Published in final edited form as:
Stat Med. 2010 May 30; 29(12): 1259–1265. doi:10.1002/sim.3607.

Empirical vs. natural weighting in random effects meta-analysis

Jonathan J Shuster, Ph.D.[1,*]
[1]Department of Epidemiology and Health Policy Research, College of Medicine, University of Florida, Gainesville, FL 32610

SUMMARY

This paper brings into serious question the validity of empirically based weighting in random effects Meta-Analysis. These methods treat sample sizes as non-random, whereas they need to be part of the random effects analysis. It will be demonstrated that empirical weighting risks substantial bias. Two alternate methods are proposed. The first estimates the arithmetic mean of the population of study effect sizes per the classical model for random effects meta-analysis. We show that anything other than an unweighted mean of study effect sizes will risk serious bias for this targeted parameter. The second method estimates a patient level effect size, something quite different from the first. To prevent inconsistent estimation for this population parameter, the study effect sizes must be weighted in proportion to their total sample sizes for the trial. The two approaches will be presented for a meta-analysis of a nasal decongestant, while at the same time will produce counter-intuitive results for the DerSimonian-Laird approach, the most popular empirically based weighted method. It is concluded that all past publications based on empirically weighted random effects meta-analysis should be revisited to see if the qualitative conclusions hold up under the methods proposed herein. It is also recommended that empirically based weighted random effects meta-analysis not be used in the future, unless strong cautions about the assumptions underlying these analyses are stated, and at a minimum, some form of secondary analysis based on the principles set forth in this article be provided to supplement the primary analysis.

Keywords

bias; DerSimonian-Laird; meta-analysis; random effects; weight

1. INTRODUCTION

Given that the DerSimonian-Laird (DSL) [1] method for random effects meta-analysis has been widely used for well over 20 years, and is often taught in classes in Epidemiology and Biostatistics, it does not seem possible that in general, empirically based methods for weighting effect sizes (including DSL), per the classical random effects model for meta-analysis, could be flawed. As biostatisticians, we are aware that the optimal linear combination of unbiased estimates is to weight them inversely proportional to the square of their standard errors. That is what the DSL method tries to emulate-on the surface a worthy objective. The major issue that these methods fail to address is that the weights themselves are volatile random variables. For example, when we “draw a study out of the hat”, we obtain not only random effect sizes, but also random study sizes, ergo random precision within studies. The ultimate weights may be correlated with these effect sizes, thereby

[*]Correspondence to: Jonathan J. Shuster, PO Box 100177, Gainesville, FL 32610-0177. jshuster@biostat.ufl.edu.

producing serious bias. This is something we shall witness and see how this can ruin the DSL approach for its intended model.

In Section 2, we shall advocate an unweighted approach for estimating the arithmetic mean of study effect sizes. The classical random effects model has as its primary parameter the "mean of means". There is nothing in the model about the means that suggest one is trying to estimate a weighted combination of these means. We shall see that if one uses any empirically based weighting system, bias will occur unless the empirical normalized weights and individual study effect size estimates are uncorrelated. To expect this in advance seems to be a leap of faith. We shall also show that the unweighted estimate can be viewed as a bias corrected weighted estimate, irrespective of the weighting system selected! In Section 3, we shall provide a method to estimate a patient level mean effect size. This will weight study effect sizes in proportion to their total sample size. This targeted population parameter is different from the one addressed in Section 2. Section 4 will deal with a real dataset from Kollar and colleagues [2]. By tweaking the data for the DSL method, we shall see a highly counterintuitive result where increasing the standard errors by a common multiple of 3.0 will switch the DSL result from a non-significant result to a significant result! In ANOVA, such a tweak would decrease the value of the F-statistic by a factor of 9.0, assuring that a much higher p-value would occur. Section 5 is devoted to a discussion of the implications of this article in practical terms. In addition, we shall rebut a letter of Wacksman and Kollar [3], where they dispute conclusions of Shuster, Jones, and Salmon [4], on whether the data in [2] represent efficacy or not.

