Biostatistics (2002), **3**, 4, *pp.* 445–457
Printed in Great Britain

Some general points in estimating heterogeneity variance with the DerSimonian–Laird estimator

DANKMAR BÖHNING*, UWE MALZAHN, EKKEHART DIETZ, PETER SCHLATTMANN
Department of Epidemiology, Free University Berlin, Haus 562, Fabeckstr. 60-62, 14195 Berlin, Germany
boehning@zedat.fu-berlin.de

CHUKIAT VIWATWONGKASEM
Department of Biostatistics, Faculty of Public Health, Mahidol University, Bangkok, Thailand
phcvw@mucc.mahidol.ac.th

ANNIBALE BIGGERI
Department of Statistics 'G Pareti', University of Florence, Italy
abiggeri@stat.ds.unifi.it

SUMMARY

In this paper we consider estimating heterogeneity variance with the DerSimonian–Laird (DSL) estimator as typically used in meta-analysis. In its general form the DSL estimator requires inverse population-averaged study-specific variances as weights, in which case the estimator is unbiased. It has become common practice, however, to use estimates of the study-specific variances instead of their population-averaged versions. This can lead to considerable bias. Simulations illustrate these findings.

Keywords: Bias in DerSimonian–Laird estimator; Estimator of heterogeneity variance; Meta-analysis; Pooled analysis; Population averaged study-specific weights; Study-specific weights.

1. INTRODUCTION

We are interested in the following simple two-level model. Let $f(x, \theta, \sigma^2)$ denote a parametric density for some random quantity X where θ is a parameter of interest and σ^2 is a nuisance parameter which might or might not be present in the model. Typically, f could be a normal density in which case a nuisance parameter (variance) is present or f could be the binomial in which case the conditional variance would be specified by the parameter of interest $\mathrm{Var}(X|\theta, n) = \theta(1-\theta)/n$. Note that is important to emphasize at this stage that moments are computed with respect to the *conditional* distribution given the parameters θ and σ^2, typically $E(X|\theta) = \theta$ and $\mathrm{Var}(X|\theta, \sigma^2) = \sigma^2 v(\theta)$, where $v(\theta)$ is a known function depending on θ only.

In the second step, it is assumed that θ is not constant, but is varying itself according to some not further specified distribution P for which only the moments $E_P(\theta) = \mu$ and $\mathrm{Var}_P(\theta) = \tau^2$ are assumed to exist. Consequently, we are led to a *marginal* or *unconditional* distribution $f(x, P) = \int f(x, \theta) P(\mathrm{d}\theta)$. It is easy to show that the unconditional mean, e.g. the mean with respect to $f(x, P)$, is $E(X) = E_P(\theta) = \mu$

*To whom correspondence should be addressed

and also the unconditional variance

$$\text{Var}(X) = \int \text{Var}(X|\theta, \sigma^2) P(\text{d}\theta) + \text{Var}_P(\theta) = \sigma^2 \int v(\theta) P(\text{d}\theta) + \text{Var}_P(\theta) \tag{1}$$

or simply

$$\text{Var}(X) = \nu^2 + \tau^2. \tag{2}$$

Equation (1) is called the *latent decomposition of variance* consisting of two terms. The first one, $\nu^2 = \sigma^2 \int v(\theta) P(\text{d}\theta)$, represents the variation conditional on the value of θ, and then averaged over θ. The second term, $\tau^2 = \text{Var}_P(\theta)$, represents the variation of θ itself. The decomposition is called *latent*, because outcomes from the (latent) distribution P are not observed directly. A classical example consists of assuming the conditional distribution of X to be normal with mean θ and variance σ^2, the latter assumed to be known. In this case, we have that $v(\theta) = 1$, and consequently, $\nu^2 = \int \text{Var}(X|\theta, \sigma^2) P(\text{d}\theta) = \sigma^2 = \text{Var}(X|\theta, \sigma^2)$, and the latent variance decomposition is simply $\text{Var}(X) = \sigma^2 + \tau^2$. Note that in this important but *special* case population-averaged and conditional variance coincide. It is usually this setting for which the DerSimonian–Laird (DSL) estimator is considered (see Brockwell and Gordon, 2001). Suppose that a random sample $x_1, x_2, \ldots, x_k$ of size k is available with associated (known) variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2$. In a meta-analytic setting this sample would represent a collection of k independent studies, and in each study a statistic x_j is measured with (known) standard error σ_j, $j = 1, \ldots, k$. Then the DSL estimator is given as (DerSimonian and Laird, 1986)

$$\hat{\tau^2} = \frac{\chi^2 - (k-1)}{\sum_{i=1}^k w_i - \sum_{i=1}^k w_i^2 \big/ \sum_{i=1}^k w_i} \tag{3}$$

where $w_i = 1/\sigma_i^2$, $\chi^2 = \sum_{i=1}^k w_i (x_i - \hat{\mu})^2$, and $\hat{\mu} = \sum_{i=1}^k w_i x_i \big/ \sum_{i=1}^k w_i$. Note that the estimator (3) is *unbiased* by construction. Since (3) can attain negative values a *truncated* version is considered, $\hat{\tau^2_+} = \max\{0, \hat{\tau^2}\}$, which is no longer unbiased. However, bias will only occur in situations close to homogeneity for which the estimator of heterogeneity variance is of diminished interest. In fact, Brockwell and Gordon (2001) suggest a 'Q-based method' in which the DSL estimator is considered only when the χ^2-statistic is significant at the 5% level of the χ^2-distribution with $(k-1)$ degrees of freedom. In this case, χ^2 is necessarily larger than its theoretical expected value $(k-1)$, and (3) is positive.

However, *two* issues are of more concern.

