (\hat{\sigma}_i^2 + \hat{\tau}^2)^{-1}$, and the distribution of $\hat{\mu}$ is approximately $\hat{\mu} \sim N\left(\mu, \left(\sum_{i=1}^n w_i^*\right)^{-1}\right)$. Hence confidence intervals and results from hypothesis tests are easily obtained. In particular, the test statistic for testing $H_0 : \mu = 0$ is given by R, the ratio of $\hat{\mu}$ and its standard error,

$$R = \frac{\sum_{i=1}^n \frac{y_i}{\sigma_i^2 + \hat{\tau}^2}}{\sqrt{\sum_{i=1}^n \frac{1}{\sigma_i^2 + \hat{\tau}^2}}}. \qquad (1)$$

The evaluated test statistic R is then compared to an appropriate percentile of a standard normal distribution for the purposes of testing the null hypothesis that $\mu = 0$ and therefore that there is no treatment effect.

2.2 The Standard Approach of DerSimonian and Laird (1986)

The simplest and most commonly used estimate of τ^2 is the DerSimonian and Laird (1986) estimate. This uses the Q statistic,

$$Q = \sum_{i=1}^n w_i (y_i - \bar{y})^2,$$

where $w_i = \hat{\sigma}_i^{-2}$, $\bar{y} = \sum_{i=1}^n w_i y_i / \sum_{i=1}^n w_i$, and n denotes the number of studies. Under the assumptions of the random effects model it can be shown that the expectation of Q is approximately

$$\mathrm{E}[Q] = (n-1) + \left(S_1 - \frac{S_2}{S_1}\right)\tau^2, \qquad (2)$$

where $S_r = \sum_{i=1}^n w_i^r$, which provides the DerSimonian and Laird estimate

$$\hat{\tau}^2 = \max\left(0, \frac{Q - (n-1)}{S_1 - \frac{S_2}{S_1}}\right).$$

If $\hat{\tau}^2$ is truncated to zero, then the resulting confidence intervals and results from hypothesis tests are the same as those from a fixed effects analysis. Hence $\hat{\tau}^2 = 0$ effectively means that a fixed effects procedure is adopted, although a random effects perspective may be maintained when interpreting the results.

In order to justify adopting the DerSimonian and Laird estimate in the usual way, however, and in particular its standard use in test statistic (1), we require that it is consistent. This is easily demonstrated using some well-known properties of the Q statistic. Biggerstaff and Tweedie (1997) obtained

$$\mathrm{var}[Q] = 2(n-1) + 4\left(S_1 - \frac{S_2}{S_1}\right)\tau^2 + 2\left(S_2 - 2\frac{S_3}{S_1} + \frac{S_2^2}{S_1^2}\right)\tau^4. \qquad (3)$$

The untruncated version of the DerSimonian and Laird estimate, $\hat{\tau}_U^2 = \{Q - (n-1)\}/(S_1 - S_2/S_1)$, is simply a linear function of Q, so that $\mathrm{E}(\hat{\tau}_U^2)$ can be obtained as τ^2 from (2) and $\mathrm{var}(\hat{\tau}_U^2)$ can be obtained from (3). Envisage a larger meta-analysis than one which has been "observed," with the same distribution of $\hat{\sigma}_i^2$ but with m times as many studies, so that the number of studies is mn and S_r becomes mS_r, where n and S_r are the values corresponding to the "observed" meta-analysis. Replacing n and S_r with mn and mS_r in (2) and (3) again provides $\mathrm{E}(\hat{\tau}_U^2) = \tau^2$, but the resulting expression for $\mathrm{var}(\hat{\tau}_U^2)$ tends towards zero as $m \to \infty$. Hence, for any distribution of within-study variances, the untruncated version of the DerSimonian and Laird estimate $\hat{\tau}_U^2$ tends in probability to τ^2 by virtue of Chebychev's inequality. The DerSimonian and Laird estimate, which is merely a truncated version which prevents the point estimate taking a value off the parameter space of τ^2, is therefore also consistent and the procedure suggested by DerSimonian and Laird is justified asymptotically.

