e-Xtra*

Phytopathology
SYMPOSIUM

Meta-Analysis for Evidence Synthesis in Plant Disease Epidemiology and Management
Presented at the Annual Meeting of The American Phytopathological Society August 4, 2009, Portland, OR

Meta-Analysis for Evidence Synthesis in Plant Pathology: An Overview

L. V. Madden and P. A. Paul

Department of Plant Pathology, Ohio State University, Ohio Agricultural Research and Development Center (OARDC), Wooster 44691.

ABSTRACT

Madden, L. V., and Paul, P. A. 2011. Meta-analysis for evidence synthesis in plant pathology: An overview. Phytopathology 101:16-30.

Meta-analysis is the analysis of the results of multiple studies, which is typically performed in order to synthesize evidence from many possible sources in a formal probabilistic manner. In a simple sense, the outcome of each study becomes a single observation in the meta-analysis of all available studies. The methodology was developed originally in the social sciences by Smith, Glass, Rosenthal, Hunter, and Schmidt, based on earlier pioneering contributions in statistics by Fisher, Pearson, Yates, and Cochran, but this approach to research synthesis has now been embraced within many scientific disciplines. However, only a handful of articles have been published in plant pathology and related fields utilizing meta-analysis. After reviewing basic concepts and approaches, methods for estimating parameters and interpreting results are shown. The advantages of meta-analysis are presented in terms of prediction and risk analysis, and the high statistical power that can be achieved for detecting significant effects of treatments or significant relationships between variables. Based on power considerations, the fallacy of naïve counting of *P* values in a narrative review is demonstrated. Although there are many advantages to meta-analysis, results can be biased if the analysis is based on a nonrepresentative sample of study outcomes. Therefore, novel approaches for characterizing the upper bound on the bias are discussed, in order to show the robustness of meta-analysis to possible violation of assumptions.

Additional keywords: Fusarium head blight, *Gibberella zeae*, wheat scab.

As recounted by Lipsey and Wilson (29), "In 1952, Hans Eysenck started a raging debate in clinical psychology by arguing that psychotherapy had no beneficial effects on patients" (14). Over the next 25 years or so, several hundred studies had been conducted, producing a "dizzying array of positive, null, and negative results" (29), with no resolution to the debate. Then, in 1977, Smith and Glass (59) analyzed the *results* (not the original individual observations) of nearly 400 studies and concluded that psychotherapy was, indeed, a very effective treatment. Glass called their statistical method meta-analysis (19), and the term has stuck. At almost the same time that Glass and Smith were assessing the effectiveness of psychotherapy, others were independently using the same general statistical methodology for collections of studies in the social sciences. Rosenthal and Rubin (52) assessed the effects of interpersonal expectations on behavior, and Schmidt and Hunter (56) assessed the validity of generalization of employment tests. Many of the important principles, concepts, and protocols underlying the analyses in these social-science research syntheses were given in important early books by Glass et al. (20), Rosenthal (51), and Hedges and Olkin (24).

As is usual in statistics, most "new" methods are not entirely new, and there were several precursors to the pioneering work of Smith, Glass, Rosenthal, Rubin, Schmidt, and Hunter (5). As early as 1904, Pearson (45) combined estimated correlation coefficients from multiple studies. Fisher (17) and Tibbett (63) showed how to combine the achieved significance levels (*P* values) from several independent studies to determine a combined significance level for the collection of studies. Importantly, for those working in the agricultural sciences, Yates and Cochran (70) and Cochran (7) showed how to combine results from several experiments to determine overall mean responses and related statistics. Despite these and related contributions (literature citation 5 provides excellent history), combining results from multiple studies to form an overall analysis was not common, and most investigators did not see the need for such efforts, until the work of Glass and others in the late 1970s.

The early applications of meta-analysis were in the social sciences, but the methodology was eventually adopted in numerous disciplines (2,4,5,21,31). In medical research, the upsurge began in the 1980s, and by the 1990s, "published meta-analyses were ubiquitous" in this field (68,69). Since then, there has been a steady rise in the effort given to this type of data analysis. For instance, based on the ISI Web of Knowledge (for science and social sciences), there were fewer than 60 articles published per year from 1980 to 1987 on meta-analysis. A decade later, there were about 1,000 new articles published per year, and most recently, there have been over 3,000 new articles per year. At the

Corresponding author: L. V. Madden; E-mail address: madden.1@osu.edu

*The *e*-Xtra logo stands for "electronic extra" and indicates that the online version contains a supplemental file demonstrating the use of SAS software for conducting meta-analysis.

doi:10.1094/PHYTO-03-10-0069
© 2011 The American Phytopathological Society

end of 2008, more than 25,000 articles had been published dealing with meta-analysis. As Chalmers and Lau (6) wrote more than 15 years ago, "It is obvious that the new scientific discipline of meta-analysis is here to stay."

Rosenberg et al. (49) were the first to strongly promote the value of meta-analysis in plant pathology, and since then, several articles have been published utilizing this method (37,39–44,57). The remainder of this article presents an overview of meta-analysis and shows how the methodology can be applied in plant disease management (or epidemiology) for evidence synthesis. In particular, after defining the term and discussing some of the controversies concerning the meta-analytic methodology, we present statistical models for performing a meta-analysis, and show how to fit the models and interpret results. Through the presented model and an example analysis, we show the high statistical power (4) that can be achieved by using this method, but also point out the dangers of the publication bias (53) that can occur when using meta-analysis. Through the consideration of statistical power, we also show the inherent fallacy of the main competitor to meta-analysis: the narrative review and "vote counting" (4,23,27). We finally discuss, very briefly, some extensions to the basic meta-analytical model (1,2,31,32,46–48).

META-ANALYSIS—GENERAL COMMENTS

Definitions and concepts. There are many similar definitions of meta-analysis (19,27,36,62). In essence, the method involves "the combination of results from multiple independent studies" (62). In standard usage, meta-analysis deals only with the results from different studies, and not with the original observations. However, it is also possible to perform a meta-analysis using all the observations from the studies (called individual participant data [IPD]), although these observations are typically not available (4,62).

The result from an individual study that becomes part of a meta-analysis is often called the estimated effect size. We defer formal definition of an effect size until later, but, as an example, a typical result being analyzed from each study involves the difference of two means (or a function of two means). Sometimes a meta-analysis is performed to test a particular hypothesis about an effect size (is it 0 or not?), and sometimes the analysis is performed to characterize the variation (distribution) of effect sizes across all studies, or to determine what factors may be influencing the magnitude of the effect sizes (4,49).

Meta-analysis is built on the principle that science is meant to be a cumulative process, where individual studies, surveys, and observations contribute to the overall total knowledge base (4,8,27). Results of individual studies can contribute something to the total, but it is the collection of results from many sources that matter in moving science forward or in informing our decision-making process. Hunter and Schmidt (27) put it elegantly and also a bit bluntly, in stating: "…a single study will not resolve a major issue. Indeed, a small sample study will not even resolve a minor issue. Thus, the foundation of science is the culmination of knowledge from the results of many studies." From the perspective of Chalmers et al. (5) and others (4,29,68), meta-analysis is the methodology that allows investigators to cumulate evidence scientifically, with the hope of improving interpretation of phenomena, developing coherent theories, or testing hypotheses.

It should be noted that meta-analysis has been considered controversial in the past (15,16,59). Chapter 43 in Borenstein et al. (4) provides an excellent review and critique of the various criticisms of this methodology, and shows that most of the criticisms are of little relevance to properly performed meta-analysis. We touch on some of the relevant issues in Boxes 1 and 2.

EFFECT SIZES AND INDIVIDUAL STUDIES

An individual study. Consider this illustration for an individual study. A researcher is interested in the effect of a fungicide on severity of a crop disease. A study is conducted consisting of at least two treatments, with label T for the treatment of interest here, and *C* for the control. There are four replications of both T and C. Disease severity (y) is assessed at a single time during the growing season, and a mixed model (30) or analysis of variance (ANOVA, a special case of a linear mixed model) is used to determine if treatment affects y. We assume that disease is measured in such a way that y is normally distributed. The mean y is determined for each group, $\bar{y}_T$ and $\bar{y}_C$; these means are estimates of the expected values for the respective populations, $\hat{\mu}_T$ and $\hat{\mu}_C$. One summary of the effectiveness of the fungicide is the difference in mean disease severity (D):

$$D = \bar{y}_C - \bar{y}_T = \hat{\mu}_C - \hat{\mu}_T \quad (1)$$

This difference is known as an estimated treatment effect in an ANOVA. In meta-analysis, it is referred to as an estimated effect size, effect size statistic, or effect size index (4,29); an estimated effect size is an estimated parameter, combination of estimated parameters (such as a difference), or a function of estimated parameters, for an individual study. Estimated effect sizes are random variables.

There are several possible effect sizes that can be used, depending on the type of response variable (y) being used and the objectives of the investigator (4,21,29,40–42). We use z as the

BOX 1

Controversy 1: "Garbage-in, garbage-out." Most criticisms of meta-analysis concern the selection (and interpretation) of the individual studies that comprise the data set for the meta-analysis. The "garbage-in, garbage-out" problem is frequently identified with published meta-analyses, going back to Eysenck (15). That is, just because a study was conducted (or even published) does not mean that it was correctly done or that the results are accurate or meaningful. Most individuals who are reading this article probably have conducted experiments that failed for various reasons (such as the failure of a growth chamber at a critical time), and are simply discarded. One would not want to be forced to analyze the data from such a study (and present the results in some form) just because it was carried out. By the same token, we all know from careful reviewing of the literature that some published studies were, in fact, poorly (or inadequately) performed or analyzed, or that the reported findings were not described well. Therefore, blindly including the results from these studies together with the results from properly conducted studies would be misleading and not advised.