- The first issue has to do with the frequently occurring fact that the variances are not known, but rather estimates $\hat{\sigma}_i^2$, $i = 1, \ldots, k$ are used in the construction and (3) becomes

$$\hat{\tau^2} = \frac{\chi^2 - (k-1)}{\sum_{i=1}^k \hat{w}_i - \sum_{i=1}^k \hat{w}_i^2 \big/ \sum_{i=1}^k \hat{w}_i} \tag{4}$$

  where $\hat{w}_i = 1/\hat{\sigma}_i^2$, $\chi^2 = \sum_{i=1}^k \hat{w}_i (x_i - \hat{\mu})^2$, and $\hat{\mu} = \sum_{i=1}^k \hat{w}_i x_i \big/ \sum_{i=1}^k \hat{w}_i$. The estimator given in (4) (untruncated) is no longer unbiased. Brockwell and Gordon (2001, p. 837) write

  > For both the fixed and random effects methods, inference is carried out ignoring the sampling errors in the individual study variances. Estimated values $\hat{\sigma}_i^2$ are used without modification to the form of $\hat{\mu}$, its variance or distribution.

  In fact, the inferential problem of replacing variance parameters by their estimates has been observed by several authors. Senn (2000) writes

> Consider the case where there are many small, equally sized centers and homoscedasticity applies. The optimal approach is to weight the centers equally. Fixed-effects meta-analysis will weight inversely proportional to the observed variance. In so doing it will produce an estimator whose true variance is higher than that produced by equal weighting, but which will appear to be lower [*x* . . .]. As a consequence the associated significance tests and confidence intervals will be anticonservative.

Similarly, Böhning and Sarol (2000) show that for the multicentre study the optimally weighted estimate of risk difference loses its optimality when the weights are replaced by their centre estimates and other (non-random) weights outperform the former one.

- The second issue is concerned with the fact that frequently outcome measures of interest have conditional variances that are functions of the parameter of interest. Consider as a simple example the standardized mortality ratio SMR $= X/e$, where e is the number of expected cases computed from an external reference population. It is not unusual to assume a conditional Poisson distribution for X with parameter $e\theta$, so that $E(X|\theta, e) = e\theta$ equals the conditional variance $\mathrm{Var}(X|\theta, e) = e\theta$, or $E(\mathrm{SMR}\,|\theta, e) = \theta$ and $\mathrm{Var}(\mathrm{SMR}\,|\theta, e) = \theta/e$. Now, in contrast to the normal case, we yield a different result when taking the population averaged value $\nu^2 = \sigma^2 \int v(\theta) P(\mathrm{d}\theta) = \mu/e$. To demonstrate the different consequences, suppose that in a meta-analysis, in occupational medicine say, k SMR-values are available: $x_1/e_1, x_2/e_2, \ldots, x_k/e_k$. Conventional practice in meta-analysis of SMR-studies would use $\hat{w}_i = smr_i/e_i$, whereas one should use $\hat{\mu}/e_i$ for $i = 1, \ldots, k$, where $\hat{\mu}$ is a suitable estimate of μ such as $\sum_i x_i \big/ \sum_i e_i$.

The first issue of the two mentioned above is specifically a problem for small sample sizes in which the variances are estimated with low precision. Kennward and Roger (1997) discuss the effect of replacing the true variance–covariance matrix for small-sample inference by its estimate for the fixed-effect model. However, if sample sizes increase the problem diminishes. This is in contrast to the second issue, for which the bias persists with increasing sample sizes unless the correct forms of the study-specific variances are used.

The paper is organized as follows. In Section 2 the problem of the correct forms of the study-specific variances is exemplified for various situations. Section 3 presents as an example a meta-analysis in which each study contributes a binomial rate as outcome measure. Conventional study-specific variance estimates are contrasted with their (correct) population-averaged forms. Section 4 provides the DSL estimator in its general form and covers its application for several examples. In Section 5 a simulation study illustrates the amount of bias that can occur if incorrect variance estimates are used.

2. THE OCCURRENCE OF HETEROGENEITY

Suppose that a quantity of interest is measured in k independent studies such that for each study an estimate for the measure of interest with its standard error is available (Petitti, 1994; Normand, 1999). This situation is outlined in Figure 1 with $k = 4$ studies. One might imagine that each study represents only a certain part of the population for which it might be representative. Considering the variation of θ in addition, might lead to a model which is adequate for a more general population. The variation of the study-specific means might be continuous or discrete, leading to a continuous or discrete heterogeneity distribution P. In the following we demonstrate the latent variance decomposition (2) with some examples.

X normal with mean θ and variance σ^2. This case has been mentioned already, so we touch upon it only briefly. We have that $v(\theta) = 1$, and consequently $\nu^2 = \int \mathrm{Var}(X|\theta, \sigma^2) P(\mathrm{d}\theta) = \sigma^2$, and the latent variance decomposition is simply $\mathrm{Var}(X) = \sigma^2 + \tau^2$.

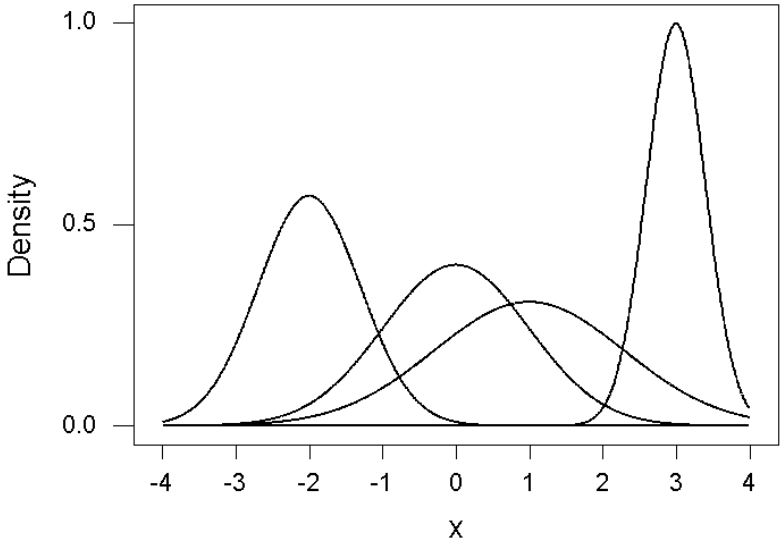

Fig. 1. Conditional distribution of measure of interest in four studies.