2. CLASSICAL STUDY POPULATION MODEL: ESTIMATION OF MEAN STUDY EFFECT SIZE

The classical random effects model, such as that posed in DerSimonian and Laird [1], presumes that studies form a random sample from a very large population of potential studies, and that within studies, we have unbiased estimates of the study-specific effect sizes. This produces the classical model:

$$Y_j=\mu+\theta_j+\varepsilon_j \quad (1)$$

where (a) θ_j is the difference between the true effect size for study j and the overall true target population effect size μ, and satisfies $E(\theta_j)=0$ and (b) ε_j is the random error associated within study j, and satisfies $E(\varepsilon_j)=0$. Note that j identifies the study, j=1,2,…,M.

The weighted estimate is given by

$$\mu^*=\Sigma W_j Y_j,$$

where

$$\Sigma W_j=1,$$

with the study weights W_j, random variables, derived from the data including sample sizes. (Note that the DSL paper uses a bit different notation, but it can be reduced to this formulation by taking their weights, which do not sum to 1.0, and dividing each by their sum of weights.)

Without loss of generality, from the model, the Y_j can be considered independent identically distributed random variables, but the W_j, are correlated because they sum to unity. If you

have a problem with the independent, identical distribution of the Y's, note that we can randomly pick labels for j in the study pairs (W_j, Y_j) j=1,2,…,M, and by doing that, we will not affect μ^*. Next, we note

$$E(W_jY_j)=Cov(W_j, Y_j)+E(W_j)E(Y_j)=Cov(W_j, Y_j)+\mu/M \quad (2)$$

Equation (2) is a direct consequence of (1) and the fact that the exchangeable weights satisfy $\Sigma W_j=1$, making $E(Y_j)=\mu$ and $E(W_j)=1/M$.

Now this in turn implies that

$$E(\mu^*)=E(\Sigma W_jY_j)=MCov(W_j, Y_j)+\mu \quad (3)$$

Conclusions about empirically based weights

Equation (3) tells one that if the estimated effect sizes and the empirically derived study weights are correlated, the summary estimate of effect size derived from the random effects analysis (e.g. DerSimonian and Laird [1]) is biased. At a very minimum, to use empirically based weights, one must clearly state the assumption that the weights are not correlated with the individual study effect sizes. Alternatively, as suggested by a colleague, Dr. Keith Muller, a bias corrected estimate μ^c for μ might be obtained from equation (3) by subtracting M times the sample covariance of W_j and Y_j from the quantity μ^*. Note that in (4) below, we use M as the denominator in the sample covariance, not the more common (M-1).

$$\mu^c=\Sigma W_jY_j - M[\{\Sigma W_jY_j/M\} - \{(\Sigma W_j/M)(\Sigma Y_j/M)\}]=\Sigma Y_j/M \quad (4)$$

Hence, the unweighted estimate can be viewed as a bias corrected weighted estimate, however the weights are formulated. We believe this argument seals any possible controversy in favor of the use of unweighted estimation for this classical model.

Note that if the weights are non-random (not derived from the sample size information or any other part of the data), then indeed μ^* is unbiased. Of course, the optimal fixed (non-random) weights to minimize the variance would be $W_j=1/M$, ergo, an unweighted analysis.

3. A PATIENT-BASED POPULATION MODEL

An attractive alternative to the classical model (1) is the following based on individual patient expectation, rather than based upon study expectation. In this patient population model, every past and future subject belongs to a conceptual or real clinical trial, and as in (1) we assume that a random sample of clinical trials is observed. The parameter that may be of interest is the mean effect size for the patients, namely the expected patient difference between the outcomes for the treatments. This is represented as

$$\nu=E(N_jY_j)/E(N_j) \quad (5)$$

where Y_j is per (1) and N_j is the total number of patients in the j-th trial sampled.

It needs to be stressed that ν and μ are completely different parameters, and the primary parameter of interest should be identified at the design point of the study.