3. The Distribution of the Standard Test Statistic Under the Null Hypothesis of no Treatment Effect

Although the conventional use of the test statistic R (Equation (1)) is justified asymptotically, the above analysis provides no guarantee that this will be appropriate for small sample sizes, that is, meta-analyses with small numbers of studies. In order to assess the suitability of comparing R to a standard normal distribution when testing the null hypothesis, using the DerSimonian and Laird estimate of τ^2, we will consider the special case where $\sigma_i^2 = \sigma^2$ for all i, that is, all studies are the same "size."

Although this is somewhat artificial, we will see later in Section 5 how the findings from this type of scenario can be used to give an indication of the significance level of the standard test for real meta-analyses where the study sizes inevitably differ. As discussed in the introduction, we will assume that the studies are sufficiently large to justify the usual normal approximation for the conditional distribution of $y_i|\mu_i$. As this type of approximation is used in many simple statistical methods, the circumstances under which this is appropriate are well understood and, as all studies are assumed to be the same size, we need only question if this is appropriate for the particular study size in question. Despite this it should be noted that if the study size is not sufficient to justify this, then the results obtained below become less or entirely invalid, depending on precisely how small the studies actually are.

Under this simplification we have that the y_i are iid from a normal distribution and therefore that all the usual standard results apply. In particular the distribution of the sample mean of the y_i, $\bar{y}$, and the corresponding sample variance, $s^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2/(n-1)$, are related to those of well-known distributions and are independently distributed. Furthermore, the untruncated DerSimonian and Laird estimate has the very simple form $\hat{\tau}_u^2 = s^2 - \sigma^2$. As the sample variance s^2 is a continuous random variable, and hence the probability that it takes the value σ^2 precisely is zero, there are two distinct possibilities for the more usual estimate $\hat{\tau}^2$: either this is truncated to zero (if $s^2 < \sigma^2$, so that a fixed effects model is adopted) or positive and equal to $\hat{\tau}_u^2$ (if $s^2 \geq \sigma^2$). Let E denote an indicator random variable for the event that a random effects procedure is adopted, that is, $E = 0$ if $\hat{\tau}^2 = 0$ and $E = 1$ otherwise. We therefore have that the distribution of R is given by

$$\begin{aligned} f(r) \;=\; & P(E=0)\, f(r|E=0) \\ & +\{1 - P(E=0)\}\, f(r|E=1). \end{aligned} \tag{4}$$

In order to obtain the distribution of R, under the null hypothesis of no treatment effect and for the case where $\sigma_i^2 = \sigma^2$ for all i, we must evaluate the three expressions on the right side of (4) under this hypothesis. As shown in the Appendix, this distribution is

$$\begin{aligned} f(r) \;=\; & \left[1 - \Gamma_1\left\{\frac{n}{2}, \frac{n-1+r^2}{2(1+w\tau^2)}\right\}\right] f_{\{t,(n-1)\}}(r) \\ & +\Gamma_1\left\{\frac{n-1}{2}, \frac{n-1}{2(1+w\tau^2)}\right\} \\ & \times \frac{1}{\sqrt{1+w\tau^2}}\phi\left(\frac{r}{\sqrt{1+w\tau^2}}\right), \end{aligned} \tag{5}$$

where

$$\Gamma_1(a,x) = \frac{1}{\Gamma(a)} \int_0^x t^{(a-1)} \exp(-t)dt,$$

$w = \sigma^{-2}$, $f_{(t,\upsilon)}(\cdot)$ denotes the density of Student's t distribution with υ degrees of freedom and $\phi(\cdot)$ denotes the standard normal density function; $\Gamma(\cdot)$ denotes the usual gamma function.