X binomial with parameters θ and n. Here, the conditional variance of X/n is $v(\theta) = \frac{\theta(1-\theta)}{n}$ and $\sigma^2 = 1$. Thus, $\nu^2 = \int \frac{\theta(1-\theta)}{n} P(\mathrm{d}\theta) = \mu/n - \int \theta^2/n P(\mathrm{d}\theta) = \mu/n - \tau^2/n - \mu^2/n = \frac{\mu(1-\mu)}{n} - \tau^2/n$, and $\mathrm{Var}(X/n) = \frac{\mu(1-\mu)}{n} + \left(1 - \frac{1}{n}\right)\tau^2$.

X Poisson with parameter θe. In the case of the Poisson with parameter θe, where e is assumed to be known, we find that $\mathrm{Var}(X/e|\theta, e) = \theta/e$. Frequently, epidemiologists consider the SMR $= X/e$, where X is the observed number of deaths in a study population and e is the expected number of deaths computed from a reference population. $E(X/e) = \theta$ is one form of relative risk (risk of death in study population relative to risk of death in the reference population) and represents one important outcome measure for many disciplines. The SMR is not only used in epidemiology, occupational medicine, or other medical disciplines, it is also prominent in demography, sociology and population studies. Keiding (1987) and Hoem (1987) provide excellent reviews of the SMR from its roots to its modern applications and theoretical developments. Here, we find $\nu^2 = \int \theta/e P(\mathrm{d}\theta) = \mu/e$, and the latent variance decomposition is $\mathrm{Var}(\mathrm{SMR}) = \mu/e + \tau^2$.

Standardized mean difference. Suppose a treatment is studied and some quantitative mean outcome is compared beween a treatment group and a control group on the basis of the *standardized mean difference* $X = (\overline{x}_T - \overline{x}_C)/s$, which estimates $\theta = (\theta_T - \theta_C)/\sigma$, where σ is the common standard deviation in treatment and control groups, and θ_T (θ_C) is the population mean in the treatment (control) group. Here, s^2 is defined in a pooled fashion, $s^2 = \{(n_T - 1)s_T^2 + (n_C - 1)s_C^2\}/(n - 2)$, as conventionally suggested (Cooper and Hedges, 1994; Shadish and Haddock, 1994). Here, s_T^2 (s_C^2) are the sample variances in treatment (control) group, n_T (n_C) are the sample sizes in the two groups, and $n = n_T + n_C$. The variance of $\hat{\theta} = (\overline{x}_T - \overline{x}_C)/s$ is readily provided as (Hedges and Olkin, 1985)

$$\mathrm{Var}(\hat{\theta}|\theta, n_T, n_C) \approx \frac{1}{n_T} + \frac{1}{n_C} + \frac{\theta^2}{2n}.$$

Table 1. *Meta-analysis of hyperdynamic therapy on the basis of* $k = 14$ *studies* (*Zhou* et al., 1999)

Study number	x_i	n_i	Study number	x_i	n_i
1	16	17	2	10	12
3	4	8	4	43	58
5	10	10	6	25	42
7	13	14	8	12	12
9	22	41	10	4	5
11	5	6	12	18	23
13	58	68	14	6	10

Therefore, the population-averaged variance is found as

$$\int \mathrm{Var}(\hat{\theta}|\theta, n_T, n_C) P(\mathrm{d}\theta) \approx \frac{1}{n_T} + \frac{1}{n_C} + \int \frac{\theta^2}{2n} P(\mathrm{d}\theta) = \frac{1}{n_T} + \frac{1}{n_C} + \frac{\mu^2 + \tau^2}{2n} \tag{5}$$

so that the unconditional variance of $X = \hat{\theta}$ is $\mathrm{Var}(X) \approx \frac{1}{n_T} + \frac{1}{n_C} + \frac{\mu^2}{2n} + \left(1 + \frac{1}{2n}\right)\tau^2$, because of the fact that $E(\hat{\theta}|\theta) \approx \theta$.

3. An example from meta-analysis

To illustrate the issues discussed above we look at a meta-analysis by Pritz *et al.* (1996) which is also used by Zhou *et al.* (1999) in their discussion on methods for combining rates. This meta-analysis investigates the effectiveness of hyperdynamic therapy in treating cerebral vasospasm. According to Zhou *et al.* (1999) there are 14 studies and Table 1 provides their data. Suppose we write the $k = 14$ proportions or success rates given in Table 1 as $\hat{\theta}_i = \frac{x_i}{n_i}$, $i = 1, \ldots, k$. Then, it has become common practice to use the estimate $\hat{\theta}_i(1 - \hat{\theta}_i)/n_i$ as an estimate of the study-specific variance. This is wrong with respect to *two* aspects. Firstly, $\hat{\theta}_i(1 - \hat{\theta}_i)/n_i$ estimates the *conditional* study-specific variance $\theta_i(1 - \theta_i)/n_i$, instead of the population-averaged variance $v_i^2 = \int \frac{\theta_i(1-\theta_i)}{n_i} P(\mathrm{d}\theta_i)$. Note that v_i^2 and $v(\theta_i) = \theta_i(1 - \theta_i)/n_i$ will only be close if the variance of the heterogeneity distribution P is small. In this example, as we will see further below, there is strong evidence of heterogeneity, and both types of variances differ considerably. Secondly, the replacement of the unknown study-specific success rate θ_i adds instability. As an extreme form of variation we note that in *two* studies this variance is estimated to be *zero*. It was mentioned by one of the referees that in this case the estimated study-specific variances need to be corrected whereas this correction is not required by the population-averaged study-specific variance, which adds to their advantage. We have seen in Section 2 that the population-averaged study-specific variance is given as $v_i^2 = \int \frac{\theta(1-\theta)}{n_i} P(\mathrm{d}\theta) = \frac{\mu(1-\mu)}{n_i} - \tau^2/n_i$. Unfortunately, this population-averaged variance involves τ^2 itself. Because $\mathrm{Var}(X_i/n_i) = \frac{\mu(1-\mu)}{n_i} + \left(1 - \frac{1}{n_i}\right)\tau^2$ it appears justified to use the approximation $v_i^2 \approx \frac{\mu(1-\mu)}{n_i}$, which can easily be estimated by $\hat{\mu}(1 - \hat{\mu})/n_i$ with $\hat{\mu}$ being the pooled estimator $\sum_{i=1}^k x_i \big/ \sum_{i=1}^k n_i$ which is in our case $\hat{\mu} = 0.7546$. Note that the pooled estimator $\sum_{i=1}^k x_i \big/ \sum_{i=1}^k n_i$ can be written as $\sum_i w_i \frac{x_i}{n_i} \big/ \sum_i w_i$, where $w_i = \frac{n_i}{\mu(1-\mu)-\tau^2} = 1/v_i^2$ which shows that this pooled estimator is the right choice for estimating μ. Figure 2 compares the two variance estimates $\hat{\mu}(1 - \hat{\mu})/n_i$ and $\hat{\theta}_i(1 - \hat{\theta}_i)/n_i$. Evidently, the variances differ unless the sample size is large as is the case for studies 4, 6, 9, 13.