Handling the above problem is relatively straight-forward, for the most part. This is done by defining strict criteria for the selection of studies in the research synthesis before the study results are actually accumulated for analysis (9,29). It is also important that the criteria for study selection are well described in the reported meta-analysis; that way, readers can make decisions about whether the criteria are reasonable or appropriate. Lipsey and Wilson (29), Cooper et al. (9), and Borenstein et al. (4) have extensive coverage of this topic.

generic symbol for any chosen estimated effect size (e.g., $z = D$ if the mean difference is used). Note that z is an estimate of a parameter ζ, the true effect size. For instance, equation 1 provides an estimate of the true mean difference, $\mu_C - \mu_T$.

Significance of a treatment effect in a single study requires an estimate of the precision of the estimated effect. For a mean difference, the precision is represented by the standard error of the difference [$SE(D)$], or its square, the (estimated) variance of the mean difference [$SE^2(D) = (SE(D))^2 = V_D$]. The ratio of D to $SE(D)$ has a Student's t distribution under the null hypothesis of no treatment effect, and this ratio serves as the basis for statistical inference. $SE(D)$ is routinely calculated and displayed in output of ANOVA programs, and $SE(D)$ or V_D can easily be determined from other statistics by hand (4,29). For instance, if V is the residual or mean square error from a one-way ANOVA, and each treatment group has the same number of replications (n), then

$$(SE(D))^2 = V_D = 2V/n \tag{2}$$

Also, the least significant difference (LSD) is equal to SE_D multiplied by the critical t value at a specified significance level ($t^*_{1-\alpha/2,df}$), so that $SE(D) = \text{LSD}/t^*_{1-\alpha/2,df}$.

Different estimated effect sizes have different variances (and different means of calculating them) (4,9). We use s^2 as the generic symbol for the variance of the estimated effect size (e.g., $s^2 = V_D$; or $s = \sqrt{V_D} = SE(D)$]. Often, s^2 is known as the sampling variance for the study.

In a meta-analysis, one analyzes the estimated effect sizes from a collection of K studies. From each study, an estimated effect size and its sampling variance are required to carry out the analysis. We use an i subscript to index the individual studies ($i = 1, \ldots, K$). In a standard meta-analysis, the ith study is represented by the pair of values (z_i, s_i^2), the estimated effect size and its (estimated) variance.

Effect sizes for treatment effects. A mean difference (equation 1) is certainly an intuitive summary of the effect of a treatment on a response variable, but there are many other possibilities (4,29). When the magnitude of the response variable (e.g., disease severity) in the control varies greatly from study-to-study, measurement of treatment effect on an absolute scale may not be the most informative. For instance, if $\bar{y}_C = 5$, then $D = 3$ ($\bar{y}_T = 2$) could be considered a large treatment effect. However, if $\bar{y}_C = 50$, then $D = 3$ (i.e., $\bar{y}_T = 47$) may not be considered that substantial of a treatment effect. It is thus often informative to quantify the effect of a treatment through the so-called percent control (C), the percentage reduction in treatment mean relative to the control mean:

$$C = 100\left(\frac{\bar{y}_C - \bar{y}_T}{\bar{y}_C}\right) = 100\left(\frac{D}{\bar{y}_C}\right) = 100\left(1 - \bar{y}_T/\bar{y}_C\right) \tag{3}$$

As seen by equation 3, C provides a scaled version of the mean difference. If $\bar{y}_C = 50$ and $\bar{y}_T = 20$, then $D = 30$ and $C = 60\%$ (i.e., use of the fungicide results in a 60% reduction in disease, on average). However, if $\bar{y}_C = 5$ and $\bar{y}_T = 2$, then $D = 3$, but C is still equal to 60% (the relative reduction is unchanged). Both D and C are 0 when the means are the same for both groups (although C is undefined if the control mean is 0).

The ratio of the means in the treatment and control is known as the estimated response ratio ($R = \bar{y}_T/\bar{y}_C$) (22). Thus, the percent control can be written simply as $C = 100(1 - R)$. R (or C) could be directly used as the effect size statistic in a meta-analysis, but, as clearly explained by Hedges et al. (22), this random variable has some undesirable statistical properties. However, the log of R [$L = \ln(R)$] is much better behaved statistically, and is a useful estimated effect size when performing a meta-analysis (4,22,41, 42). The estimated variance (sampling variance) of L is given by

$$V_L = \frac{V}{n}\left(\frac{1}{\bar{y}_C^2} + \frac{1}{\bar{y}_T^2}\right) \tag{4a}$$

where, as before, V is the residual variance from the study (when y is being analyzed directly) and n is the number of replications. One can also write the variance of the log response ratio in terms of the variance of the difference:

$$V_L = \frac{V_D}{2}\left(\frac{1}{\bar{y}_C^2} + \frac{1}{\bar{y}_T^2}\right) \tag{4b}$$

BOX 2

Controversy 2: "Mixing apples and oranges." Typically, the population of studies being included in a research synthesis is diverse. The individual studies are performed by many different individuals at many different locations and at many different times. Specific methods used in individual studies, and the response variables being measured, are chosen for different reasons by different investigators. Thus, one could ask whether or not the studies being considered are too different from each other to conduct an overall single meta-analysis.

A hypothetical example is helpful. Consider a research synthesis of the effect of the fungicide captan on apple scab. Given that this fungicide has been used for many decades around the world, it would be quite an undertaking to find all the papers and reports, but we assume that this can be done (or that we can obtain a random sample of the population of studies). Many factors could contribute to the diversity of studies, some of which we list here. Captan is produced by several manufacturers, is available in multiple formulations (e.g., 80 WG, 50 WP), and can be applied at different rates (a.i./ha) and with different spraying equipment. Studies can also vary in terms of the timing of the applications, the number of sprays, and whether or not captan is mixed with other fungicides. The apple variety (varieties) will vary across studies, as well as the horticultural practices used for production. The other diseases and pests of concern also will vary among studies. Moreover, the form of disease assessment, and the timing of assessment(s), will not be the same for the different studies. Readers can probably think of other types of diversity.

Clearly, studies can be quite different from each other. The important thing to keep in mind, however, is that meta-analysis involves the analysis of related (or similar) studies, not identical studies. In a technical sense, a random-effects meta-analysis (4,68) can be used to account for the heterogeneity of results (effect sizes) that may be found with diverse studies. Characteristic features of the individual studies (e.g., a.i./ha of captan) can be recorded (as so-called moderator variables) and formally assessed in the meta-analysis to see if they significantly affect the effect size of interest (e.g., is one fungicide formulation more effective than another?) (4,41,43). However, it is also possible that the meta-analyst will need to be more specific in the research questions being addressed. For instance, one could ask: What is the effect of captan on apple scab when applied no later than a certain growth stage (e.g., "green tip") and not used in combination with other fungicides? This will narrow down the list of possible studies to consider for inclusion in the analysis, and potentially reduce the variability.

Formulae for unequal n for each group, or separate variances for each group, are also available (4,22). It is instructive to note that V_L depends on the variability of y in a study and also, in an inverse manner, on the magnitude of the means for the two groups.

Use of the log response ratio is also valuable when the form of the response variable, or conditions related to how the response variable is measured, is not the same in every individual study that makes up the K studies. For instance, if y represents counts of lesions per leaf in one study, but y is the proportion of diseased leaves per plant in another study (a form of disease incidence), then there is no common scale to determine D for each study (D_i). However, use of L (or R or C) reduces the impact of the scale differences so that the direct comparisons of treatment effects can be more readily made across all studies, at least as an approximation.

The standardized mean difference (Cohen's d), or its adjustment (Hedges' g), is especially common as an effect size in the social sciences and other fields when the measurement scale varies among studies. For binary data (e.g., diseased or not), the log of the risk ratio, log of the odds ratio, and the risk difference are all popular measures (4). Details on the estimation of each effect size and its sampling variance are given in Borenstein et al. (4) and most other meta-analysis textbooks.

Effect sizes when treatments are not evaluated. An effect size could characterize the relationship between two variables (27,43,44,49,57), in the form of an estimated correlation coefficient (r) or an estimated slope of a regression model (b). Most statistical programs automatically calculate and display the standard error of the estimated slope, SE(b), when fitting a linear model to data in a study; the square of this standard error is the sampling variance for b (V_b) that is utilized together with b in a meta-analysis.

The strength of the relationship between two variables is represented by r. Most meta-analysts prefer (for statistical reasons) to work with a transformation of r, known as the Fisher Z_r transformation, as the estimated effect size (4,43). This transformation is given by

$$Z_r = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \quad (5)$$

Interestingly, the variance of this r transformation is very simple: $V_{Zr} = 1/(n - 3)$, where n here is the number of pairs of observations in a study used to determine the correlation coefficient. Thus, because the variance only involves n, if the correlation coefficient is reported in a study, one can easily determine the sampling variance (a critical part of a meta-analysis) if the sample size is reported.

MODELS AND PARAMETER ESTIMATION—AN EXAMPLE ANALYSIS

As part of the U.S. Wheat and Barley Scab Initiative, Uniform Fungicide Trials have been conducted for over a decade on the use of various fungicides for the control of Fusarium head blight (FHB) of wheat. An expert panel has developed standardized protocols for the experiments, so that all studies were conducted in a similar manner across multiple states and years. Usually, four to seven fungicide treatments were evaluated in a study (including a control), but sometimes a higher number of treatments were considered (34,42). Because there was a proactive effort from the start to have all investigators report their findings in reports or proceedings (whether there was disease present or not, or whether the treatments appeared to work or not), results from most studies were obtained. This permitted the calculation of estimated effect sizes and their sampling variances for the K studies.