3.1 Interpreting the Distribution of R

It is interesting to note that the distribution of R in (5) depends only on the number of studies and $w\tau^2$, which provides a measure of the degree of between-study variance relative to the size of the studies. The dependence of a key test statistic on the product of w and τ^2, rather than these two variables separately, has already been noted by Jackson (2006b). Furthermore, Higgins and Thompson (2002) described an I^2 statistic that estimates the proportion of the studies' variance that is due to between-study variation. The corresponding true proportion is given by $w\tau^2/(1+w\tau^2)$, hence we will define $I^2 = w\tau^2/(1+w\tau^2)$. As the distribution of $f(r)$ under the null hypothesis depends directly on $w\tau^2$, and I^2 is a one-to-one function of this, the distribution of R can also be parameterized in terms of I^2, which is considered to provide a nicely interpretable value. For example, values of $I^2 < 0.3$ are considered mild degrees of between-study variation, while values well in excess of 0.5 are considerable. To express (5) in terms of I^2, $(1+w\tau^2)$ is replaced with $1/(1-I^2)$ giving

$$\begin{aligned} f(r) \;=\; & \left[1 - \Gamma_1\left\{\frac{n}{2}, \frac{(n-1+r^2)(1-I^2)}{2}\right\}\right] f_{\{t,(n-1)\}}(r) \\ & +\Gamma_1\left\{\frac{n-1}{2}, \frac{(n-1)(1-I^2)}{2}\right\} \\ & \times\sqrt{1-I^2}\,\phi\left(r\sqrt{1-I^2}\right). \end{aligned} \tag{6}$$

The standard two-tailed test of $H_0 : \mu = 0$ versus the alternative $H_1 : \mu \neq 0$ is provided by comparing the observed value of $-|R|$ with $Z_{\alpha/2}$, where $Z_{\alpha/2}$ denotes the $\alpha/2$ percentile of a standard normal distribution and α is the *nominal* significance level of the test. If $-|R|$ is less than $Z_{\alpha/2}$, then the null hypothesis is rejected, otherwise we accept this hypothesis. The *actual* significance level of the two-tailed standard test for a treatment effect, with a nominal significance level of α, is therefore

$$\alpha_s(I^2, n, \alpha) = 1 - \int_{Z_{\alpha/2}}^{-Z_{\alpha/2}} f(r)dr, \tag{7}$$

where $f(r)$ is given in (6). As the density in (6) is written in terms of well-known expressions and densities, the

Table 1. The actual significance level of the standard test for a treatment effect, assuming all studies are the same size and using a nominal level of 0.05.

I^2	$\alpha_s(I^2, 4, 0.05)$	$\alpha_s(I^2, 8, 0.05)$	$\alpha_s(I^2, 16, 0.05)$	$\alpha_s(I^2, 32, 0.05)$
0	0.037	0.038	0.040	0.042
0.15	0.049	0.049	0.050	0.050
0.3	0.064	0.061	0.058	0.056
0.5	0.088	0.076	0.066	0.059
0.75	0.123	0.089	0.069	0.059
0.9	0.140	0.091	0.069	0.059

actual significance level can easily be evaluated numerically. The actual significance level of a corresponding one-tailed test can also be obtained in a similar fashion.

4. Investigating the Significance Level of the Standard Test

As shown in the previous section, if all studies are the same size, then the actual significance level of the standard test, α_s, is a function of I^2, n, and α. It is therefore of considerable interest to adopt the standard nominal significance level of $\alpha = 0.05$ and use some indicative values of I^2 and n in order to illustrate the actual significance level of the test. Values of I^2 of 0 (no between-study variation), 0.15, and 0.3 (mild between-study variation), 0.5, 0.75, and 0.9 (considerable between-study variation) will be used for this process, in conjunction with $n = 4, 8, 16$, and 32, the same sample sizes discussed by Follmann and Proschan (1999). The resulting actual significance levels, $\alpha_s(I^2, n, 0.05)$, obtained directly from (7), are shown in Table 1.

The standard test is merely justified asymptotically and is not expected to provide the nominal significance level exactly. Under such circumstances, an actual significance level in the region of 0.04–0.06 is likely to be considered satisfactory. Evident from Table 1 is that mild heterogeneity provides such significance levels, as does $n = 32$. The difficulties are presented by considerable heterogeneity and sample sizes of 16 or less. The results indicate that the standard test does not provide a suitable significance level if there are only a small number of studies and considerable heterogeneity.