It might be interesting at this stage to relate the issue under discussion to the area of *heterogeneity* tests. In this case, the χ^2-test of heterogeneity can be given in several versions. We mention two of these.

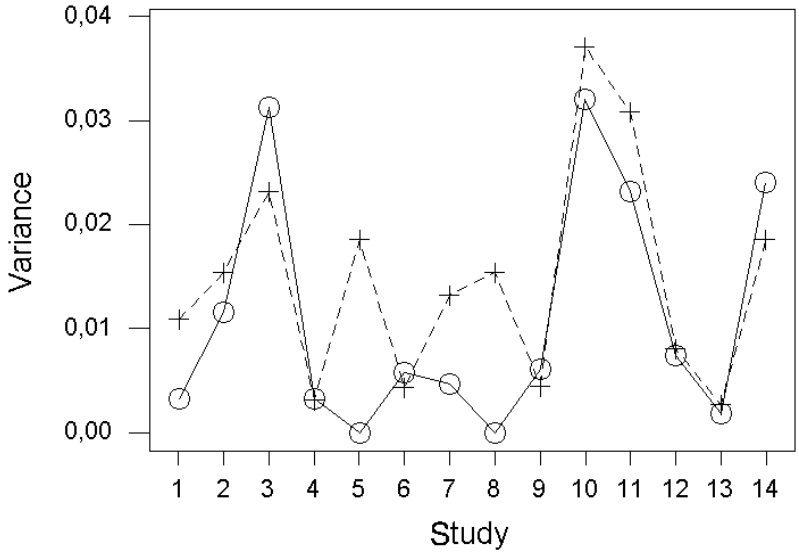

Fig. 2. Two different types of variance from a meta-analysis with 14 studies on the success rate of hyperdynamic therapy in treating cerebral vasospasm: study-specific variance (circle) and population-averaged variance (+).

Let $\chi^2 = \sum_{i=1}^k \hat{w}_i (x_i/n_i - \hat{\mu})^2$, with $\hat{\mu} = \sum_{i=1}^k x_i \big/ \sum_{i=1}^k n_i$. Then, with $\hat{w}_i = 1/\hat{v_i^2} = [\hat{\mu}(1-\hat{\mu})/n_i]^{-1}$ one version of this test is given which is approximately χ^2-distributed with $k-1$ degrees of freedom under homogeneity (P is a one-point distribution) as discussed in Collett (1999). This is even true for small n_i, but large number of studies k (Potthoff and Whittinghill, 1966). This property breaks down completely if a different version of this test is used, namely one using the χ^2-statistic with weights $\hat{w}_i = [\hat{\theta}_i(1-\hat{\theta}_i)/n_i]^{-1}$, where $\hat{\theta}_i = x_i/n_i$. (Here, the additional problem arises that the variance is estimated to be 0 if $x_i = 0$ or $x_i = n_i$. If this occurs 0 is replaced by 0.5 and $x_i = n_i$ by $n_i - 0.5$, respectively, as conventionally recommended.) This statistic is only χ^2-distributed with $k - 1\,df$ if k is fixed and the n_i become large, but not if n_i is small and k becomes large. For the data of Table 1 we find a value of $\chi^2 = 37.375$ with p-value $= 0.000\,36$, using the first version, and a value of $\chi^2 = 59.193$ with p-value $= 0.000\,00$, using the second version. This indicates that the results of the two tests can be quite different, and only using the first version will avoid the occurrence of artefacts (Potthoff and Whittinghill, 1966; Böhning, (2000, The flaw in χ^2-heterogeneity tests: a revisit of the Neyman–Scott problem? Technical Report)).

4. THE DERSIMONIAN–LAIRD ESTIMATOR

In this section the DSL estimator is considered in its general form. Suppose that a random sample $x_1, x_2, \ldots, x_k$ of size k is available with associated variances $v_1^2, v_2^2, \ldots, v_k^2$. In a meta-analytic setting this sample would represent a collection of k independent studies, and in each study a statistic x_j is measured with standard error v_j, $j = 1, \ldots, k$. Then the following result holds:

$$E(\chi^2) = (k-1) + \tau^2 \left(\sum_{i=1}^k w_i - \frac{\sum_{i=1}^k w_i^2}{\sum_{i=1}^k w_i} \right) \tag{6}$$

where $w_i = 1/v_i^2$, $\chi^2 = \sum_{i=1}^k w_i(x_i - \hat{\mu})^2$, and $\hat{\mu} = \sum_{i=1}^k w_i x_i \big/ \sum_{i=1}^k w_i$. The proof is along the lines of the proof given in Böhning (1999) where the simpler case $v^2 = \sigma^2$ is reviewed. Equating the expected value (6) to the empirical observed χ^2-value leads to the *moment estimator*

$$\hat{\tau^2} = \frac{\chi^2 - (k-1)}{\sum_{i=1}^k w_i - \sum_{i=1}^k w_i^2 \big/ \sum_{i=1}^k w_i}. \tag{7}$$