The Uniform Fungicide Trials have served as the basis for several meta-analyses (30–44) as well as other statistical modeling (34). Here, we use the results for the effects of the fungicide tebuconazole (Folicur 3.6F; Bayer Crop Science, Research Triangle Park, NC), applied at wheat growth stage 10.5.1 (beginning of anthesis), on deoxynivalenol (DON) toxin in harvested wheat grain (41) to demonstrate a meta-analysis and to examine several issues of relevance in conducting this type of analysis. Many of the specific results are presented here for the first time.

The response variable in each individual study was DON (ppm); Folicur treatment was coded as T, and the control (no fungicide) was coded as C. For the ith study, the estimated effect size of primary interest was the percent control (C_i), but as discussed above, the log response ratio (L_i) was analyzed directly. The sampling variance for the log ratio was determined with equation 4a. There were 101 studies (K = 101) analyzed.

Before conducting any type of statistical modeling, it is a good practice to visualize the data (30,33). Histograms of the effect size statistics (z_i) can be of value. A histogram of the L_i values was fairly symmetrical, with values between −1.8 and 1.2 (Fig. 1A). Note that a log response ratio below 0 indicates a positive percent control. The corresponding percent control values were more asymmetrical, with values between −233 and 100% (Fig. 1B). Negative values occur when mean DON in the control is lower than mean DON in the treatment. Although the largest positive value of C is 100% (when the mean response in the treatment is 0), the lower limit can be less than −100%.

Although histograms of z_i values (e.g., L_i) are useful, they should not be over-emphasized (or used in isolation) when summarizing the estimated effect sizes. This is because histograms can be misleading due to the fact that the precision of the individual z_i values can vary greatly, as indicated by the sampling variances (s_i^2 values). The histogram does not provide information on the s_i^2 for the displayed estimated effect sizes.

Meta-analysts often recommend a so-called Forest plot for data visualization, which shows the individual estimated effect sizes together with either standard errors [s_i; equal to $\sqrt{V_{Li}}$ (equation 4a) here] or 95% confidence intervals for the individual studies. One variation of this type of graph is given in Figure 1C, which shows the $z_i \pm s_i$ for the 101 studies. Since there is no natural ordering of these studies (they come from many different states and years), the studies can be ordered (as here) based on the magnitude of z_i. The full range of estimated effect sizes can be easily seen with this graph as well as the precision of the estimates. The values on the far left are the ones with large (positive) percent control, and the values on the right are those with lower DON in the control than in the treatment (negative percents control); the many effect sizes around 0 are the ones where the treatment and control means were very similar. The wide range of estimated effect sizes is a visual demonstration of the variability (heterogeneity) in the treatment effects. It is very apparent that the precision (sampling variation) varies a great deal in these 101 studies. This is partly because the standard error (or the sampling variance; equation 4a) is inversely proportional to the means in the two groups. Very low mean y (near 0) will result in a very large standard error because the estimate of a log ratio (or a ratio) is imprecise when the numerator and denominator are close to 0.

Model. A statistical model for a meta-analysis is given by

$$z_i = \zeta + u_i + \varepsilon_i \quad (6)$$

where z_i is the estimated effect size (L_i in this case), ε_i is the within-study (sampling) variability term (the residual), u_i is the among-study variability term, and ζ is the unknown expected effect size (the expected z) for the population of studies (21,41,65,68). Both u and ε are considered random effects that are

normally distributed (although the normality assumption can be relaxed) with zero mean; it is also assumed that u and ε are independent. The variance of ε for the ith study is s_i^2, which is the sampling variance (V_{Li} in this case; equation 4a); it is assumed that s_i^2 values are known for each study and are, therefore, not estimated during model fitting. The variance for the among-study random effect (u_i) is given by σ^2. The distributional assumptions for equation 6 can be written succinctly as

$$u_i \sim N(0, \sigma^2)$$
$$\varepsilon_i \sim N(0, s_i^2)$$

With these assumptions, the marginal-distributional result shows that z_i is normally distributed with mean ζ and variance $\sigma^2 + s_i^2$. One can write this simply as

$$z_i \sim N(0, \sigma^2 + s_i^2) \quad (7)$$

Equation 6 is the classical model for meta-analysis (68); it is a hierarchical model, with randomness manifested at two levels. This can be seen by writing the equation in components. The estimated effect size for the ith study can be first expressed as

$$z_i = \zeta_i + \varepsilon_i \quad (8a)$$

where ζ_i is the true (but unknown) effect size for the ith study, and ε_i is the residual or within-study variability term [$\varepsilon_i \sim N(0, s_i^2)$]. Then, at the second level, it is assumed that the true effect size is not constant but varies randomly among the studies. This is written as

$$\zeta_i = \zeta + u_i \quad (8b)$$

where ζ (zeta without a subscript) is the expected value of variable ζ_i values, and u_i is the random effect of the ith study on the true effect size [$u_i \sim N(0,\sigma^2)$]. This formulation is especially instructive because it emphasizes the fact that the true effect size can vary from study-to-study due to study heterogeneity. Some sources of heterogeneity include variation in how the studies were conducted, how the response variable was measured, or, for field studies, differences in environmental conditions. Substituting equation 8b in equation 8a gives equation 6.

Furthermore, based on equation 8a, one can specify the distribution of z_i given the true effect size as

$$z_i|\zeta_i \sim N(\zeta_i, s_i^2) \quad (9a)$$

and, based on equation 8b, the distribution of the true effect size as

$$\zeta_i \sim N(\zeta, \sigma^2) \quad (9b)$$

The marginal distribution resulting from equations 9a and 9b is equation 7.

The classical meta-analytical model can thus be expressed as: (i) equation 6 (with associated definitions of the random effects); (ii) equations 8a and 8b; or (iii) equations 9a and 9b. These are all equivalent, although it is likely only one of these formulations will be used in a particular paper. It should be further noted that the symbols used vary greatly in different papers and books, and there is no single best choice for notation.

A special case of equation 6 is for the situation when the among-study variance is zero or assumed to be zero ($\sigma^2 = 0$); this means that $u_i = 0$ in all studies. Based on equation 8b, this also indicates that the true effect size is identical for all studies (i.e., $\zeta_i = \zeta$). Therefore, ζ would no longer be viewed as a (true) expected effect size, but a common effect size across all studies (4,26).

The model with zero for σ^2 in equation 6 is often called a fixed-effect model in meta-analysis (9,26,68), and the model with nonzero σ^2 is called a random-effect model. However, meta-analysts use these labels somewhat differently from the way they are used by other statisticians (4). So, it is probably more appropriate to call the "fixed-effect" version of equation 6 the *common-effect* model (26). However, the fixed and random labels are used so often, that it is unlikely that there will be any major changes to the names given to the models.

Model fitting methods. The essence of a meta-analysis involves the estimation of ζ and σ^2, and the utilization of these

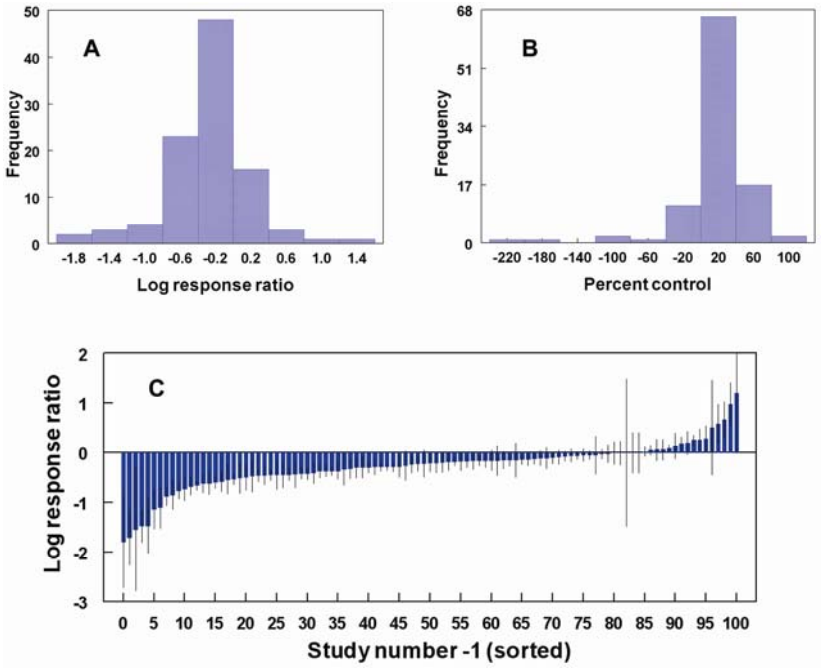

FIGURE 1

Summary of estimated effect sizes for the influence of tebuconazole (applied at growth stage 10.5.1) on deoxynivalenol (DON) toxin concentration in wheat grain, based on the data given in Paul et al. (41). **A,** Frequency distribution of the log response ratio (log of the ratio of mean in the fungicide treatment divided by mean in the fungicide-free check) for each of the studies (L_i or z_i). **B,** Frequency distribution of the percent control (relative reduction in mean DON by the fungicide treatment). **C,** Version of a forest plot for the individual z_i values, with the effect sizes sorted from low to high. Bars in **C** represent standard errors.

estimates in statistical inference and general quantification of the effect-size distribution. There are several methods of parameter estimation (11,58,68,69). The traditional approach to random-effect parameter estimation is known as the method of moments (12); this approach may still be the most common approach and is the basis for some popular specialized computer programs. This method is easy to perform and parameter estimation is also less dependent on some of the distributional assumptions.