A suggested modification by Follmann and Proschan (1999) is to use the t distribution with $(n - 1)$ degrees of freedom as the null distribution of R. The significance levels resulting from this modification can easily be found from (7) but replacing $Z_{\alpha/2}$ by $t_{\{(n-1),\alpha/2\}}$, where $t_{\{(n-1),\alpha/2\}}$ denotes the corresponding percentile of the t distribution with $(n-1)$ degrees of freedom. As this percentile is a function of n and α alone, the actual significance level of the modified test remains a function of these variables and I^2 and will be denoted by α_m. The actual significance levels resulting from this modification, for the same range of values illustrated for the standard test in Table 1, are shown in Table 2.

The results in Tables 1 and 2 are inconsistent, indicating that the modification has a considerable impact. The difference is not particularly marked for $n \geq 16$ as the effect of using a t distribution for the null distribution is not so great when the degrees of freedom are sufficiently large. The important differences are for $n = 4$ and 8, which is a comparable sample size to many meta-analyses encountered in practice. For such sample sizes and mild heterogeneity, the standard test appears preferable on the basis of the actual significance level, while for considerable heterogeneity the modification is to be preferred.

Table 2. The actual significance level of the modified test for a treatment effect, assuming all studies are the same size and using a nominal level of 0.05.

I^2	$\alpha_m(I^2, 4, 0.05)$	$\alpha_m(I^2, 8, 0.05)$	$\alpha_m(I^2, 16, 0.05)$	$\alpha_m(I^2, 32, 0.05)$
0	0.001	0.013	0.026	0.034
0.15	0.002	0.019	0.033	0.042
0.3	0.004	0.026	0.041	0.047
0.5	0.010	0.037	0.047	0.050
0.75	0.028	0.048	0.050	0.050
0.9	0.044	0.050	0.050	0.050

The reasons for these observations are easily described, as the modification can be justified by adopting the untruncated version of the DerSimonian and Laird estimate and assuming that all studies are the same size. This provides a test statistic R that has an exact a t distribution under the null hypothesis (Follmann and Proschan 1999). The difficulty is, however, that this test does not generalize appropriately when all studies are not the same size while the DerSimonian and Laird procedure can be applied to any distribution of study sizes. Large values of I^2 provide substantial heterogeneity and therefore a test statistic that is distributed almost identically to a t distribution under the null hypothesis, as it is very unlikely that the DerSimonian and Laird estimate will require truncation under these circumstances. Hence in such cases a t distribution for the null distribution is more appropriate than a standard normal.

However, another extreme case is provided by $I^2 = 0$, indicating that a fixed effects model is appropriate. In this scenario the DerSimonian and Laird testing procedure is necessarily conservative, because the true variance of the sample mean, which is used to construct the test statistic, is approximately $1/S_1$. The DerSimonian and Laird procedure provides a nonnegative estimate of τ^2 which, when effectively used as the true value, provides a variance for the sample mean that is greater than or equal to true variance and hence inevitably results in a conservative test. This conservative nature of the standard test is exacerbated by using the modification of adopting a t distribution, with very few degrees of freedom, for the null distribution. We are forced to conclude that neither null distribution consistently outperforms the other.

5. More Realistic Scenarios

Although the analysis in the previous section is revealing, real meta-analyses do not have studies that are all the same size. An approach similar to that suggested by Jackson (2006b) will therefore be adopted. Using a representative study size providing a "typical" σ^2, or equivalently w, the results obtained in the previous section can be applied to a particular dataset as an approximation. Hence the results obtained above can be used to give an indication of the significance level of the standard test for meta-analyses more generally.