Similarly to Section 1, where the simple case of a normal with *known* variances for the conditional study-specific distribution was considered, the more general estimator (7) is *unbiased* by construction as well. This result is unaffected by the distributional properties of the χ^2-statistic. Since (7) is a direct generalization of (3) we still refer to the estimator (7) as the DSL estimator. In practice, it might occur that $\hat{\tau^2} < 0$, in which case the truncated version $\hat{\tau^2_+} = \max\{0, \hat{\tau^2}\}$ is used instead, as before. Note that the result requires *population-averaged* study-specific variances v_i^2 (or, at least, estimates for them). This is fulfilled in the case of the normal distribution, since $v_i^2 = \sigma_i^2$, for all $i = 1, \ldots, k$. However, it should be kept in mind that bias will occur if study-specific variances are estimated with error. This problem vanishes if the study-specific sample sizes are large enough. More importantly, however, it has become a common practice to use also the conditional study-specific variances instead of their population-averaged versions in situations other than the normal. This can lead to considerable bias which will not disappear even with large study sizes. This kind of bias will be demonstrated in the next section. First, we want to exemplify the difference between *conditional study-specific* variances and *population-averaged study-specific* variances.

X binomial with parameters θ and n. Suppose a sample of k proportions $\hat{\theta}_i = \frac{x_i}{n_i}$, $i = 1, \ldots, k$ is available. Then, it has become common practice to use the inverse of $\hat{\theta}_i(1-\hat{\theta}_i)/n_i$ as weights when computing the estimator (7). We have seen in Section 2, however, that $v_i^2 = \int \frac{\theta(1-\theta)}{n_i} P(\mathrm{d}\theta) = \mu(1-\mu)/n_i - \tau^2/n_i$, which is the one to be used in the computation of (7). Unfortunately, this population-averaged variance involves τ^2 itself. Because $\mathrm{Var}(X_i/n_i) = \frac{\mu(1-\mu)}{n_i} + \left(1 - \frac{1}{n_i}\right)\tau^2$ it appears justified to use the approximation $v_i^2 \approx \mu(1-\mu)/n_i$, which can easily be estimated by $\hat{\mu}(1-\hat{\mu})/n_i$ with $\hat{\mu}$ being the pooled estimator $\sum_{i=1}^k w_i \frac{x_i}{n_i} \big/ \sum_{i=1}^k w_i = \sum_{i=1}^k x_i \big/ \sum_{i=1}^k n_i$, where $w_i = \frac{1}{v_i^2} \approx \mu(1-\mu)/n_i$.

Standardized mean difference. Suppose a sample of k studies is available comparing a treatment group and a control group on the basis of the *standardized mean difference*: for example, there are k independent study results $X_i = \hat{\theta}_i = (\overline{x}_{T,i} - \overline{x}_{C,i})/s_i$, for $i = 1, \ldots, k$. Since

$$\mathrm{Var}(\hat{\theta}_i | \theta_i, n_{T,i}, n_{C,i}) \approx \frac{1}{n_{T,i}} + \frac{1}{n_{C,i}} + \frac{\theta_i^2}{2n_i},$$

it has become common practice to use

$$\widehat{\mathrm{Var}}(\hat{\theta}_i | \theta_i, n_{T,i}, n_{C,i}) \approx \frac{1}{n_{T,i}} + \frac{1}{n_{C,i}} + \frac{\hat{\theta}_i^2}{2n_i}$$

when computing the weights in the DSL estimator. However, the population-averaged variance was found in (5) as

$$\int \mathrm{Var}(\hat{\theta}_i | \theta_i, n_{T,i}, n_{C,i}) P(\mathrm{d}\theta_i) \approx \frac{1}{n_{T,i}} + \frac{1}{n_{C,i}} + \frac{\mu^2 + \tau^2}{2n_i}$$

which could be estimated by $\frac{1}{n_{T,i}} + \frac{1}{n_{C,i}} + \frac{\hat{\mu}^2}{2n_i}$ with $\hat{\mu} = \overline{\theta}$.

Table 2. *Bias and SD for DSL estimator using weights*

	$w_i^{-1} = \hat{\mu}(1-\hat{\mu})/n_i$		$w_i^{-1} = \hat{\theta}_i(1-\hat{\theta}_i)/n_i$	
τ^2	Bias[a]	SD	Bias	SD
0.0025	−0.190	0.172	−0.182	0.132
0.0625	0.160	0.809	−3.852	0.652
0.0625	−0.180	0.906	−5.563	0.277
0.1600	−0.350	0.995	−15.224	0.317

[a] All entries for bias and SD are multiplied by 100.

5. AN ILLUSTRATION BY MEANS OF SIMULATION

In this section we will show some of the consequences when these facts are ignored, for example, if estimated conditional study-specific variances are used, instead of population-averaged study-specific variances. We do not want to provide a complete simulation study at this stage, but instead provide some illustrative examples. The DSL estimator has been recommended for various reasons (Biggerstaff and Tweedie, 1997; NRC Committe on Applied and Theoretical Statistics, 1992) including its *unbiasedness*. Therefore, we will concentrate the evaluation of the simulation study on the $bias = \hat{\tau^2} - \tau^2$. We will first consider the binomial distribution.

Binomial distribution Let X_i be binomially distributed with parameter θ_i and $n_i = 20$, and $i = 1, \ldots, 40$; in other words, it is assumed that $k = 40$ studies are available. It is assumed further in this example that θ_i follows a discrete distribution with two masspoints θ_1 and θ_2 receiving equal weights $1/2$. Four populations are studied:

- $\theta_1 = 0.1, \theta_2 = 0.2$ with associated $\tau^2 = 0.0025$
- $\theta_1 = 0.25, \theta_2 = 0.75$ with associated $\tau^2 = 0.0625$
- $\theta_1 = 0.1, \theta_2 = 0.6$ with associated $\tau^2 = 0.0625$ and
- $\theta_1 = 0.1, \theta_2 = 0.9$ with associated $\tau^2 = 0.16$.