We shall estimate ν by the ratio estimate

$$\nu^* = \Sigma(N_j Y_j)/\Sigma(N_j) \quad (6)$$

For a large sample of M studies, based on the "Delta Method" per Serfling [5], ν^* has the following asymptotic normal distribution (AN):

$$\nu^* \sim AN(\nu, \sigma^2/M) \quad (7)$$

where the asymptotic variance is given by

$$\sigma^2/M = [\{Var(U_j)/E^2(N_j)\} + \{Var(N_j)E^2(U_j)/E^4(N_j)\} - 2\{Cov(U_j, N_j)E(U_j)/E^3(N_j)\}]/M \quad (8)$$

where $U_j = N_j Y_j$

A consistent estimator σ^{*2} for σ^2 can be obtained using the sample variances and covariances in place of the actual variances and covariances for U_j and N_j, respectively.

Note that if one estimated ν but replaced the weights in (6) by any weighting system other than proportional to the total sample size for the trial, there is no reason to believe that the resulting estimate would be consistent for ν. Nonparametrically, the method in (6) is valid when the number of studies is large, and its validity does not depend upon balanced randomization between the treatment arms in parallel studies.

4. ANOTHER LOOK AT THE KOLLAR META-ANALYSIS [2]

In order to illustrate the various methods, we shall line up four methods side by side. The Kollar et. al. meta analysis [2] contrasted the efficacy in terms of nasal airway resistance (NAR) in seven small randomized crossover studies, with respect to a single dose of phenylephrine 10 mg compared with placebo to treat nasal congestion. While the number of studies might not meet the requirements for asymptotic approximations, these studies do offer an excellent platform to assess the methods. As in Kollar and colleagues [2], we shall look at each of eight time points separately. Table I provides the summary statistics for the seven studies and eight time points. As a surrogate for the standard error, we used 25% of the length of the confidence interval for effect size as given in Table III of Kollar and colleagues [2]..

Table II provides the estimated effect size, standard errors, and in the random effect analyses, two-sided P-values for testing the effect size is zero. P-values are calculated from the t-distribution with degrees of freedom equal to one less than the number of studies involved. This is a conservative adjustment to the asymptotically equivalent normal distribution cut points to account for the small number of studies. Fixed effects analysis use weights inversely proportional to the standard error. The random effects analyses are (1) unweighted per section 2, (2) DerSimonian and Laird [1], and (3) weighted by sample size per Section 3. The DerSimonian-Laird approach applies to the same model as the unweighted, and is therefore not seen as a competitor to the method that weights by sample size.

The DSL method tries to estimate the effect size by weighting the Y_j in (1) by the inverse of its variance: ideally

$$V_j = \sigma^2(Y_j) = Var(\theta_j) + Var(\varepsilon_j) \quad (9)$$

Of course $Var(\varepsilon_j)$, the within study variance of the estimator is estimated by the square of the standard error for the j-th study, SE_j^2 The DSL method estimates the between study variation $Var(\theta_j)$ by Δ^2, (See [1] for details). This quantity depends heavily upon the diversity of the studies, as measured by the Cochran Q chi-square statistic.

The DSL weights are therefore defined by

$$Wj=Uj/\Sigma Uj \quad (10)$$

$$\text{where} Uj=(\Delta^2+SE_j^2)^{-1} \quad (11)$$

When the diversity is “small”, with the chi-square statistic below the degrees of freedom, the DSL and the fixed effects analysis will yield identical estimates. When the diversity leads to chi-square tending to infinity, the DSL will correspond closely to the unweighted analysis. The problem is that the Cochran Q chi-square statistic depends heavily on the sample sizes in the population of trials. Hence if we had two populations of trials, with study per study identical study target population means (θ_j) in (1), but every study in population B had ten times the sample size of the one corresponding to population A, the diversity of population B will be perceived as far greater than population A, and hence the DSL estimate will tend to be closer to the unweighted mean in population B and closer to the fixed effects mean in population A. If the target parameters of the fixed and unweighted scenarios are systematically different, the DSL estimates in populations A and B will be consistent for different numbers, when they should logically be estimating the same number. Neither of the estimates proposed in Sections 2 and 3 have this issue. We shall illustrate this in detail for the 45 minute data in Tables I and II, while at the same time demonstrating a highly counterintuitive property of the DSL estimate.