Three examples will be illustrated, all of which use the log odds ratio as the measure of treatment effect. Briefly, the first of these is from Follmann and Proschan (1999) and concerns eight studies that investigate the effectiveness of treatments for reducing serum cholesterol for preventing death in patients with no history of heart attacks. The second is from Jackson (2006b) and concerns nine studies that investigate the use of glycerol for preventing death in patients who have suffered an acute stroke. Finally, the third is from the Cochrane Collaboration and concerns just four studies investigating the use of amisulpride, compared to typical antipsychotics, to treat schizophrenia where the outcome of interest is "less than much improved clinical global impression" (Mota Neto et al. 2002, analysis 02.04). These examples provide some really rather disparate study sizes in order to exert considerable pressure on any approximation that requires studies of similar size; see Table 3 for further details of the values of σ_i^2 for these studies, where the total number of patients in each study is also provided in parentheses. Larger studies (smaller variances) generally involve more patients, but note that all three examples involve meta-analyses of binary events and hence study variances depend on the event rates, as well as patient numbers. Table 3 serves to emphasize that study "size" refers directly to the studies' variances. It will be assumed that these study sizes are suitable for adopting the usual normal approximations as, although some of the studies are really rather small, the values used are merely intended to provide some realistically variable distributions of study sizes that can be encountered in practice.

Table 3. The within-study variances, that is, the "size" of the studies, for the three examples. The total number of patients in each study are also shown in parentheses.

Study (i)	Serum cholesterol	Glycerol	Amisulpride
1	0.011 (10627)	0.09 (216)	0.09 (199)
2	0.013 (3806)	0.16 (173)	0.09 (191)
3	0.016 (9057)	0.17 (106)	0.13 (132)
4	0.030 (4081)	0.23 (113)	0.15 (129)
5	0.040 (846)	0.24 (93)	
6	0.053 (2278)	0.39 (56)	
7	0.076 (8245)	0.54 (62)	
8	1.353 (118)	0.62 (27)	
9		2.77 (38)	

A representative σ^2 is the "typical" within-study variance provided by Higgins and Thompson (2002) which is given by

$$\sigma_t^2 = \frac{(n-1)S_1}{S_1^2 - S_2}. \tag{8}$$

In addition to the justification given by Higgins and Thompson, there is a further motivation for using (8) as an indicative value. E$[Q]$ is given by (2), and the same E$[Q]$ is obtained if all studies are the same size and have σ_t^2 as given in (8). Hence a suitable study size, in relation to the value of τ^2, is given in a natural way by (8). For meta-analyses with inevitably different study sizes, we will therefore define $w = \sigma_t^{-2}$, and $I^2 = w\tau^2/(1 + w\tau^2)$ as before. As the serum cholesterol example has eight studies, and provides $w = 36.9$,

Table 4. Actual significance levels, obtained approximately by simulation, for the standard test using the three examples and a nominal significance level of 0.05

I^2	Serum cholesterol	Glycerol	Amisulpride
0	0.036	0.037	0.037
0.15	0.056	0.054	0.050
0.3	0.073	0.072	0.066
0.5	0.093	0.088	0.090
0.75	0.103	0.095	0.126
0.9	0.104	0.094	0.142

Table 5. Actual significance levels, obtained approximately by simulation, for the modified test using the three examples and a nominal significance level of 0.05

I^2	Serum cholesterol	Glycerol	Amisulpride
0	0.013	0.014	0.001
0.15	0.022	0.025	0.002
0.3	0.037	0.037	0.005
0.5	0.050	0.050	0.011
0.75	0.061	0.057	0.029
0.9	0.061	0.057	0.044

the third columns of Tables 1 and 2 give an indication of the significance levels of the respective tests in terms of $I^2 = 36.9\tau^2/(1 + 36.9\tau^2)$. For the glycerol example, $w = 3.95$ and, although $n = 9$ is not tabulated in Tables 1 and 2, the significance levels for this sample size do not differ much from the case where $n = 8$. Hence the third columns of Tables 1 and 2 also give an indication of the significance level of the tests in terms of $I^2 = 3.95\tau^2/(1 + 3.95\tau^2)$. Finally, for the amisulpride data, $w = 9.0$, and hence the second columns of Tables 1 and 2 give an indication of the significance levels of the tests in terms of $I^2 = 9.0\tau^2/(1 + 9.0\tau^2)$.