The results are provided in Table 2. (All simulation results refer to a replication size of 1000.) The bias for estimated conditional study-specific weights is clearly visible in column 4 of Table 2.

Standardized mean difference. Finally, we consider again the standardized mean difference which has already been discussed in Sections 2 and 3. A design of sample sizes was used as given in Table 3 (which was taken from a published meta-analysis on standardized mean differences). It was further assumed that the mean of the heterogeneity distribution was fixed to be $\mu = 0.5$, whereas τ^2 varied between 1 and 10. On the one hand, we considered the conditional study-specific variance estimate of the standardized mean difference $\frac{1}{n_{T,i}} + \frac{1}{n_{C,i}} + \frac{\hat{\theta}_i^2}{2n_i}$, commonly used when computing the weights in the DSL estimator. On the other hand the population-averaged study-specific variance was found in (2) as $\frac{1}{n_{T,i}} + \frac{1}{n_{C,i}} + \frac{\mu^2+\tau^2}{2n_i}$. This could be estimated by $\frac{1}{n_{T,i}} + \frac{1}{n_{C,i}} + \frac{\hat{\mu}^2}{2n_i}$ with $\hat{\mu} = \overline{\theta}$, which was used alternatively, when computing the weight in the DSL estimator. The results are shown in Figure 3, which clearly indicates the considerable amount of bias occurring when using estimated conditional study-specific weights.

In both cases, it can be seen that there appears considerable bias when the study-specific variances are replaced by their estimates.

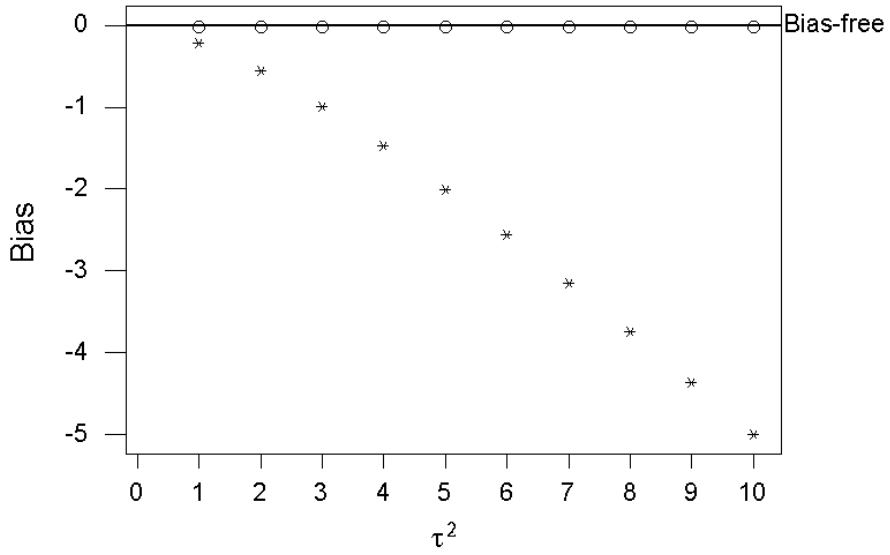

Fig. 3. Comparison of bias for DSL estimator with estimated population-averaged weights (circle) with bias for DSL estimator with estimated conditional study specific weights (asterisk) in the case of the standardized mean difference.

Table 3. *Layout of sample sizes for* $k = 15$ *studies*

Study number	$n_{T,i}$	$n_{C,i}$	Study number	$n_{T,i}$	$n_{C,i}$
1	7	13	2	18	32
3	35	13	4	8	16
5	19	10	6	13	7
7	8	15	8	13	7
9	4	9	10	15	8
11	31	20	12	24	37
13	27	39	14	11	29
15	32	18			

6. EXAMPLE FROM META-ANALYSIS (CONTINUED)

We have seen in Section 3 that $\hat{\mu} = \sum_{i=1}^{k} x_i \big/ \sum_{i=1}^{k} n_i = 0.7546$. It might be of interest and importance to have the variance of $\hat{\mu}$ available. According to Section 3, we have that $\mathrm{Var}(X_i/n_i) = \frac{\mu(1-\mu)}{n_i} + \left(1 - \frac{1}{n_i}\right)\tau^2$ or, $\mathrm{Var}(X_i) = n_i\mu(1-\mu) + n_i(n_i-1)\tau^2$. This leads to

$$\mathrm{Var}\left(\frac{\sum_{i=1}^{k} X_i}{\sum_{i=1}^{k} n_i}\right) = \frac{1}{\sum_{i=1}^{k} n_i}\mu(1-\mu) + \frac{\sum_{i=1}^{k} n_i(n_i-1)}{\left[\sum_{i=1}^{k} n_i\right]^2}\tau^2. \tag{8}$$