Table III gives the normalized weights for the DSL estimates by study for the actual standard error, and for standard errors tweaked to be exactly three times the originals. Intuitively, one would certainly expect that the p-value would by higher under the tweaked model. If this would happen in a one-way ANOVA setting, the F-statistic would be reduced by a factor of 9.0, making the p-value much higher than in the original setting. Yet the DSL point estimate moves radically from being relatively close to the equal weight estimate (DSL =−1.33, equal weight =−1.13) for the actual standard errors to being relatively close to the fixed effect estimate (DSL =−2.87, fixed effect =−3.54) for the tweaked triple standard errors. Curiously, this tweak converted a non-significant result to a significant result.

To illustrate numerically that the unweighted estimate can be viewed as a bias corrected weighted estimate, consider the tweaked weights in Table III against the Point estimates at 45 minutes from Table I. Note that the correlation coefficient between these weights and point estimates is high at −0.87, leading to a covariance of −0.2484. The bias correction of M times the covariance (M=7 studies) is 7(−0.2484)=−1.74 and hence μ^c = −2.87−{7*(−0.2484)}=−1.13, the unweighted estimate. Note that the bias correction is major.

Note that while the author wrote his own software for the computations involved in this paper, with the help of Dr. Alexander Wagenaar, the DSL method was cross-checked via the Comprehensive Meta-Analysis package. See http://www.meta-analysis.com/. Its default random effects method is DSL.

5. DISCUSSION

The major conclusion that can be reached from this article is that when one uses empirically based weighting to conduct a random effects meta-analysis, it is dangerous to ignore the sampling errors involved in the weights. We argue that all such past articles need to be reexamined to determine if their conclusions hold up qualitatively. Although some of these articles may contain a fixed effects analysis, with or without diagnostics for homogeneity, we reject these as viable alternatives to new random effects analyses in terms of their conclusions. It is true that the fixed effects supplementary analyses validly test the null hypothesis that all effect sizes are zero (ergo a fixed effect). But this hypothesis is far too narrow for practical application. Random effects proponents allow for null situations where the effect sizes have a non-trivial probability distribution, with both positive and negative effect sizes. Diagnostic testing for homogeneity of effect sizes should be deemed as completely irrelevant to this question. First, these tests have very low power. As statisticians, it is improper to accept a null hypothesis as true, unless a very tight confidence set about the null hypothesized value is obtained. This will not occur in meta-analysis. Second, the error properties must be assessed from before the point where the diagnostic test creates two analytic branches. The type I error is only conditional on the assumption that the correct path was taken, and therefore risks being incorrectly assessed.

Another issue of fixed effects, typically based on weights inversely proportional to the square of the standard error, is that often, the meta-analysis is dominated by a small number of studies. For example, for the data in Table II, Studies 1 and 2 consume 84%, 79%, 78%, 83%, 86%, 78%, 92%, and 95% of the fixed effects weights at times 15,30,45,60,90,120,180, and 240 minutes, respectively. Yet these studies have only 16/113 (10/ 81) of the patients at times <180(180+) minutes.

There are many good reasons to expect that the effect size and weights will be correlated. For example, in drug development, early smaller studies may be pure (drug only vs. placebo), while later larger studies may use the drug vs. placebo in an adjuvant setting, with larger studies expecting a smaller difference in efficacy. For side effects, however, an adjuvant therapy interaction with the experimental drug may trigger a larger differential, yielding the opposite correlation. Better designed studies may lower the sampling error, thereby increasing the weight, while the greater skills of these investigators over those contributing to less well run studies, may lead to larger advantages for the experimental therapy over the controls. This is especially problematic in surgery device trials. Unknown to the meta-analyst, some studies may have been terminated early for efficacy, yielding smaller weights than those that run to completion. These arguments suggest that these correlations can be expected, and it is statistically risky to assume they do not exist.

Getting over the use of empirically weighted methods

Assuming the standard model for random effects meta-analysis, we provide three related arguments that should help practitioners to select one of the methods proposed in Sections 2 or 3 over empirically based weighted methods in random effects meta-analysis. It is true that weighted methods can do a better job at reducing within study variability. But this is at a cost of estimating the wrong quantity, according to standard models.

