Annals of Internal Medicine | Research and Reporting Methods

Random-Effects Meta-analysis of Inconsistent Effects: A Time for Change

John E. Cornell, PhD; Cynthia D. Mulrow, MD, MSc; Russell Localio, PhD; Catharine B. Stack, PhD, MS; Anne R. Meibohm, PhD; Eliseo Guallar, MD, DrPH; and Steven N. Goodman, MD, PhD

A primary goal of meta-analysis is to improve the estimation of treatment effects by pooling results of similar studies. This article explains how the most widely used method for pooling heterogeneous studies—the DerSimonian–Laird (DL) estimator—can produce biased estimates with falsely high precision. A classic example is presented to show that use of the DL estimator can lead to erroneous conclusions. Particular problems with the DL estimator are discussed, and several alternative methods for summarizing heterogeneous evidence are presented. The authors support replacing universal use of the DL estimator with analyses based on a critical synthesis that recognizes the uncertainty in the evidence, focuses on describing and explaining the probable sources of variation in the evidence, and uses random-effects estimates that provide more accurate confidence limits than the DL estimator.

Ann Intern Med. 2014;160:267–270. www.annals.org
For author affiliations, see end of text.

The basic premise of meta-analysis is that the average of estimates provided by a group of studies is closer to the truth than the estimate provided by an individual study. This premise rests on the assumption that each study is a near-replication of a single experiment and that differences among study results are due only to chance. The technical jargon for this fundamental assumption is that each of the studies is estimating the same "fixed effect," and the corresponding meta-analytic approach is dubbed the "fixed-effects model."

When studies are statistically heterogeneous and differences among their results cannot be explained by chance alone, the meta-analyst faces a conundrum. Qualitative heterogeneity among study designs, patient characteristics, and treatment and comparator regimens may be so great that it does not make sense to combine studies to derive a single summary estimate. However, when the qualitative and quantitative heterogeneity is not so great that a single number summarizing the evidence would be misleading, statistical models that incorporate the extra variability across studies not believed to be due to chance may be used to summarize the data. These models assume that the observed treatment effect for a study is a combination of a treatment effect common to all studies plus a component specific to that study alone. This extra, study-specific component is assumed to be random, hence the jargon that it is a "random effect," with accompanying mathematical models dubbed "random-effects models." The most widely used random-effects model is based on an estimator developed by DerSimonian and Laird in the mid-1980s and is known as the DerSimonian–Laird (DL) estimator (1).

An Example

The **Figure** depicts a statistically heterogeneous set of studies followed by several methods of estimating their average effect. The example is from a 1985 meta-analysis by Collins and colleagues on the effect of administering a diuretic to women at risk for preeclampsia (11), and it is frequently used to illustrate different methods for estimating a common treatment effect when the body of evidence is heterogeneous (12, 13). The effect estimates from the individual studies range from a more than 4-fold statistically significant decrease in the odds of eclampsia with diuretics observed in the study by Fallis and colleagues (5) to an almost 3-fold nonsignificant increase in the study by Tervilä and Vartiainen (9). A visual clue that these studies are statistically heterogeneous is that the confidence limits of several pairs of studies do not overlap.

The **Figure** shows that different statistical approaches to combining data can produce results leading to different conclusions. The fixed-effects model, which is not appropriate for these data, shows a summary effect of 0.67, with 95% confidence limits (0.56 and 0.80) that are 19% less than and greater than that value. The DL random-effects estimate shows a slightly larger effect (odds ratio, 0.60), but the confidence limits are substantially wider—33% less than (0.40) and greater than (0.89) the summary effect, albeit still highly statistically significant. Use of any of the other 3 random-effects estimators depicted in the **Figure** shows identical point estimates for the odds ratio of 0.60 but dramatically wider confidence limits that are 73% less than and greater than 0.60, with the upper limits all exceeding 1.00. The corresponding *P* values range from less than 0.001 for the fixed-effects model to 0.011 for the DL estimator and 0.070 or greater for the other random-effects models.

Statistical Heterogeneity and Uncertainty

The differences noted in the example are due to the ways that the models handle statistical heterogeneity. Statistical heterogeneity refers to variation in the true effects being estimated by each study. We characterize this variation by its SD, a statistic called τ. Assuming normality, we expect 95% of true effects to fall within $\pm 2 \times \tau$ of the central estimate. When odds ratios or relative risks are

Key Summary Points

The decision to calculate a summary estimate in a meta-analysis should be based on clinical judgment, the number of studies, and the degree of variation among studies.

A random-effects model is a meta-analytic approach that incorporates study-to-study variability beyond what would be expected by chance.