Note that conventional textbook formulas for the variance of $\hat{\mu}$ use only the first term in (8), e.g. $\mu(1-\mu)/N$ with $N = n_1 + n_2 + \cdots + n_k$. Replacing μ and τ^2 in (8) by their estimates leads to

$$\widehat{\mathrm{Var}}\left(\frac{\sum_{i=1}^{k} x_i}{\sum_{i=1}^{k} n_i}\right) = \frac{1}{\sum_{i=1}^{k} n_i}\hat{\mu}(1-\hat{\mu}) + \frac{\sum_{i=1}^{k} n_i(n_i-1)}{\left[\sum_{i=1}^{k} n_i\right]^2}\hat{\tau}^2 \tag{9}$$

where $\hat{\tau^2}$ is the DSL estimator, with value $\hat{\tau^2} = 0.015\,79$ in our case. Note that $\hat{\tau^2}$ has been computed on the basis of $\hat{v_i^2} = \hat{\mu}(1-\hat{\mu})/n_i$, for $i = 1, \ldots, k$. We find a value of $\widehat{\mathrm{Var}}\big(\sum_{i=1}^k x_i \big/ \sum_{i=1}^k n_i\big) = 0.002\,459$ which is at least four times higher than the one based on the conventional variance formula $\hat{\mu}(1-\hat{\mu})/N = 0.000\,57$. This leads to 95% confidence intervals based on (9) as (0.657 40, 0.851 80) and (0.707 89, 0.801 32) using the conventional variance. It might be also interesting to look at the variance using the *wrong* χ^2, e.g. the one using the weights $\hat{w_i}^{-1} = \frac{x_i}{n_i}\big(1-\frac{x_i}{n_i}\big)\big/n_i$, and thus the *wrong* $\hat{\tau^2}$. If χ^2 is based on the estimated weights $\hat{\theta_i}(1-\hat{\theta_i})/n_i$ we find that $\hat{\tau^2} = 0.020\,14$, considerably larger than the result based upon the unbiased estimate. Using this value of $\hat{\tau^2}$ in (9) we are led to a confidence interval of (0.647 58, 0.861 63) for μ.

7. DISCUSSION

Generality of the approach. We have seen that the DSL estimator in its general form requires inverse population-averaged study-specific variances as weights, in which case the estimator is unbiased. It has become common practice, however, to use estimates of the study-specific variances instead of their population-averaged versions. This can lead to considerable bias. It has been also demonstrated for several examples such as the example of a normally distributed measure of interest, a binomial proportion, a Poisson distributed SMR and the effect measure of the standardized difference, how to construct the population-averaged study-specific variances which lead to the unbiased version of the DSL estimator. Indeed, for practical applications this is the most important issue. It is often possible to construct estimates of the population-averaged study-specific variances, and thus avoid the occurrence of severe bias. However, there are also several important examples for which it is more difficult to do so, such as the situation of the effect measure of relative risk or risk difference. Here, the difficulty lies in the fact that the exact, conditional study-specific distribution is not so easy to determine. To be more precise let us consider as the parameter of interest the risk difference $\theta = p_T - p_C$ where $p_T(p_C)$ is the event risk in the treatment (control) group. Similarly, let $n_T(n_C)$ be the number under risk in the treatment (control) group. Then, the conditional variance is given as $\mathrm{Var}(\hat{\theta}|n_T, n_C, p_T, p_C) = \frac{p_T(1-p_T)}{n_T} + \frac{p_C(1-p_C)}{n_C} = \mathrm{Var}(\hat{\theta}|n_T, n_C, \theta, p_C) = \frac{(\theta+p_C)(1-\theta-p_C)}{n_T} + \frac{p_C(1-p_C)}{n_C}$ which clearly shows that the conditional variance is not only a function of θ but also of a further nuisance parameter. This is rather typical for two-sample problems such as risk difference, relative risk, or odds ratio (the standardized difference is exceptional in this respect). These areas will require future research.

Interpretation of meta-analysis under heterogeneity. We have mentioned and demonstrated previously (for example, see Section 6) that the random effects approach by DSL will lead to the most different result from the fixed-effect model if there is strong heterogeneity. If there is small heterogeneity both approaches will provide similar results. This should be kept in mind, especially in the following context. Sometimes critical voices of meta-analysis raise the point that the computation of a summary measure is inappropriate in the presence of heterogeneity. The core of the objection argues that if there is heterogeneity there is no common ground for computing a summary measure and it is unclear what this summary measure is estimating. We think, however, that such a summary statistic may be computed though its interpretation is different. In a homogeneous population the summary measure will provide an estimate for all parts of the population, in the same way that in a meta-analysis it would provide an estimate of the measure of interest for all possible studies. In a heterogenous population this summary statistic will also provide an estimate for the overall mean in the population, knowing that this measure of interest will have different mean values in certain parts of the population. More precisely, when we reconsider the two-level model, the summary statistic will estimate the mean of the heterogeneity distribution P. In other words, if we

have a sample of studies the overall measure will provide an estimate of the mean of the study means though there might be variation between study means.

Truncation of DSL estimator. In the simulation study the estimator $\hat{\tau}^2$ was used and *not* $\hat{\tau}^2_+$. The reason was simply that we had always $\hat{\tau}^2 \geqslant 0$ which is typically the case if one simulates under heterogeneity, and consequently both estimators agree. Recall that the denominator of (7) can be written as $\{\sum_{i=1}^k w_i\}\{1-\sum_{i=1}^k w_i^2/(\sum_{i=1}^k w_i)^2\}$ which shows that the denominator is always positive if all w_i are positive. Therefore, $\hat{\tau}^2$ can only be negative if χ^2 is smaller than $k-1$, which typically occurs under homogeneity. In consequence, simulation studies under *homogeneity* will show distributional deviations between $\hat{\tau}^2$ and $\hat{\tau}^2_+$, though this is less the case under heterogeneity and not at all in our simulation study. It was speculated that the observed bias in the DSL estimator stems from using the *truncated* version (Hartung, 1999). Though this idea is intuitively reasonable, it is quite clear from the above that the genesis of the observed bias is from a different source. In fact, the analysis has shown that it is not the question of using a different *DSL estimator*, but rather of using it in a way which is appropriate for a two-level model.

Alternative estimators. Occasionly, different estimators are proposed as alternatives to the DSL estimator. Hardy and Thompson (1996) suggest a marginal maximum likelihood estimator. In this approach $f(x_j, \theta, \sigma_j^2)$ is assumed to be the normal density (for available measures x_j, $j = 1, \ldots, k$), and P is assumed to be normal as well with parameters μ and τ^2. Then, the marginal distribution is normal as well with parameters μ and $\sigma_j^2 + \tau^2$ and the maximum likleihood estimator $\widehat{\tau^2_{ML}}$ of τ^2 is determined by iteratively maximizing the marginal likelihood function. For details see Brockwell and Gordon (2001). The question arises to that extent the problems descibed here are also inherent in the Hardy and Thompson (1996) approach. This is certainly the case for the first issue mentioned in Section 1, namely replacing known variances by their estimates, since

> it still assumes that the individual study variances are known, when in practice they too must be estimated. (Hardy and Thompson, 1996, p. 627)

Hardy and Thompson argue, however,

> Except when all the trials are small, the additional uncertainty would not therefore be expected to have a great impact on the results ...