A potentially superior approach is maximum likelihood (ML) or restricted maximum likelihood (REML) estimation (30). Although more dependent on assumptions (such as normality) regarding the data and random effects, the likelihood methods result in parameter estimates with many desirable statistical properties (30,66,68) and powerful tests of hypotheses when there are many studies (large K). The distributional properties of the estimated among-study variance ($\hat{\sigma}^2$) can also be characterized with the likelihood-based approaches. Moreover, likelihood-based methods allow investigators to use the full suite of statistical tools developed in the last two decades for fitting mixed models to data (30); this makes it possible to fit models much more complex than the one given in equation 6 (and much more complex than ones that could be fitted by the method of moments). We prefer likelihood-based approaches for these reasons.

Although theory shows that ML will give biased variance estimates, the bias is small with large numbers of studies (K); ML and REML methods will yield very similar results with a large K (e.g., $K > 30$). With a small number of studies, REML-based methods are preferable because they produce unbiased variance estimates. The likelihood-based methods are all accessible using several mixed-model software programs, such as the MIXED procedure in SAS (SAS Institute, Inc., Cary, NC), although users of these programs must learn certain "tricks" for fitting the meta-analytical model (65,68).

A third important way to fit equation 6 to effect-size statistics is to use Bayesian methods (26). This approach is becoming more and more popular in medical statistics and other disciplines. Mila and Ngugi (35) present a discussion on this approach. Interestingly, with certain choices for a prior distribution of a parameter, and other distributional assumptions, Bayesian point estimates are very close to REML estimates (68). However, Bayesian methods have the additional advantage of more fully accounting for uncertainty in parameter estimates in conducting statistical inference (tests of significance, as an example).

It is instructive to see parameter estimation formulae for the fit of equation 6. These equations are appropriate for both moment- and likelihood-based methods. The estimated ζ is given as

$$\hat{\zeta} = \frac{\sum w_i z_i}{\sum w_i} \tag{10}$$

where w_i is the weight for each study:

$$w_i = \frac{1}{\hat{\sigma}^2 + s_i^2} \tag{11}$$

Because, in general, the sampling variance is different for different studies, the study weights are also different. Note that the summation is over all K studies. The standard error for the estimated ζ is given by

$$SE(\hat{\zeta}) = \left(\sum w_i\right)^{-1/2} \tag{12}$$

With the distributional assumption in equation 7, the estimated expected (mean) effect size is also normally distributed (at least asymptotically), with estimated variance of $[SE(\hat{\zeta})]^2$.

As can be seen from equations 10 and 11, at its core, meta-analysis is a method of obtaining weighted averages of effect sizes. If the sampling variance was the same for all studies, then equation 10 would produce the simple arithmetic average of the z_i values. But because s_i^2 do vary, studies with small sampling variance receive proportionally more weight in determining the estimated expected effect size. If sample sizes within studies vary, then, all other things being equal, studies with larger n will have more weight than studies with smaller n (e.g., equations 2 and 4a). Also, studies with smaller residual variance will have more weight than studies with larger residual variance.

Estimation of ζ requires an estimate of σ^2, but, although not shown here, estimation of σ^2 requires an estimate of ζ. This makes ML and REML estimation an iterative process, where estimates of ζ and σ^2 are incrementally updated in a series of steps (Chapter 4 in literature citation 68). Fortunately, programs such as PROC MIXED in SAS make this a straight-forward endeavor. We do not show the formulae for ML or REML estimation of the parameters, because these are standard iterative methods for mixed models (and are too complex to help intuitively). Pages 94–96 in Whitehead (68) succinctly present the details. In the Appendix we show how the moment method is used for parameter estimation.

Expected effect size. We return now to the example, which deals with the effect of Folicur on DON in harvested grain (41). The ML-based estimate of the expected effect size, mean log response ratio, was $\hat{\zeta} = -0.24$, with a standard error of $SE(\hat{\zeta}) = 0.0276$ (Table 1). The estimated among-study variance was $\hat{\sigma}^2 = 0.0365$; we defer discussion of this term until the next subsection. The ratio of the estimated mean effect size and its standard error [$t = \hat{\zeta} / SE(\hat{\zeta})$] is known as a Wald statistic, and this serves as the main basis for hypothesis testing. That is, under the null hypothesis that the expected effect size is 0 (i.e., H_0: $\zeta = 0$), t has a Student's t distribution. There is some disagreement about what degrees of freedom (df) are appropriate for this test because of the hierarchical nature of the model (see discussion starting on page 145 of literature citation 26). A good compromise appears to be $df = K - 2$. Of course, with $K > 30$, the Student's t and the standard normal (Z) variates are very similar, and it is quite common (and correct) to compare the Wald statistic to a standard normal variate (65) for testing the null hypothesis. The achieved significance level (P) for the example was <0.001 (much less than a preassigned critical significance level of $\alpha = 0.05$). Clearly, the evidence is very strong that the expected log ratio is different from 0.

It is important to note that the effect sizes in *individual* studies were often not significantly different from 0 [at a prespecified significance level (α) of 0.05]. In fact, ζ_i was significantly different from 0 in only 30% of the 101 studies (L. V. Madden, *unpublished data*). It is quite possible for the test of the individual effect size to be nonsignificant in every study, but for the test of the expected effect size (across all studies) to be highly significant (4). This is related, in part, to low power of the tests with small number of replications or blocks in individual studies (see Power section, below).

The 95% confidence interval for ζ is easily calculated as

$$\hat{\zeta} \pm t^*_{0.975,df} SE(\hat{\zeta})$$

where $t^*_{0.975,df}$ is the upper 97.5th percentile of the Student's t distribution with df as defined above. With large K (as here), one can use the standard normal 97.5th percentile ($Z^*_{0.975} = 1.98$) instead of t value. The confidence interval extended from –0.299 to –0.189. Because the range does not include 0, this supports (as it must) the conclusion that the expected log ratio was different from 0. However, in other ways, this log-scale result is not very intuitive. One can easily convert the log ratios (the point estimate or the confidence limits) to percent control by simple back-trans-

formation (starting with the original equation 3). For instance, the point estimate of the overall percent control (which is really a median estimate of C) is given by: $\hat{C} = 100(1 - \exp(\hat{\zeta}))$. These estimates are given in Table 1. In this case, one is 95% confident that the interval from 17.2 to 25.8% contains the true (but unknown) expected effect size.

These results based on ML estimation repeat what was given in Paul et al. (41). For comparison purposes, we also give the results for other estimation methods in Table 1. Normally, an investigator should simply use one estimation method. The results based on REML estimation are virtually identical to those found for ML estimation. This is because of the large number of studies. With $K < 30$, it is generally preferable to use REML (68), although some authors prefer ML for a wider range of circumstances (65). When the type of likelihood method makes a difference, the ML estimates of variability [$SE(\hat{\zeta})$, $\hat{\sigma}^2$] will be smaller than the REML estimates because the ML variability estimates are biased downward. The method-of-moments (MM) results were also similar to those found by the likelihood methods, as is usually the case (4,9). However, there is some noticeable discrepancy in the variability estimates [$SE(\hat{\zeta})$, $\hat{\sigma}^2$] between the MM and the likelihood-based estimates. For other analyses, the discrepancy can be either in the positive or negative direction. It should be noted that there is no standard error estimate for the estimated among-study variance with the MM approach.

The comparison of ML, REML, and MM all involve the same model (equation 6). Finally (for now), we show the parameter estimates for the fit of a different type of model, the fixed-effect model (where $\sigma^2 = 0$ or $u_i = 0$, by definition). This is a poor approach for estimating ζ for this data set because the among-study variance was clearly larger than 0 (as discussed in more detail below). A major negative consequence of using the fixed-effect model is that the calculated $SE(\hat{\zeta})$ is much too low, resulting in artificially large Wald statistic (t) and artificially narrow range of the confidence interval (Table 1). In other data sets, the estimate of ζ itself may also be poor (because the estimate depends on the estimate of σ^2; equation 10). In agreement with many authors (4,26,27,65,68), we see little advantage in using a fixed-effect meta-analysis for most investigations. If the estimated σ^2 is actually 0, or close to 0, then the random-effect approach will automatically give about the same results as a fixed-effect approach, without the need for taking a modeling approach that can be severely biased and misleading when there is, in fact, substantial among-study variability.

Heterogeneity. As stated by Higgins et al. (26), "The naïve presentation of inference only on the mean of the random-effects distribution [expected effect size] is highly misleading." Estimation of σ^2 may be just as important, because this parameter quantifies the variability of the true effect sizes among studies. However, many published meta-analyses give little information on the extent of the variability or impact of the variability on the expected effect size (4,25,36).

Hypothesis tests for heterogeneity are possible. For those who use the MM parameter estimates, a chi-square test for σ^2 is routine, based on the calculation of the Q statistic (equation A1 in Appendix). For our example, the estimate of Q was 250.4. Comparing this to quantiles of a chi-square distribution with $K - 1 = 100$ degrees of freedom gave a P value of <0.001, confirming that the among-study variance was larger than 0. For those who prefer likelihood-based parameter estimation (as we do), a likelihood ratio test can be performed (30,65). This involves fitting equation 6 and a version of equation 6 with no u_i term (i.e., no random effect of study), and determining the difference in −2 times the log-likelihood for both model fits (which is the likelihood ratio statistic [LRS]). Under the null hypothesis of $\sigma^2 = 0$, it is traditional to assume that LRS has a chi-square distribution with 1 degree of freedom. For the example, LRS = 54.04, which corresponded to a P value of <0.001. Recent research has shown that the distribution of LRS under the null hypothesis is more complex than that represented by the simple chi-square distribution (66), but we do not expand on this here.