The DerSimonian–Laird (DL) method, the earliest and most commonly used random-effects model, is the default method in many software packages.

The DL method produces confidence bounds that are too narrow (and *P* values that are typically too small) when the number of studies is small or when there are substantive differences among study estimates.

Alternative random-effects estimates based on small-sample adjustments, the profile likelihood, or hierarchical Bayesian models that perform better than the DL method are readily available in software packages.

When it is appropriate to pool studies whose estimates vary widely, meta-analytic methods that provide a better accounting of uncertainty than the DL estimator should be used.

used, the normality is on a log scale, so that true study odds ratios or relative risks fall within a range of the estimate multiplied by $e^{\pm 2 \times \tau}$. In the example, τ equals 0.48, so the true study effects are estimated to fall within $0.60 \times e^{\pm 0.96}$, or 2.6 times greater than or less than 0.60 (0.23 to 1.56). This range should be smaller than the actual smallest and largest study estimates, as is the case in this example, with the remainder of the variation assumed to be due to chance.

The models vary in their assumption of how certain we are about τ; this uncertainty is included in the meta-analytic CIs. The DL method assumes that our guess about τ is exactly correct, with no uncertainty; thus, confidence limits are too narrow and the *P* values are too small. In Collins and colleagues' meta-analysis, which pooled a modest number of studies ($n = 9$) with statistically heterogeneous effects, the DL estimator provided the narrowest confidence limits among the random-effects options.

In addition to τ, meta-analysts commonly use statistical tests, such as the Cochran *Q* test, or indices, such as the I^2 index, to help gauge heterogeneity of effects. Both the Cochran *Q* test and the I^2 index are dimensionless measures of statistical heterogeneity. Neither conveys information about actual variation in effect size, and both have low power to detect heterogeneity in situations involving 10 or fewer studies (14).

The DL Estimator and Alternative Approaches

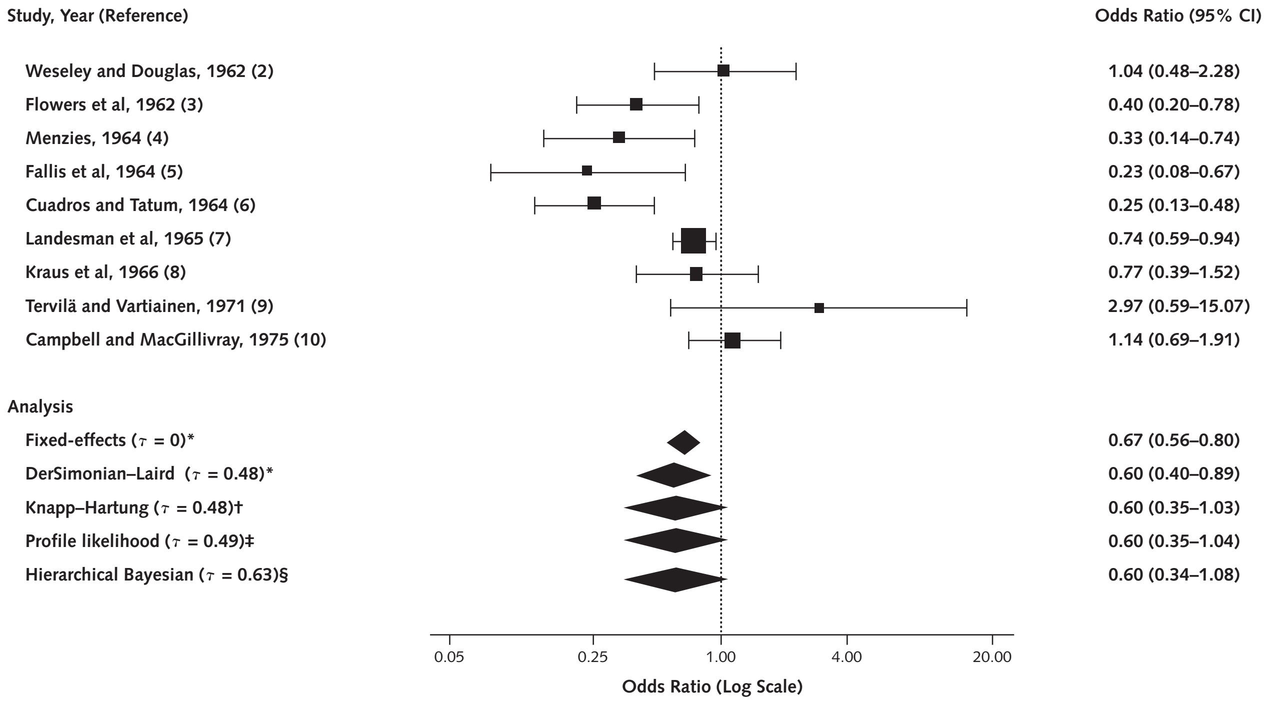
The DL method appeared in the literature just as meta-analytic methods were being adopted to help reviewers quantitatively summarize evidence about medical interventions. It was relatively simple to compute and is still the standard estimator programmed into many meta-analysis software packages, including the RevMan software developed by the Cochrane Collaboration (15). As statisticians began in the 1990s to recognize the problems with the DL approach, they—including DerSimonian and Kacker (16)—proposed a wide range of alternatives that better capture the uncertainty associated with statistical heterogeneity. These included random-effects estimators based on small-sample adjustments, such as the Knapp–Hartung approach (17), likelihood-based methods (13, 18, 19), and hierarchical Bayesian models (20).