1. Imagine that you were about to publish a random effects meta-analysis, when you discover that the data you had for one study (the smallest in your collection) was in fact based on an interim analysis, and that there were far more patients, making it the largest of the studies. You will reanalyze the data, and the weights will change considerably. This will change the target population mean value of your estimator. The expected value of the unweighted estimator would not change.

2. Think of the parameter we are trying to estimate in a physical sense. Imagine that we are dealing with a collection of parallel two-treatment randomized drug vs. placebo studies. The parameter μ represents the expected outcome in the following experiment. Draw a study at random from the target population of studies. Treat one patient on drug and one patient on placebo. What is the expected effect size (treatment-placebo)? Note that all studies are equal partners. When we select the study, the standard random effects model does not weight the studies to bring this to a patient level. So neither should the analysis. In fact, it can be argued that sample sizes are just accidents of fate in the overall context of the meta-analysis. The parameter ν in Section 3 represents the average effect size at a patient level. It also has a physical interpretation. For example, in parallel two treatment randomized studies, it estimates the patient level difference in expectation if you treated all patients on Treatment A vs. treating all patients on Treatment B.

3. Equal weighting avoids the “Bill Gates/Warren Buffet effect” that was seen in Kollar et. al. [2] and Nissan-Wolski [6], where a very few studies play a dominant role. This seems to be contrary to the real spirit of a meta-analysis. Just as in our electoral system, every “person” has one vote, irrespective of their financial weight.

In Waksman and Kollar [3], the claim is made that the data from Table I demonstrate efficacy for the decongestant. They argue that a random effects analysis showed significance at three of the eight time points, and if they dropped the most influential study in terms of weight, study #2, they still saw a significant difference at five time points. Dropping Study #2 in fact makes Study #1 the dominant one. But the issue of empirical weighting renders these analyses as yielding highly biased estimates of effect size, and should be discounted. As can be seen in Table II, the point estimates for fixed effects and unweighted analyses differ markedly at several time points, a red flag for correlation between the weights and the estimated effect sizes, something Waksman and Kollar discount in [3]. Based on our rigorous analysis at eight time points, and two endpoints, we did not see a single P-value below 0.05. The 30 minute time point was close, however. We conclude that these crossover studies do not provide sufficient evidence to conclude that phenylephrine 10 mg has efficacy with respect to nasal airway resistance.

It needs to be pointed out that the methods of this paper should not be used in rare event binomial outcomes where relative risk is the endpoint. This is because of the likelihood of zero events. See Shuster and Colleagues [7], who provide a parallel development for an unweighted random effects meta-analysis to cover that situation.

Acknowledgments

This work was partially supported by grant M01RR00082 from the National Institute of Research Resources, National Institutes of Health.

The author also wishes to thank Drs. Keith Muller, Alexander Wagenaar, Almut Winterstein, and Leslie Hendeles of the University of Florida for their helpful discussion.

REFERENCES

1. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986; 7(3):177–188. [PubMed: 3802833]
2. Kollar C, Schneider H, Waksman J, Krusinska E. Meta-analysis of the efficacy of a single dose of phenylephrine 10 mg compared with placebo in adults with acute nasal congestion due to the common cold. Clinical Therapeutics. 2007; 29(6):1057–1070. [PubMed: 17692721]
3. Waksman J, Kollar C. Comments on ‘Rebuttal to Carpenter *et al*. Comments on “Fixed vs. random effects meta-analysis in rare event studies: The rosiglitazone link with myocardial infarction and

cardiac death” ’ By J.J. Shuster, L.S. Jones, and D.A. Salmon. Statistics in Medicine. 2009; 28(3): 534–536. [PubMed: 19125393]

4. Shuster JJ, Jones LS, Salmon DA. Rebuttal to Carpenter *et al*. Comments on ‘Fixed vs. random effects meta-analysis in rare event studies: the rosiglitazone link with myocardial infarction and cardiac death’. Statistics in Medicine. 2008; 27:3912–3914.
5. Serfling, RJ. Approximation theorems in mathematical statistics. New York: John Wiley Publication; 1980. p. 118-125.
6. Nissen SE, Wolski K. Effect of Rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. N. Engl. J. Med. 2007 Jun 14; 356(24):2457–71. [PubMed: 17517853]
7. Shuster JJ, Jones LS, Salmon DA. Fixed vs. random effects meta-analysis in rare event studies: the rosiglitazone link with myocardial infarction and cardiac death. Statistics in Medicine. 2007; 26:4375–4385. [PubMed: 17768699]