Confidence intervals can be calculated for σ^2, although this is much more complicated than with ζ. This is partly because the distribution of the test statistic can be quite complex (and possibly poorly defined) when the null hypothesis is not true (3,36,67), especially if K is small. No single method is best for this purpose (36,67). The profile likelihood method is a good choice for those who use likelihood-based methods (3,65,67). Until recently, this approach entailed fitting equation 6 multiple times, with different fixed values of σ^2 for each fit (to determine cut-off limits of σ^2

TABLE 1
Estimated expected effect size (log response ratio), among-study variance, and related statistics for the effect of a single application of tebuconazole (Folicur) on the deoxynivalenol toxin content in harvested wheat grain, based on a meta-analysis of 101 studies (41)

Estimation method[a]	Effect size statistics[b]				Among-study variance	Percent control[c]	
	$\hat{\zeta}$ ($SE(\hat{\zeta})$)	t	P	Confidence limits for ζ	$\hat{\sigma}^2$ ($SE(\hat{\sigma}^2)$)	$\hat{C}$	Confidence limits for C
ML	−0.244 (0.0276)	−8.85	<0.001	−0.299 ↔ −0.189	0.0365 (0.0206)	21.6%	17.2% ↔ 25.8%
REML	−0.244 (0.0278)	−8.80	<0.001	−0.299 ↔ −0.189	0.0374 (0.0108)	21.6%	17.2% ↔ 25.8%
MM	−0.245 (0.0285)	−8.60	<0.001	−0.301 ↔ −0.189	0.0407 (–)[d]	21.7%	17.2% ↔ 26.0%
FIXED[e]	−0.223 (0.0163)	−13.70	<0.001	−0.255 ↔ −0.192	–[f]	20.0%	17.5% ↔ 22.5%
ML(MI)[g]	−0.249 (0.0287)	−8.68	<0.001	−0.305 ↔ −0.193	0.0469 (0.0156)	22.0%	17.6% ↔ 26.3%

[a] ML: maximum likelihood; REML: restricted maximum likelihood; MM: method-of-moments (4,12); FIXED: fixed-effect model (equation 6 without u_i term); ML(MI): multiple-imputation with ML estimation for each imputation.
[b] $\hat{\zeta}$: estimated expected effect size (from fit of equation 6, with L_i used for z_i); t: Student's t statistic; P: significance level of t test for the equality of ζ to 0; Confidence limits: limits of a 95% confidence interval for ζ.
[c] Percent control estimate ($\hat{C}$) based on back-transformation of the estimated expected effect size and the confidence-interval limits.
[d] Standard error of estimated among-study variance not calculated with method of moments (MM).
[e] The fixed-effect (or common-effect) model is not appropriate for these data, but results are shown for comparison purposes.
[f] By definition, the among-study variance is 0 for the fixed-effect model.
[g] Statistics for ML(MI) determined from 10 imputations of missing sampling variances followed by ML parameter estimation. Twenty percent of the studies were randomly selected and assigned a missing value for the sampling variance for the purpose of the multiple-imputation (MI) analysis.

based on achieved log-likelihood values for each fit). The calculation of the LRS and also the profile likelihood confidence interval is now straightforward using the GLIMMIX procedure in version 9.2 of SAS. For the FHB example, a 95% confidence limit is 0.020 to 0.063. Note, in general, the confidence interval for a variance is not symmetrical; typically, the width of the interval above the point estimate is larger than the width below the point estimate.

When K is not large (<30), tests of σ^2 are not powerful (4,28) and confidence intervals can be very wide. In contrast, tests of the variance are very powerful at large K. For instance, with a small number of studies, a large estimated σ^2 may be considered nonsignificant, but with a large number of studies, a very small σ^2 may be considered significant. As pointed out elegantly by Higgins and Thompson (25), quantification of the impact of heterogeneity is probably more useful than determining if there is (significant) heterogeneity. Higgins and Thompson developed three interrelated indices of impact, although two of these are most easily expressed (through a function of Q) when the MM estimation method is used.

The R^2 index (which is not a coefficient of determination) is especially straightforward no matter which estimation approach is taken. The index involves the estimated standard error of the estimated ζ when the random-effect model is fitted (equation 6), which we temporarily label as $SE(\hat{\zeta})_{RAN}$, and the standard error when the fixed-effect (common-effect) model is fitted (equation 6 with no u_i term), which we label as $SE(\hat{\zeta})_{FIX}$. The index is written as

$$R^2 = \left(\frac{SE(\hat{\zeta})_{RAN}}{SE(\hat{\zeta})_{FIX}} \right)^2 \quad (13)$$

The square-root of equation 13 represents the inflation in the confidence interval for ζ under a random-effect model compared with a fixed-effect model; the inflation is attributable to the among-study variability of the true effect size (25). If the sampling variance (s_i^2) was the same in all studies ($s_i^2 \equiv \bar{s}^2$), then equation 13 would reduce to

$$\frac{(\hat{\sigma}^2 / K) + (\bar{s}^2 / K)}{(\bar{s}^2 / K)} = (\hat{\sigma}^2 + \bar{s}^2) / \bar{s}^2$$

an expression that clearly shows the inflation due to σ^2. Indices such as R^2 are advantageous because they are scale free, which means their magnitude can be compared across different meta-analyses. When the among-study variation is zero, R^2 equals 1.

It is a good practice for investigators to always present one of the indices of Higgins and Thompson (25) when reporting on a meta-analysis. Based on the ML method of model fitting with the example used here, one finds $R^2 = 2.9$. Based on analogies of R^2 with related indices discussed in Higgins and Thompson (25), the among-study variation has a large impact on the meta-analytical results whenever R^2 is larger than ~1.5. For our example, study variability must clearly be accounted for in the analysis.

Missing values. Missing data in studies considered to be part of a meta-analysis are not uncommon. Missing information on variability (e.g., V or s_i^2) may be the most common problem, because authors often do not report any measure of variation when there is a nonsignificant result (18). The conservative approach to a missing sampling variance (or other measure of variation that can be used to calculate s_i^2) is to omit the data set; however, this eliminates valuable information on the estimated effect size and will most likely increase the bias of the estimated expected effect size. A useful alternative is to impute values for the sampling variance (18), based on relationships between s_i^2 (or V) and other variables in the data set (for the studies without missing values). Along these lines, Paul et al. (41) found a strong linear relationship between the log of V and the log of the mean DON in the study (across all treatments); from a fitted regression model, missing values of s_i^2 could be estimated (equation 4a) based on predicted V, using a single imputation approach.

It turns out that a random-effect meta-analysis is often not overly sensitive to a moderate number of missing sampling variances. We demonstrate this with our example. Using the SURVEYSELECT procedure of SAS, we randomly selected 20% of the studies and assigned them a missing value for s_i^2 (and for V_i, the residual variance for the observations in the ith study). We then utilized the MI procedure in SAS to perform multiple imputations using the MCMC method, based on a multivariate normal distribution for the variables on a log scale (V_i, mean DON and mean field disease severity). The missing data were then filled in 10 times, by drawing random samples of missing value from its estimated distribution, to obtain 10 complete data sets. A meta-analysis (using ML estimation) was then performed on these 10 completed' data sets, and then the results (estimated ζ and σ^2) were combined using the multiple-imputation methods of Rubin (54) to produce composite results for inference.

As shown in Table 1, the results [ML(MI)] from the multiple imputations were very close to those found for the original complete data set. There was a slight trend for a larger among-study variance, which is expected given the extra uncertainty associated with the imputations. Confidence intervals for ζ were little affected by imputation.

The insensitivity of the results to missing sampling variances is likely due to the strong relationship among the observed variables that made up the imputations. With other systems, there may be greater sensitivity to missing sampling variances. An additional cause of the insensitivity may be related to the weight function (w_i) involved in the estimation of ζ (equations 10 to 12), which involves the estimated among-study variance (fixed across all studies) added to the sampling variance (dependent on study). The larger the $\hat{\sigma}^2$ relative to the sampling variances, the less the weights vary with study; under these circumstances, inaccuracy in sampling variances will not affect the weights very much because a reasonable range of predicted s_i^2 values will give about the same w_i values. We expect that when R^2 is large (equation 13), imputation of missing variance data will work well even when there is not an overly strong relationship between variances and other observed variables.

PREDICTION

Prediction intervals. As stated recently by Higgins et al. (26), "Predictions are one of the most important outcomes of a meta-analysis, since the purpose of reviewing research is generally to put knowledge gained into future application. Predictions also offer a convenient format for expressing the full uncertainty around inferences, since both magnitude and consistency of effects may be considered." Unfortunately, formal predictions and the uncertainty in the predictions are seldom reported in published meta-analyses.

When equation 6 is fitted to data from K studies, the estimated expected effect size ($\hat{\zeta}$) is the best predictor of the true effect size in a randomly-selected new study (ζ_{new}). This new study would have to be done in the same way as the studies represented in the analysis. By implication, in the FHB example this prediction also applies to commercial wheat fields treated in the same manner. Note that ζ_{new} is a random variable, not a constant. Although the point estimate of the effect size is important, the variance of ζ_{new} and the width of the so-called prediction interval for ζ_{new} are just as important. It can be shown that the estimated variance of ζ_{new} is $\hat{\sigma}^2 + (SE(\hat{\zeta}))^2$. The second term reflects the uncertainty in the estimate of the expected effect size and the first term reflects the heterogeneity in the true effect sizes. Note that $SE(\hat{\zeta})$ is also a

function of $\hat{\sigma}^2$ (equations 11 and 12); thus, the among-study variance enters the formula twice for the variance of ζ_{new}.