Now that the studies are not all the same size, however, the sampling distribution of R is more complicated and Tables 1 and 2 merely provide an indication of the significance level. Despite this, the actual significance levels of the tests can be obtained approximately by simulation, following the type of procedure Hardy and Thompson (1998) adopted when investigating the standard test for heterogeneity. Under the null hypothesis that $\mu = 0$, the distributions of the estimates are $y_i \sim N(0, \sigma_i^2 + \tau^2)$. As $I^2 = w\tau^2/(1 + w\tau^2)$, the distributions of the estimates are equivalently $y_i \sim N(0, \sigma_i^2 + I^2/\{w(1 - I^2)\})$. These observations can be simulated using a particular value of I^2, from which $\hat{\tau}^2$ and then R can be evaluated. This procedure can be repeated, providing many evaluations of R under the null hypothesis, and the approximate significance level is provided by the proportion of significant results. The approximate significance levels of the standard test using 100,000 simulations, and a nominal significance level of $\alpha = 0.05$, are shown for all three example datasets in terms of I^2 in Table 4. The corresponding significance levels of the modified test are also shown in Table 5.

The results are encouraging as the examples have very disparate study sizes and yet the significance levels obtained by simulation are similar to the values shown in the corresponding columns of Tables 1 and 2. In particular note the very large actual significance level, assuming $I^2 = 0.9$, for the amisulpride example and using the standard test; there are only four studies present, so this finding is predicted by the theory developed in the previous section. The power of a statistical test depends broadly on the amount of information available, hence by using the results obtained in the previous section, with the typical within-study variance, one may consider an analogous meta-analysis with the same number of studies of comparable size as one which has been observed. This provides a similar amount of information and hence power, suggesting that the results obtained in Section 4 can be used to give an indication of the significance level of the standard and modified tests for real meta-analyses where the study sizes inevitably differ.

6. Conclusions

The standard meta-analysis models include the more basic model where the study estimates are iid from a normal distribution, which implies that all studies are the same size. By considering this special case, analytical results can be obtained and used to inform meta-analyses where the study sizes differ. This has been illustrated using three examples with really rather disparate study sizes but it is of interest to investigate just how different these sizes can be for the analytical results to be used as an approximation. For example, a cautionary tale is told by a simulation study, conducted in the same way as those described in Section 5, involving the scenario where there are four studies providing estimated within-study variances of 0.1, 0.1, 5, and 5. Using $I^2 = 0.75$, for example, gave estimated significance levels of 0.215 and 0.092 for the standard and modified tests, with nominal levels of 0.05, respectively. Although these values compare poorly with those shown in Tables 1 and 2, this is a rather unusual situation, and one that was created especially to put very considerable pressure on the assumption made. Despite this, it is anticipated that the occasional real meta-analysis will provide circumstances under which the analytical results provide a poor approximation. If the study sizes are truly disparate, or provide a

peculiar empirical distribution, simulation studies should also be performed in order to inform those contemplating conducting the standard test.

The results indicate that the standard test for a treatment effect is not suitable for datasets with small numbers of studies and considerable between-study variation. For analyses where there are a moderate number of studies, or convincing evidence that the between-study variation is mild, the standard test is appropriate. By contrast, the modified testing procedure requires moderate numbers of studies or considerable between-study variation, and hence neither approach consistently outperforms the other. It is not surprising that small numbers of studies provide most of the difficulties, as the standard procedure relies heavily on asymptotic results. The findings indicate that the modification suggested by Follmann and Proschan provides a considerable improvement under the circumstances described above.

It should be emphasized again, however, that the modeling assumes that the usual normal conditional distributions are appropriate, which requires sufficiently large studies. If this is not the case, then more appropriate conditional distributions should be used to obtain significance levels. For example, a noncentral hypergeometric distribution could be used for studies whose results are summarized as two-by-two tables. Indeed, the "actual" significance levels obtained might be more reasonably described as the rather unwieldy "approximately actual significance levels, depending on the size of the studies."

As this is the one of the first articles to examine analytically the exact distributions resulting from standard meta-analysis procedures, it has considered perhaps the simplest type of formal statistical inference, a hypothesis test. Further investigation is possible, such as exploring the coverage probability of confidence intervals. Bayesian analyses may also be carried out using Markov chain Monte Carlo methodology, as described by Normand (1999). Although these do not require the kind of asymptotic results conventionally used in the frequentist framework adopted here, they are subject to other issues and concerns.