The Knapp–Hartung approach, one of the more recent methods, assumes that variances are estimated from small samples, makes small-sample adjustments to the variance estimates, and constructs confidence limits based on the *t* distribution with $k - 1$ degrees of freedom. This estimator produces a wider confidence limit than the DL estimate. It may slightly overestimate the amount of uncertainty in some cases, particularly when dealing with 5 or fewer studies. It is available in some specialized meta-analysis programs and packages, such as the metareg program (21) in Stata (StataCorp, College Station, Texas) and the metafor package (22) in R (R Foundation for Statistical Computing, Vienna, Austria).

Likelihood estimates, which are readily available in such commonly used statistical packages as SAS (SAS Institute, Cary, North Carolina), are computed using standard mixed-effects linear models (18, 19). The profile likelihood is a good method for computing confidence bounds. Unlike estimators based on maximum likelihood or restricted maximum likelihood methods, the profile likelihood allows for asymmetrical intervals and uncertainty in estimation of the between-study variance (τ^2). Simulation studies show that it provides a substantially better accounting of uncertainty than the DL estimator (13, 23). The profile likelihood estimates are available in the metaan package (24) in Stata and the metaLik package (25) in R. The latter provides a more accurate but possibly conservative small-sample profile likelihood estimate of uncertainty (26).

Bayesian random-effects models, which are based on an exact binomial distribution, perform well in many situations where others do poorly, particularly with sparse data and few studies (27, 28). A hierarchical Bayesian equivalent to the mixed-effects model can be fitted using WinBugs or related packages (OpenBugs or JAGS). A hierarchical Bayesian model augmented by careful consideration of priors on τ may provide a better accounting of the uncertainty than non-Bayesian approaches, particularly when the number of studies is small (29). Because selection

Figure. **Heterogeneous evidence from Collins and colleagues' meta-analysis of the effects of diuretics on preeclampsia (11).**



* The metafor package in R was used to compute the fixed-effects estimate and the DerSimonian–Laird random-effects estimate. † The metafor package in R was used to compute the Knapp–Hartung small-sample adjustments, based on the DerSimonian–Laird estimate. ‡ The small-sample (Skovgaard) estimate from the metaLik package in R was used to compute the profile likelihood estimate. The large-sample profile likelihood estimate produced a narrower CI that indicates a statistically significant effect (95% CI, 0.37 to 0.95). § The hierarchical Bayesian estimate was computed using WinBugs and assumed a vague uniform (10, 10) prior distribution for τ. A sensitivity analysis assuming a vague γ (0.001, 0.001) on precision ($1/\tau^2$) produced a slightly smaller but statistically significant 95% CI (0.36 to 0.98).

of a prior distribution for τ or τ^2 is critical to any Bayesian analysis (27), it is important to conduct sensitivity analyses based on different choices for the prior distribution.

Recommendations for Moving Forward

None of the random-effects methods provide a universal solution to the problem of heterogeneity. The decision to summarize data mathematically depends on critical judgment, and the reasons for that decision should be articulated as part of any meta-analysis. Random-effects estimates are most appropriate when it is difficult to attribute observed heterogeneity of effects to clinical or methodological differences among the studies. Proper selection and implementation of a random-effects model requires careful consideration of how many studies are available, the extent to which estimates vary from study to study (τ), and study-specific clinical and methodological factors that contribute to heterogeneity. Large variation in study design, conduct, population, measurements, and analyses suggests that it may be unwise to estimate an average effect. When the number of studies is sufficiently large, organizing analyses around clinically or methodologically important study-level characteristics through stratification or meta-regression may be more informative than a single summary estimate. When there are too few studies to stratify by study-level characteristics, whether pooling is reasonable must be addressed. A critical synthesis that highlights the variations in the evidence and describes the possible sources of variation will almost always be more useful than one that averages over these dimensions and can point the way toward improvement of future studies.