A 95% prediction interval for the effect size of a randomly selected new study is estimated by

$$\hat{\zeta} \pm t^*_{0.975,df}\left(\hat{\sigma}^2 + (SE(\hat{\zeta}))^2\right)^{0.5} \quad (14)$$

where $t^*_{0.975,df}$ is the 97.5th percentile of the Student's t distribution with $df = K - 2$. Of course, when K is large (≥30, for instance), one could use the standard normal variate ($Z^*_{0.975} = 1.96$) instead of the Studen's t value. Based on the ML estimates for the example (Table 1), and use of the standard normal, the 95% prediction interval is

$$-0.244 \pm 1.96\left(0.0365 + 0.0276^2\right)^{0.5} = -0.244 \pm 1.96 \cdot 0.193$$

which is −0.622 to 0.134. Back-transforming these log ratios, one obtains a prediction interval for percent control of −14.3 to 46.3%. One interpretation of this interval is that if a large number of future studies are conducted (or if commercial fields are treated) in basically the same manner as in the studies comprising the original analysis, about 95% of the individual true percents control will be between −14 and 46%. This is in general agreement with the histogram of the individual estimated effect size values shown in Figure 1A and B. The agreement between prediction intervals and raw percentiles of frequency distributions will never be more than approximate, however, because the latter does not account for, among other things, the different sampling variances of the studies.

Even though there is very strong evidence—from the confidence interval—that the expected log ratio is truly less than 0 (meaning that the expected percent control is truly greater than 0), the meta-analysis also gives strong evidence—from the prediction interval—that a wide range of individual results could be experienced, when applying Folicur to susceptible wheat at one growth stage of the crop, under high inoculum density. Clearly, Folicur is not an overly effective fungicide for managing the level of DON in harvested grain (41). Meta-analyses show that other triazole fungicides are much more effective (42).

One can consider any prespecified significance level (α) in determining prediction intervals. For instance, for an 80% prediction interval, one would use $t^*_{1-0.2/2,df}$ (i.e., $t^*_{0.9,df}$) or, for large K (as here), the standard normal variate ($Z^*_{0.9} = 1.28$). When there is substantial heterogeneity, for instance, when $R^2 > 1.5$ (equation 13), $\hat{\sigma}^2$ will be substantially larger than $(SE(\hat{\zeta}))^2$. Thus, the last term in equation 14 [i.e., $\left(\hat{\sigma}^2 + (SE(\hat{\zeta}))^2\right)^{0.5}$] can be approximated by $\hat{\sigma}$. For example, the last term is given fully as 0.193 above, which is very similar to $\hat{\sigma} = \sqrt{0.0365} = 0.191$. Therefore, one can simplify calculations by using $\hat{\sigma}$ in determining prediction intervals. Technically, strict use of the simplification implies that the expected effect size is estimated without uncertainty.

Risk probabilities. The above prediction intervals were two-sided, but one could also calculate one-sided intervals. One can also turn this idea around and specifically estimate the probability that ζ_{new} is below (or above) a critical limit or value (such as a log ratio of 0), rather than determine the limits of an interval for a pre-specified probability (65). We demonstrate this for a situation where we can use the standard normal as an approximation for the Student's t distribution. For simplicity, we use the square-root of the estimated among-study variance (i.e., $\hat{\sigma}$) instead of the more complex function used in equation 14.

Following van Houwelingen et al. (65) and Paul et al. (41,42), the estimated probability that ζ_{new} is less than a constant, ϑ, can be expressed as

$$p_\vartheta = \Pr(\zeta_{new} < \vartheta) = \Phi\left((\vartheta - \hat{\zeta})/\hat{\sigma}\right) \quad (15a)$$

where $\Phi(\bullet)$ is the cumulative distribution function of the standard normal distribution. The probability that ζ_{new} is greater than ϑ is estimated as

$$p_\vartheta = \Pr(\zeta_{new} > \vartheta) = 1 - \Phi\left((\vartheta - \hat{\zeta})/\hat{\sigma}\right) \quad (15b)$$

For the example, the probability that ζ_{new} (specifically here, the log ratio L_{new}) is less than $\vartheta = 0$ (equivalent to the probability that C_{new} is greater than 0) is estimated as $p_0 = \Phi(0.244/0.191) = 0.90$. If one wanted to estimate the probability that percent control of DON is greater than 25%, one needs to estimate the probability that the log ratio is less than $\vartheta = \ln(1-0.25) = -0.288$. This gives $p_{-0.288} = \Phi((-0.288+0.244)/0.191) = 0.41$, much less than the probability that any positive control will occur.

We have found equations 15a and 15b, and their expansions for more complex scenarios, to be extremely valuable in interpreting results from a random-effect meta-analysis (40–42). The excellent article by van Houwelingen (65) gives additional background on use of these equations, although they use a different notation. It should be noted that there is a typing mistake on page 599 of their article (65), where they inadvertently used the estimated among-study variance instead of the estimated among-study standard deviation in their numerical example. The correct result, based on our notation is $p_0 = \Phi\left((0-(-0.742))/\sqrt{0.302}\right) = 0.911$.

POWER IN META-ANALYSIS

An obvious reason to conduct a meta-analysis is to increase our knowledge base for some phenomenon, which will hopefully lead to better (or more appropriate) predictions of a future effect size (such as the effect of some treatment), as well as to better informed decision-making regarding management of a system (8,29). In this context, hypothesis testing can be very important. Individual studies are often conducted to test a null hypothesis (H_0) versus an alternative hypothesis (H_a). Using our notation, we can write these for the ith study as

$$H_0: \ \zeta_i = 0 \ \ versus \ \ H_a: \ \zeta_i \neq 0 \quad (16a)$$

In our example, the null hypothesis would be that the effect of Folicur on DON in wheat grain, expressed here in terms of log ratios (L_i), equaled 0, and the alternative hypothesis would be that the treatment effect was not 0 (i.e., that Folicur did actually affect DON). The effect size could also be the mean difference (D_i) or standardized mean difference (d_i). For correlation studies, the null hypothesis would be that there was no relation between two variables (a correlation coefficient of 0), and the alternative hypothesis would be that there was a relationship (a nonzero correlation coefficient). As discussed above, hypothesis testing extends directly to a meta-analysis. With a random-effect meta-analysis, the null and alternative hypotheses are typically expressed in terms of the true expected effect size (for the population of studies):

$$H_0: \ \zeta = 0 \ \ versus \ \ H_a: \ \zeta \neq 0 \quad (16b)$$

For individual studies or for the population of studies, one can also be more specific and provide a direction for the alternative hypothesis (e.g., that the effect size is less than 0).

In the usual practice of statistical inference, one gathers evidence (such as data from experiments or surveys) that will allow the investigator to reject the null hypothesis in favor of the alternative based on the results of a test. Statistical tests are designed to control, usually at a low level (e.g., $\alpha = 0.05$), the probability of rejecting the null hypothesis when it is in fact true. However, the probability of rejecting the null hypothesis when the

alternative is true, known as power, is often overlooked. It can be argued that statistical power is a key aspect of hypothesis testing, in general (30,38,61). Unfortunately, the power of tests of many interesting hypotheses is low for individual studies (4,27); the low power may be due to a small number of replicates, a high degree of variability (s_i^2), or a small true effect size (ζ_i). As it turns out, however, the power of statistical tests in a meta-analysis can be very high. In fact, high statistical power could be considered a major (or even the primary) reason to conduct a meta-analysis (27). We considered below several aspects of power, using the DON example as a guide.

Power of individual studies. If one assumes that the alternative hypothesis is true (such as a nonzero log response ratio), one can estimate the power for an individual study (study *i*) based on the assumed or specified effect size and its variance (or standard error), pre-determined significance level for the test (α), and residual degrees of freedom (30,38). For demonstration purposes, we consider only the two-sided situation where the null hypothesis is zero and the alternative is nonzero (not specifically less than or greater than 0). We take the approach described in the SAS/STAT POWER User's Guide (especially page 16 of Chapter 1 in citation 55); the analogous approach for a meta-analysis is described explicitly below.

We can get a good sense of the power of the test for an effect of Folicur on DON (as a log ratio) by using the estimated effect sizes (z_i) and their variances for the true values. Based on the results in Figure 1, we do not believe that DON is truly reduced in every study, so that it is not realistic to assume that the alternative hypothesis is true in all cases. However, the exercise serves to demonstrate the range of powers for studies like this (i.e., studies with the kind of estimated effect sizes and sampling variances found in the 101 studies). A more technical description of this exercise is that, assuming the alternative hypothesis is always correct, we are calculating the power that would be achieved if another (future) collection of studies were conducted in the same manner with the same levels of sampling variability and magnitude of effect sizes.

Figure 2A shows a histogram of the estimated powers for the example studies, given the assumptions and qualifications presented above. It is clear that if there truly was an effect of Folicur in every study, that the power to detect the effect would often be quite low. The mean power was only 0.36, and the median was 0.23 (i.e., half the studies had power less then 0.23). In many disciplines, a power of 0.8 is considered to be a desired goal in designing studies. Only 15% of the studies had an estimated power of 0.8 or higher.

Power in a meta-analysis. One can take the same approach to estimate the power of a meta-analysis as done for individual studies. Note, with a random-effect meta-analysis, the hypotheses are different (equation 16b) compared with the hypotheses for individual studies (equation 16a). In particular, the hypotheses in the meta-analysis involve the true *expected* effect size (ζ, across all studies), not the true effect sizes for individual studies (ζ_i). To calculate power for the meta-analysis, one simply assumes that the true expected effect size is nonzero; one does not need to make any assumptions about the true effect sizes in the individual studies (in other words, one can fully expect that the individual true effect sizes be positive, negative, or nil in the different studies).