Finally, it seems reasonable to require that any meta-analysis methodology should perform satisfactorily when all studies are the same size, and should generalize appropriately when they are not. It is therefore a necessary condition that any proposed methodology must be suitable in this simplified situation, which may provide the means to remove particular methods from the meta-analysis armory and this is currently being investigated. It is also of considerable interest to go back through the medical literature and determine if random effects models have been used with less than, say, ten studies. If such analyses are commonplace, then it is probable that inferences, and therefore practices which are implemented because of these, are open to unexpected questions, which has worrying implications for us all.

A. Appendix: The Derivation of the Density of R Under the Null Hypothesis of no Treatment Effect and Assuming that all Studies are the Same Size

We evaluate the terms on the right side of (4). First note that $P(E=0) = P(\hat{\tau}^2 = 0) = P(\hat{\tau}_u^2 < 0) = P(s^2 < \sigma^2)$. Furthermore, $(n-1)s^2/(\sigma^2+\tau^2) \sim \chi^2_{n-1}$, so $P(E=0) = P\{\chi^2_{n-1} < (n-1)/(1+w\tau^2)\}$, where $w = \sigma^{-2}$. Defining the incomplete gamma function as

$$\Gamma_1(a,x) = \frac{1}{\Gamma(a)} \int_0^x t^{(a-1)} \exp(-t)dt,$$

where $\Gamma(\cdot)$ denotes the usual Gamma function, this provides

$$P(E=0) = \Gamma_1\left(\frac{n-1}{2}, \frac{n-1}{2(1+w\tau^2)}\right). \qquad \text{(A.1)}$$

Note that the derivation of this probability does not assume the null hypothesis, and hence this is true for all values of μ and τ^2.

Conditioning on $E=0$ provides $\hat{\tau}^2 = 0$; substituting this estimate into (1), and noting that $\sigma_i^2 = \sigma^2$ for all i, gives $R = \bar{y}/\sqrt{\sigma^2/n}$, where $\bar{y}$ denotes the sample mean. The distribution of $\{(\bar{y}/\sqrt{\sigma^2/n})|E=0\}$ is equivalent to the distribution of $\bar{y}/\sqrt{\sigma^2/n}$ conditional on the event that $s^2 < \sigma^2$, and the unconditional distribution of $\bar{y}/\sqrt{\sigma^2/n}$, as $\bar{y}$ and s^2 are independent. Under the null hypothesis, $\bar{y} \sim N\{0, (\sigma^2+\tau^2)/n\}$, so we have that $(R|E=0)$ is normally distributed, centered at zero but with variance $(1+w\tau^2)$. We obtain

$$f(r|E=0) = \frac{1}{\sqrt{1+w\tau^2}}\phi\left(\frac{r}{\sqrt{1+w\tau^2}}\right) \qquad \text{(A.2)}$$

under the null hypothesis, where $\phi(\cdot)$ denotes the standard normal density function.

Conditioning on $E=1$, and noting that $\sigma^2 = \sigma_i^2$ for all i, provides $\hat{\tau}^2 = \hat{\tau}_u^2 = s^2 - \sigma^2$. Substituting this estimate into (1) provides $R = \bar{y}/\sqrt{s^2/n}$. This is a similar expression as for the case where $E=0$; s^2 has replaced σ^2 in the denominator of R. A little rearrangement gives

$$R = \frac{\frac{\sqrt{n}\bar{y}}{\sqrt{\sigma^2+\tau^2}}}{\sqrt{\frac{s^2(n-1)}{(\sigma^2+\tau^2)(n-1)}}}.$$