When the decision has been made to pool studies in the face of heterogeneity, the extra uncertainty due to that heterogeneity must be adequately represented. All of the alternative approaches to random-effects modeling more accurately incorporate the uncertainty associated with statistical heterogeneity than does the DL estimator. With a small number of studies, the Knapp–Hartung or small-sample profile likelihood estimator may be the best choices, even if they are conservative. The Bayesian methods are good but require knowledge of Bayesian software and perform best with informed choice of a prior distribution for τ (that is, the range of plausible values for τ).

Insightful synthesis recognizes the qualitative and quantitative heterogeneity and uncertainty of evidence; focuses on describing and explaining the probable sources of variation in the evidence; and, when summarizing heterogeneous evidence quantitatively, uses random-effects estimates that properly represent statistical uncertainty. The original DL estimator from 1986 made it easy to calculate random-effects estimates with computers or spreadsheets, leading to its rapid adoption and incorporation into meta-analytic software. However, after more than 25 years of improvement in methods and software, it is time to move forward and use random-effects methods that provide a more adequate accounting of uncertainty in estimating an average effect when heterogeneity is present.

From University of Texas Health Science Center at San Antonio, San Antonio, Texas; University of Pennsylvania and American College of Physicians, Philadelphia, Pennsylvania; Johns Hopkins School of Public Health, Baltimore, Maryland; and Stanford University School of Medicine, Stanford, California.

Potential Conflicts of Interest: Disclosures can be viewed at www.acponline.org/authors/icmje/ConflictOfInterestForms.do?msNum=M13-2886.

Requests for Single Reprints: John E. Cornell, PhD, Department of Epidemiology and Biostatistics, University of Texas Health Science Center at San Antonio, 7703 Merton Minter Boulevard, San Antonio, TX 78229; e-mail, cornell@uthscsa.edu.

Current author addresses and author contributions are available at www.annals.org.