A so-called noncentrality parameter (assuming 0 for the null hypothesis) for a specified true expected effect size is given by

$$\phi = \frac{\zeta}{\left(\sum\left(\frac{1}{\sigma^2 + s_i^2}\right)\right)^{-1/2}} \quad (17)$$

where the summation is over all K studies, and the denominator is the same as $SE(\hat{\zeta})$ when the estimate of the among-study variance ($\hat{\sigma}^2$) is substituted for the true variance (σ^2) (equations 11 and 12). The value of ϕ indicates how far, on a standardized scale, the true mean is from the hypothesized value under H_0. If the estimate of ζ ($\hat{\zeta}$) and its standard error [$SE(\hat{\zeta})$] are substituted in equation 17, one obtains the Student's t statistic described previously [$t = \hat{\zeta} / SE(\hat{\zeta})$]. For α = 0.05 and a two-sided test, power is estimated as

$$Power = 1 - F_F(F^*_{0.95,1,df}; 1, df, \phi^2) \quad (18)$$

where $F^*_{0.95,1,df}$ is the 95th percentile of the central-F distribution with 1 and df for the numerator and denominator degrees of

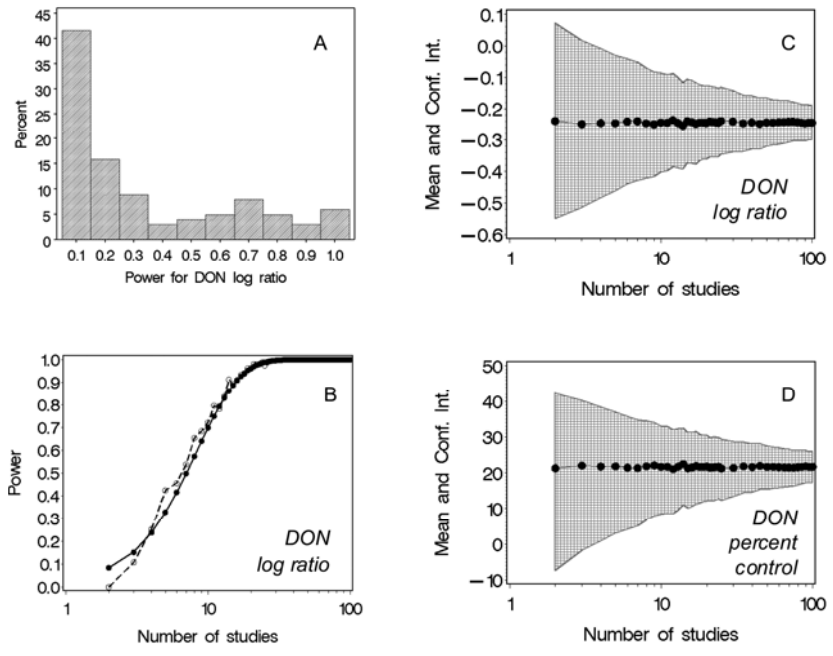

FIGURE 2

Power and precision in the meta-analysis of the effects of tebuconazole on deoxynivalenol (DON) toxin in wheat grain, based on the data given in Paul et al. (41). Results are based on the use of the log response ratio as the effect size (L_i or z_i). **A,** Frequency distribution of the estimated power to detect an effect of the fungicide for the individual studies, with the (unrealistic) assumption that the alternative hypothesis (a nonzero effect of the fungicide) was true for every study. **B,** Estimated power to detect an effect of the fungicide on the expected value (across all studies; ζ), with the (realistic) assumption that the alternative hypothesis (a nonzero effect of fungicide on the expected value) was true overall (not necessarily for individual studies), in relation to number of studies in the meta-analysis (ranging from 2 to 100). Solid symbols are for theoretical calculations, and open symbols are for simulations. **C,** Estimated mean effect size ($\hat{\zeta}$) (points) and 95% confidence interval (hatched area) in relation to number of studies (ranging from 2 to 100). Results based on simulations. **D,** Estimated percent control (relative reduction in DON), and 95% confidence interval, based on a transformation of $\hat{\zeta}$ and the limits of the confidence interval for $\hat{\zeta}$ in **C**.

freedom, respectively; $F_F(\bullet; 1, df, \phi^2)$ is the cumulative noncentral-F distribution with 1 and df for the numerator and denominator degrees of freedom, respectively; and ϕ is as defined in equation 17. More details can be found in Stroup (61) for a more complex model, and in the SAS POWER instructional book (55) for a wide range of situations. For a meta-analysis, we chose here to use $df = K - 1$ for the denominator degrees of freedom, but as discussed above, other choices are possible.

As discussed for power calculations with single studies, we do not know the true status of the expected effect size for this population of studies (nor the variability in the effect sizes). However, we can substitute the ML estimates to get a sense of the power that would be achieved (assuming, as always here, that the alternative hypothesis is correct) for another population of studies with the same distribution of effect sizes. Taking this approach, the power for this meta-analysis was estimated as >0.999. Thus, there is a very high (almost certain) chance that the correct decision would be made regarding a nonzero expected effect size (when there is truly a nonzero expected value). Clearly, the power advantages over individual analyses are obvious. Investigators considering only individual studies would have a difficult time discovering the overall, but modest, benefits of Folicur for reducing DON.

Number of studies. In classical power analysis applied to meta-analysis (4), one can explore the impact of a range of (true) expected effect sizes, variability, and number of studies on power. We consider only the latter here. Given the very high power found above in the example, it is natural to ask: How many studies would be required to achieve a power of 0.8 or 0.9? This can be done in a standard (classical) theoretical way, and also with simulation. The theoretical approach involves adjustments to the denominator of equation 17 for ϕ (the standard error) to reflect the number of studies (K). If the sampling variance was the same for all studies (i.e., $s_i^2 \equiv \bar{s}^2$), then the denominator of equation 17 would be equivalent to $K^{-1/2}\left(\sigma^2 + \bar{s}^2\right)^{1/2}$ or $K^{-1/2}\Psi$, where $\Psi = \left(\sigma^2 + \bar{s}^2\right)^{1/2}$ (a type of standard deviation). The impact of different K values is then easy to obtain at a fixed Ψ (which reflects the sampling and among-study variability, not dependent on number of studies). With variable sampling variances, K cannot be removed from the standard-error equation. However, as an approximation, we can estimate Ψ (roughly) as $\hat{\Psi} = SE(\hat{\zeta})/K^{-1/2}$, and fix this in the power calculations; for the example with $K = 101$, $\hat{\Psi} = 0.0276/101^{-1/2} = 0.277$.

Power was estimated for values of K from 2 to 100 using the noncentrality parameter in equation 18 with $K^{-1/2}\hat{\Psi}$ as the denominator (an approximate standard error). As shown in Figure 2B, estimated power increased rapidly with K, and exceeded 0.80 with 13 or more studies, and exceeded 0.95 with 18 or more studies.

To determine if these results were overly dependent on the approximation for the standard error, a simulation was conducted (30). For our example data set, effect-size data were generated using equation 6, using normal distributions for u_i and ε_i. The ML estimate of ζ was used for the expected effect size and the ML estimate of σ^2 was used for the variance of u_i. The variances of ε_i depended on the study (as required). In separate work, we previously found that the distribution of the reciprocal of the s_i^2 values was well described by a gamma distribution with estimated scale parameter of 32.8 and shape parameter of 1.1 (L. V. Madden, *unpublished data*). We thus generated an s_i^2 value for each simulated study from a gamma distribution, and used this sampling variance to generate a normally distributed ε_i value. One thousand simulations were performed for selected values of K between 2 and 100. For each simulation, equation 6 was fitted to the data, and the expected effect size and its standard error were calculated, as well as the P value of the t test for ζ. The proportion of P values less than or equal to α across all simulations was an estimate of the power (Fig. 2B).

The estimates of power based on simulation were very similar to those found by the more approximate theoretical approach (Fig. 2B) at most values of K. For instance, power was over 0.8 for 13 or more studies, and above 0.95 for 19 or more studies. The greatest difference in the approaches was at K values of 2 and 3, where a meta-analysis should not normally be done. This analysis does show that the simpler theoretical approach was adequate to explore power of a meta-analysis for situations represented by the DON example data set. Analysis of other examples would be necessary before more general recommendations could be given about the best approach to evaluate power.

Precision. If all an investigator wished to do was determine if a treatment was effective (on average), use of 101 studies in a meta-analysis would be overkill, based on the power results for our example. However, a meta-analysis is valuable for many reasons, not just for hypothesis testing. For instance, we can consider the precision of the estimated ζ, as quantified by the width of the estimated 95% confidence interval around $\hat{\zeta}$ ($\pm t^*_{0.975,df} SE(\hat{\zeta})$). Using the above simulated data, we calculated the point estimate and 95% confidence-interval limits for each of the simulated data sets for each chosen K, and then determined the means for these at each K (across all simulations). As shown in Figure 2C, the (mean) estimate of ζ depended very little on K, with an estimate around –0.244 (the prespecified value), even with two or three studies. The back-transformed mean and the confidence-interval limits are given in Figure 2D; the intervals are nonsymmetrical (wider below the mean, especially at small K) because of the nonlinear transformation.

Graphs such as Figure 2C and D can be used in different ways for determining the influence of study number on precision. For instance, if one wanted to determine the number of studies that would result in a lower bound of 15% on the confidence interval for percent control (when expected percent control was 21%), one would need 45 studies (with the among-study and sampling variances considered here). Or, if one wanted to determine the number of studies that gave a width of the confidence for expected percent control of no more than 10% (e.g., from 16 to 26%), one would need a K of 75.

The fallacy of counting individual *P* values. Meta-analysis is generally considered to be an alternative to so-called narrative review for research synthesis. In a classical narrative review, "An expert in a given field would read the studies that addressed a question, summarize the findings, and then arrive at a conclusion…" (4). Meta-analysts have written extensively about problems with such narrative reviews; we recommend interested readers start by consulting Borenstein et al. (4). We deal only with one aspect of this approach, which is related to statistical power.