Under the null hypothesis $X_1 = \frac{\sqrt{n}\bar{y}}{\sqrt{\sigma^2+\tau^2}} \sim N(0,1)$ and $X_2 = \frac{s^2(n-1)}{\sigma^2+\tau^2} \sim \chi^2_{n-1}$, where X_1 and X_2 are independent. Furthermore, $R = X_1/\sqrt{X_2/(n-1)}$, conditional on $E = 1$, which is equivalent to the event that $s^2 \geq \sigma^2$ and therefore the event that $X_2 \geq (n-1)/(1+\omega\tau^2) = c$. We therefore require the probability density function of $\left(X_1/\sqrt{X_2/n-1}\middle| E=1\right)$. This expression can be derived as the density $f(x_1, x_2|E=1)$ can be found as $f(x_1)f(x_2)I_{(x_2\geq c)}/P(X_2 \geq c)$, where $I_{(x_2\geq c)} = 1$ if $x_2 \geq c$ and $I_{(x_2\geq c)} = 0$ otherwise. Furthermore, $P(X_2 \geq c) = 1 - \Gamma_1\{(n-1)/2, c/2\}$. Changing variables in the usual way with $R = X_1/\sqrt{X_2/(n-1)}$ and $R_2 = X_2$, and integrating out R_2 provides

$$f(r|E=1) = f_{\{t,(n-1)\}}(r)\left[\frac{1-\Gamma_1\left\{\frac{n}{2}, \frac{n-1+r^2}{2(1+\omega\tau^2)}\right\}}{1-\Gamma_1\left\{\frac{n-1}{2}, \frac{n-1}{2(1+\omega\tau^2)}\right\}}\right], \tag{A.3}$$

under the null hypothesis, where $f_{(t,\upsilon)}(\cdot)$ denotes the density of Student's t distribution with υ degrees of freedom. Substituting (A.1), (A.2), and (A.3) into (4) gives (5).

[Received November 2006. Revised August 2007.]

References

Biggerstaff, B.J., and Tweedie, R.L. (1997), "Incorporating Variability of Estimates of Heterogeneity in the Random Effects Model in Meta-analysis," *Statistics in Medicine*, 16, 753–768.

Dersimonian, R., and Laird, N. (1986), "Meta-Analysis in Clinical Trials," *Controlled Clinical Trials*, 7, 177–188.

Emerson, J. D., Hoaglin, D. C., and Mosteller, F. (1996), "Simple Robust Procedures for Combining Risk Differences in Sets of 2 × 2 Tables," *Statistics in Medicine*, 15, 1465–1488.

Follmann, D.A., and Proschan, M.A. (1999), "Valid Inference in Random Effects Meta-Analysis," *Biometrics*, 55, 732–737.

Hardy, R.J., and Thompson, S.G. (1996), "A Likelihood Approach to Meta-analysis With Random Effects," *Statistics in Medicine*, 15, 619–629.

——— (1998), "Detecting and Describing Heterogeneity in Meta-analysis," *Statistics in Medicine*, 17, 841–856.

Hartung, J., and Knapp, G. (2001), "A Refined Method for the Meta-analysis of Controlled Clinical Trials with Binary Outcome," *Statistics in Medicine*, 20, 3875–3889.

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

Jackson, D. (2006a), "The Implications of Publication Bias for Meta-analysis' Other Parameter," *Statistics in Medicine*, 25, 2911–2921.

——— (2006b), "The Power of the Standard Test for the Presence of Heterogeneity in Meta-analysis," *Statistics in Medicine*, 25, 2688–2699.

Mota Neto, J.I.S., Lima, M.S., and Soares, B.G.O. (2002), "Amisulpride for Schizophrenia," *Cochrane Database of Systematic Reviews*, Issue 2

Normand, S.T. (1999), "Meta-analysis: Formulating, Evaluating, Combining, and Reporting," *Statistics in Medicine*, 18, 321–359.

Sutton, A.J., Abrams, K.R, Jones, D.R., Sheldon, D.R., and Song, F. (2002), *Methods for Meta-analysis in Medical Research*, New York: Wiley.

Van Houwelingen, H.C., and Zwinderman, K.H. (1993), "A Bivariate Approach to Meta-analysis," *Statistics in Medicine*, 12, 2273–2284.

About the Authors

Dr. Dan Jackson is Research Scientist, MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR United Kingdom (E-mail: *daniel.jackson@mrc-bsu.cam.ac.uk*).