With respect to the second issue, the problem of using the right variances is inherently taken care of when using the marginal (population averaged) likelihood. Here, however, another problem arises. It might be doubted that for measures such as the binomial proportion or the standardized mortality ratio the normal is an appropiate conditional distribution. In addition, the assumption of normality for the heterogeneity distribution is critical. Hardy and Thompson (1998) suggest a number of techniques including normal plots and departure tests such as Anderson–Darling and Shapiro–Francia to check the distributional assumption of the normal–normal model. Brockwell and Gordon (2001, p. 829) comment

> The assumption of normality poses problems, first in its validity, and secondly in our ability to check that validity for meta-analyses based on a small number of studies. In particular, the assumption of normally distributed random effects ... is not easily verified or justified.

When using a parametric marginal model the crucial part consists in choosing the right parametric forms. Frequently, for binomial proportions the beta-binomial is used (Collett, 1999), whereas for the standardized mortality ratio the Gamma-Poisson is employed. These models seem to be more appropriate

for the mentioned outcome measures than the normal–normal model. However, it should be pointed out that their preferred choice might be influenced by the fact that they form conjugate families and are thus particularly suitable for Bayesian procedures. In addition, as Brockwell and Gordon (2001) have commented, it is difficult (if not impossible) to check whether the associated random effects distribution follows the specific distribution of a beta or gamma. In conclusion, nonparametric procedures, to which the DSL estimator belongs, retain their importance.

ACKNOWLEDGEMENTS

This research has been conducted with support from the German Research Foundation. In addition, the work of D. B. and A. B. is supported within the Vigoni programme of the German–Italian science co-operation. The authors would like to express their sincerest thanks to the editor Peter Diggle as well as to two reviewers for their thoughtful, helpful and constructive comments.

REFERENCES

BIGGERSTAFF, B. J. AND TWEEDIE, R. L. (1997). Incorporating variability in estimates of heterogneity in the random effects model in meta-analysis. *Statistics in Medicine* **16**, 753–768.

BÖHNING, D. (1999). *Computer-assisted Analysis of Mixtures and Applications: Meta-analysis, Disease Mapping, and Others*. London: Chapman and Hall.

BÖHNING, D. AND SAROL, J. (2000). Estimating risk difference in multicenter studies under baseline heterogeneity. *Biometrics* **56**, 304–308.

BROCKWELL, S. E. AND GORDON, I. R. (2001). A comparison of statistical methods for meta-analysis. *Statistics in Medicine* **20**, 825–840.

COLLETT, D. (1999). *Modelling Binary Data*. London: Chapman and Hall.

COOPER, H. AND HEDGES, L. (1994). *The Handbook of Research Synthesis.* New York: Russell Sage Foundation.

DERSIMONIAN, R. AND LAIRD, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials* **7**, 177–188.

HARDY, R. J. AND THOMPSON, S. G. (1996). A likelihood approach to meta-analysis with random effects. *Statistics in Medicine* **15**, 619–629.

HARDY, R. J. AND THOMPSON, S. G. (1998). Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine* **17**, 841–856.

HARTUNG, J. (1999). Personel communication.

HEDGES, L. V. AND OLKIN, I. (1985). *Statistical Methods for Meta-analysis*. New York: Academic.

HOEM, J. M. (1987). Statistical analysis of a multiplicative model and its application to the standardization of vital rates: a review. *International Statistical Review* **55**, 119–152.

KEIDING, N. (1987). The method of expected number of deaths, 1786-1886-1986. *International Statistical Review* **55**, 1–20.

KENNWARD, M. G. AND ROGER, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* **53**, 983–997.

KLEINBAUM, D. G., KUPPER, L. L. AND MORGENSTERN, H. (1982). *Epidemiologic Research: Principles and Quantitative Methods*. Belmont, CA: Lifetime Learning.

NORMAND, S.-L. T. (1999). Tutorial in biostatistics: meta-analysis: formulating, evaluating, combing, and reporting. *Statistics in Medicine* **18**, 321–359.

NRC COMMITTEE ON APPLIED AND THEORETICAL STATISTICS (1992). *Combining Information: Statistical Issues and Opportunities for Research*. Washington, DC: National Academy Press.

PETITTI, D. B. (1994). *Meta-analysis, Decision Analysis and Cost—Effectiveness Analysis. Methods for Quantitative Synthesis in Medicine*. Oxford: Oxford University Press.

POTTHOFF, R. E. AND WHITTINGHILL, M. (1966). Testing for homogeneity I: the binomial and multinomial distributions. *Biometrika* **53**, 167–182.

PRITZ, M. B., ZHOU, X. H. AND BRIZENDINE, E. J. (1996). Hyperdynamic therapy for cerebral vasospasm: a meta-analysis of 14 studies. *Journal of Neurovascular Disease* **1**, 6–8.

SENN, S. (2000). Letter to the editor. *Controlled Clinical Trials* **21**, 589–592.

SHADISH, W. R. AND HADDOCK, C. K. (1994). Combining estimates of effect size. In Cooper, H. and Hedges, L. (eds), *The Handbook of Research Synthesis*, New York: Russell Sage Foundation, pp. 261–281.

ZHOU, X.-H., BRIZENDINE, E. J. AND PRITZ, M. B. (1999). Methods for combining rates from several studies. *Statistics in Medicine* **18**, 557–566.

[*Received December 4, 2000; revised June 11, 2001; accepted for publication July 17, 2001*]