References

1. **DerSimonian R, Laird N.** Meta-analysis in clinical trials. Control Clin Trials. 1986;7:177-88. [PMID: 3802833]
2. **Weseley AC, Douglas GW.** Continuous use of chlorothiazide for prevention of toxemia of pregnancy. Obstet Gynecol. 1962;19:355-8. [PMID: 14006267]
3. **Flowers CE Jr, Grizzle JE, Easterling WE, Bonner OB.** Chlorothiazide as a prophylaxis against toxemia of pregnancy. A double-blind study. Am J Obstet Gynecol. 1962;84:919-29. [PMID: 13893680]
4. **Menzies DN.** Controlled trial of chlorothiazide in treatment of early pre-eclampsia. Br Med J. 1964;1:739-42. [PMID: 14102017]
5. **Fallis NE, Plauche WC, Mosey LM, Langford HG.** Thiazide versus placebo in prophylaxis of toxemia of pregnancy in primigravid patients. Am J Obstet Gynecol. 1964;88:502-4. [PMID: 14123429]
6. **Cuadros A, Tatum HJ.** The prophylactic and therapeutic use of bendroflumethiazide in pregnancy. Am J Obstet Gynecol. 1964;89:891-7. [PMID: 14207556]
7. **Landesman R, Aguero O, Wilson K, LaRussa R, Campbell W, Penaloza O.** The prophylactic use of chlorthalidone, a sulfonamide diuretic, in pregnancy. Br J Obstet Gynaecol. 1965;72:1004-10.
8. **Kraus GW, Marchese JR, Yen SS.** Prophylactic use of hydrochlorothiazide in pregnancy. JAMA. 1966;198:1150-4. [PMID: 5332983]
9. **Tervilä L, Vartiainen E.** The effects and side effects of diuretics in the prophylaxis of toxaemia of pregnancy. Acta Obstet Gynecol Scand. 1971;50:351-6. [PMID: 4945572]
10. **Campbell DM, MacGillivray I.** The effect of a low calorie diet or a thiazide diuretic on the incidence of pre-eclampsia and on birth weight. Br J Obstet Gynaecol. 1975;82:572-7. [PMID: 1096930]
11. **Collins R, Yusuf S, Peto R.** Overview of randomised trials of diuretics in pregnancy. Br Med J (Clin Res Ed). 1985;290:17-23. [PMID: 3917318]
12. **Biggerstaff BJ, Tweedie RL.** Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. Stat Med. 1997;16:753-68. [PMID: 9131763]
13. **Hardy RJ, Thompson SG.** A likelihood approach to meta-analysis with random effects. Stat Med. 1996;15:619-29. [PMID: 8731004]
14. **Higgins JP, Thompson SG.** Quantifying heterogeneity in a meta-analysis. Stat Med. 2002;21:1539-58. [PMID: 12111919]
15. Review Manger (RevMan) [computer program]. Version 5.2. Copenhagen, Denmark: Nordic Cochrane Center, Cochrane Collaboration; 2012.
16. **DerSimonian R, Kacker R.** Random-effects model for meta-analysis of clinical trials: an update. Contemp Clin Trials. 2007;28:105-14. [PMID: 16807131]
17. **Knapp G, Hartung J.** Improved tests for a random effects meta-regression with a single covariate. Stat Med. 2003;22:2693-710. [PMID: 12939780]
18. **Normand SL.** Meta-analysis: formulating, evaluating, combining, and reporting. Stat Med. 1999;18:321-59. [PMID: 10070677]
19. **Whitehead A.** Meta-analysis of Controlled Clinical Trials. New York: J Wiley; 2002.
20. **Higgins JP, Thompson SG, Spiegelhalter DJ.** A re-evaluation of random-effects meta-analysis. J R Stat Soc Ser A Stat Soc. 2009;172:137-59. [PMID: 19381330]
21. **Sterne JAC.** Meta-analysis in Stata: An Updated Collection from the Stata Journal. 1st ed. College Station, Texas: Stata Pr; 2009.
22. **Viechtbauer W.** Conducting meta-analysis in R with the metafor package. J Stat Softw. 2010;36:1-48.
23. **Brockwell SE, Gordon IR.** A comparison of statistical methods for meta-analysis. Stat Med. 2001;20:825-40. [PMID: 11252006]
24. **Kontopantelis E, Reeves D.** metaan: Random-effects meta-analysis. Stata J. 2010;10:395-407.
25. **Guolo A, Varin C.** The R package metaLik for likelihood inference in meta-analysis. J Stat Softw. 2012;50:1-14.
26. **Guolo A.** Higher-order likelihood inference in meta-analysis and meta-regression. Stat Med. 2012;31:313-27. [PMID: 22173666]
27. **Warn DE, Thompson SG, Spiegelhalter DJ.** Bayesian random effects meta-analysis of trials with binary outcomes: methods for the absolute risk difference and relative risk scales. Stat Med. 2002;21:1601-23. [PMID: 12111922]
28. **Lambert PC, Sutton AJ, Burton PR, Abrams KR, Jones DR.** How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. Stat Med. 2005;24:2401-28. [PMID: 16015676]
29. **Higgins JP, Whitehead A.** Borrowing strength from external trials in a meta-analysis. Stat Med. 1996;15:2733-49. [PMID: 8981683]

Current Author Addresses: Dr. Cornell: Department of Epidemiology and Biostatistics, University of Texas Health Science Center at San Antonio, 7703 Merton Minter Boulevard, San Antonio, TX 78229.

Dr. Mulrow: University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Drive, San Antonio, TX 78229.

Dr. Localio: Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania, 635 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104-6021.

Drs. Stack and Meibohm: American College of Physicians, 190 N. Independence Mall West, Philadelphia, PA 19106.

Dr. Guallar: Welch Center for Prevention, Epidemiology and Clinical Research, Johns Hopkins Medical Institutions, 2024 East Monument Street, Room 2-645, Baltimore, MD 21287.

Dr. Goodman: Stanford University School of Medicine, 259 Campus Drive, T265 Redwood Building/HRP, Stanford, CA 94305.

Author Contributions: Conception and design: J.E. Cornell, C.D. Mulrow, R. Localio, S.N. Goodman.

Analysis and interpretation of the data: J.E. Cornell, R. Localio.

Drafting of the article: J.E. Cornell, C.D. Mulrow, R. Localio, C.B. Stack, A.R. Meibohm, S.N. Goodman.

Critical revision of the article for important intellectual content: J.E. Cornell, C.D. Mulrow, R. Localio, C.B. Stack, A.R. Meibohm, E. Guallar, S.N. Goodman.

Final approval of the article: J.E. Cornell, C.D. Mulrow, R. Localio, C.B. Stack, A.R. Meibohm, E. Guallar, S.N. Goodman.

Statistical expertise: J.E. Cornell, R. Localio, A.R. Meibohm, S.N. Goodman.

Administrative, technical, or logistic support: C.D. Mulrow.

Collection and assembly of data: J.E. Cornell.