Table I

Point estimate (SE) per Kollar et. al [2]

SE defined as 25% of length of Confidence Interval

Time	Study 1 (N=16)	Study 2 (N=10)	Study 3 (N=16)	Study 4 (N=15)	Study 5 (N=15)	Study 6 (N=16)	Study 7 (N=25)
15 Min	−1.26(0.3050)	−0.05(0.1975)	−0.17(0.7650)	−0.13(0.8625)	−0.58(0.6750)	−0.57(1.1250)	0.99(.9825)
30 Min	−3.11(0.4275)	−1.68(.3250)	−2.24(1.0600)	0.31(0.8550)	−0.21(1.1175)	−0.06(1.6150)	−0.36(1.6250)
45 Min	−5.74(0.4325)	−3.51(0.4325)	−1.90(1.3150)	0.13(1.3025)	−0.07(1.1925)	1.11(1.1625)	2.09(1.4825)
60 Min	−5.44(0.5975)	−3.82(0.4075)	−3.14(1.9375)	−1.81(1.5475)	−0.13(1.3075)	1.53(1.9500)	1.44(2.1275)
90 Min	−4.70(0.6625)	−2.90(0.3750)	−4.75(2.0775)	0.39(1.6550)	0.15(1.5400)	0.17(1.8950)	−0.18(1.9075)
120 Min	−3.44(0.7375)	−2.09(0.3550)	−4.88(1.9625)	1.05(2.1325)	0.93(1.5600)	2.70(2.5725)	2.89(1.7925)
180 Min	N/A	−1.17(0.2700)	−6.81(2.1425)	0.63(2.6225)	N/A	0.83(2.5400)	1.49(1.2675)
240 Min	N/A	−0.38(0.3375)	−6.66(2.8600)	0.68(3.2175)	N/A	−1.65(3.7850)	1.61(2.2125)

Stat Med. Author manuscript; available in PMC 2013 July 01.

Table II

Contrast of Estimates. Entries are Point Estimate (SE)[Two-sided P-Value]

	15 Min	30 Min	45 Min	60 Min	90 Min	120 Min	180 Min	240 Min
# of Studies	7	7	7	7	7	7	5	5
Fixed Effect	−0.37(0.15)	−1.85(0.23)	−3.54(0.27)	−3.71(0.31)	−2.94(0.30)	−2.00(0.30)	−1.10(0.26)	−0.42(0.27)
DSL	−0.39(0.31)[.25]	−1.36(053)[.044]	−1.33(1.10)[.27]	−2.15(0.91)[.057]	−2.07(0.80)[.041]	−0.95(0.89)[.32]	−0.93(1.16)[.47]	−0.75(1.07)[.52]
Unweighted	−0.25(0.26)[.37]	−1.05(0.49)[.076]	−1.13(1.04)[.32]	−1.62(1.02)[.16]	−1.69(0.89)[.11]	−0.45(1.16)[.74]	−1.01(1.52)[.54]	−1.28(1.45)[.43]
Weighted by Sample Size	−0.16(0.33)[.64]	−0.98(0.49){.090]	−0.76(1.12)[.52]	−1.28(1.09)[.29]	−1.53(0.89)[.14]	−0.08(1.25)[.95]	−0.75(1.61)[.67]	−1.07(1.60)[.54]

Table III

DSL Weights (Normalized)

	Study 1	Study 2	Study 3	Study 4	Study 5	Study 6	Study 7	Total
Real SE	0.162	0.162	0.134	0.135	0.139	0.140	0.128	1.00
Tweaked (Tripled SE)	0.319	0.319	0.069	0.070	0.082	0.085	0.056	1.00

Real DSL estimate=−1.33 SE=1.10 P=0.27 (NB Cochran Q=74.26 with 6 df)
Tweaked DSL estimate= −2.87 SE=1.10 P=0.041 (NB Cochran Q=8.25 with 6 df)