A typical narrative or qualitative summary of a topic, at least at the simplest level, is to review the published studies and count the number of significant results (studies where the test of a treatment effect gives $P \leq 0.05$). For instance, if one wanted to know if DON in wheat grain was related to head blight symptoms on spikes in the field, one could count the number of studies where the investigator found a significant relationship between symptoms and DON. If interested in the effect of a biocontrol agent, one could count the number of studies where agent "X" significantly reduced disease severity compared with a control. One would conclude that there truly was a relationship, or there truly was a treatment effect, if at least half the studies had significant results. Using our DON example, with K = 101 studies under consideration, P would have to be less than or equal to 0.05 (α) in a least 51 studies to declare that this fungicide had any true effect. This general approach is known sometimes as vote counting, and

it is easy to show that it is a fatally flawed approach to test hypotheses (23,27).

We demonstrate this with an instructional example where the alternative hypothesis (equation 16a) is true in each of the K studies. In other words, the true treatment effect (the effect size of interest) is nonzero in every study. We also suppose that power to test for the treatment effect is 0.40 in every study. This is not a high value, but it is not uncommon; it is actually higher than the mean individual-study power in the DON example (Fig. 2A) (with the assumption that effect size was truly nonzero in every study). With a large number of studies ($K > 100$) under these circumstances, about 40% of the studies will have a significant result (27). Thus, the criterion for overall true effect in the vote-counting approach would not be met, and one would falsely conclude that treatment was *not* effective, even though it was truly effective in every study!

Whenever the power of tests in individual studies is not high, the narrative-review approach is problematic, and bound to fail under many circumstances. In other words, the power of the vote-counting method is low, and often (much) lower than the power of the individual studies on which it is based. In fact, as shown by Hunter and Schmidt (27) and Hedges and Olkin (23), the power of the vote-counting method can tend towards 0 as the number of studies increases. As stated by Borenstein et al. (4), "vote counting is not only misleading, it tends to be more misleading as the amount of evidence (the number of studies) increases!"

There is ample evidence that naïve vote-counting approaches can lead to disastrous conclusions under many circumstances (23,27). It should be pointed out that there are valid statistical methods for combining P values from independent studies, going back to the work of Fisher (17) and Tippett (63). Chapter 36 in Borenstein et al. (4) and, in more detail, Chapter 3 in Hartung et al. (21), describes these approaches. These methods are usually considered as special types of procedures under the general approach of meta-analysis. These specialized methods are not as powerful or useful as the methods described here, but can be utilized when the actual effect size estimates are not available.

PUBLICATION BIAS

General issues. If statistical power is one of the major advantages of meta-analysis, then publication bias is one of the major potential disadvantages (4,26). Most meta-analyses make the tacit assumption that the studies under review are a random sample from a hypothetical population of possible studies, or, more realistically, that the effects in each study comprise a random sample from an imaginary population of effects (26). From a Bayesian perspective, this is equivalent to assuming that the study effects meet the criterion of exchangeability. Despite common misconceptions, there is no requirement that results from all studies be included; the principle of randomness, applied to random samples of effects (out of a larger population of effects), assures that statistical inference is valid. Of course, the assumption of a random sample of effects is rather strong, and is unlikely to be fully met for any meta-analysis.

Study results are published or made available for review for many reasons. Other study results are discarded or stored away in a file cabinet for other reasons. There is a considerable literature on this topic, going back at least to Rosenthal (50). Borenstein (4) and Lipsey and Wilson (29) are good places to start reading about the issue, and many more details are given in Rothstein et al. (53). It is likely that larger studies, or studies with significant results, or studies with small standard errors, or studies with novel results will be published, compared with smaller studies, or those with high variability, or those with nonsignificant or nonnovel results. van Houwelingen (64) considered the selective reporting of study results to be the nightmare of meta-analysis.

It is always a good practice to use quality and relevance criteria (unrelated to the actual estimated effect sizes) when selecting studies for a meta-analysis (Boxes 1 and 2) (9,29). However, if inclusion of studies in the data set for analysis depends on the realized effect sizes, then the meta-analytical results will be biased. A biased result means that the expected value of $\hat{\zeta}$ does not equal the true value, ζ; that is, $E(\hat{\zeta}) \neq \zeta$. This is the publication-bias problem. In general, the direction of the bias will be to favor the alternative hypothesis. Bias is not a real concern for the DON example data set because the national wheat scab initiative encouraged the reporting of all results (not necessarily in journal articles).

Some solutions. By far, the most common approach to the problem is to ignore the selection bias of studies (4,9). This means that the population of studies under consideration is actually a restrictive subset of the larger population of studies. Among other things, this limits the scope of inference about the effect size and can lead to higher type I errors (too frequently rejecting H_0 when H_0 is true). More informative approaches involve the use of weights for the observed effect sizes (in addition to the nominal weights based on variability, as given in equation 11), based on various assumptions regarding the study-selection (publication) process, followed by a sensitivity analysis to determine the implications of the hypothesized selection process on the calculated statistics (53).

However, based on the available studies, it is impossible to determine the selection process that makes certain studies avail-

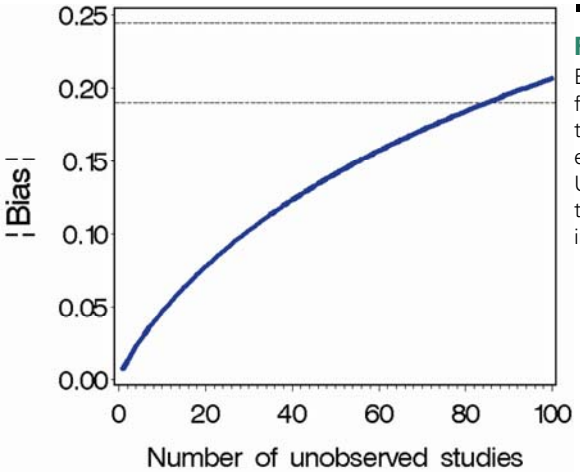

FIGURE 3

Estimated maximum possible bias of the mean effect size ($\hat{\zeta}$) when there are from 1 to 100 studies not used in the meta-analysis. Results are based on the theoretical work of Copas and Jackson (10), utilizing the data published in Paul et al. (41) for the effects of tebuconazole on deoxynivalenol toxin in wheat grain. Upper horizontal line is the absolute value of the estimated expected effect size for the data set, and the lower horizontal line is the limit of the 95% confidence interval for the expected effect size

able for analysis and others unavailable. Studies get published, or become available in reports and so on, for numerous reasons. The size of the observed effect size may be one reason, but not the only reason. A very interesting new alternative is to determine the upper bound on the bias (in absolute units) for any number of unpublished studies. Copas and Jackson (10) show that this worst-case bias (for any selection mechanism) is straightforward to calculate based on the statistics from the analysis of the available studies.

One assumes that there are K_0 unpublished (or unavailable) studies in addition to the K published studies. Thus, the probability of selection (i.e., the probability that study results become available) is $K/(K + K_0)$. Of course, K_0 is not known. The approach of Copas and Jackson (10) is not to predict K_0, rather, their approach is to determine the maximum bias $[\,|E(\hat{\zeta})-\zeta|\,]$ or bound for the bias for a range of possible K_0 values. The formula to estimate this bound for normal distributions is given as

$$|bias| \le \left(\frac{K+K_0}{K}\right)\varphi\left\{\Phi^{-1}\left(\frac{K}{K+K_0}\right)\right\}\frac{\sum(s_i^2+\hat{\sigma}^2)^{-1/2}}{\sum(s_i^2+\hat{\sigma}^2)^{-1}} \qquad (19)$$

where $\varphi\{\bullet\}$ is the probability-density function for the standard normal distribution, and $\Phi^{-1}(\bullet)$ is the inverse cumulative density function for the standard normal. The summations in the last term are over the K observed studies. It is important to point out that the bias (in absolute value) will be *less than* or equal to the right side of equation 19, and possibly much less; thus, the right side of this equation gives the upper bound (in absolute value).

The bound for the bias for the DON example (Table 1, Fig. 1) was calculated using equation 19 (utilizing the ML estimates of ζ and σ^2), with K_0 values from 1 to 100. Note that if $K_0 = 20$, for example, this means that there were, in fact, a total of 121 studies ($K + K_0$). As required, the bias bound increased with K_0, but the increase in the bias slowed as K_0 became larger (Fig. 3). Also shown on the graph is the absolute value of the estimated ζ and the absolute value of the confidence limit that was closest to 0 (Table 1). There would be a bias problem in the meta-analysis if the calculated $|bias|$ bound was close to or above the confidence limit, or, especially, close to or above the point estimate of ζ at small K_0. With the example, the bias bound only crosses the confidence limit at $K_0 \approx 85$ additional studies, and does not reach the estimate of ζ even with $K_0 = 100$ additional studies.

We clarify a few points by considering the situation with $K_0 = 20$ unpublished studies. The bias bound is 0.077. This means that $\hat{\zeta}$ could be as high as –0.167 (= –0.244 + 0.077) or as low as –0.321 (= –0.244 – 0.077), instead of equal to –0.244 (Table 1), if there were actually 121 and not 101 studies. Back-transforming to percent control, this means that expected C could be decreased to 15% or increased to 27%, instead of being equal to 22% (Table 1). Given the distance between the bias bound and $\hat{\zeta}$ for reasonable values of K_0, we consider publication bias to be of little concern for this example.

EXPANSIONS

The among-study variance reflects, among other things, the diversity of study conditions or characteristics in the analyzed data set (26,36,58). One may be able to simultaneously reduce the among-study variability term and increase our understanding of the phenomenon being studied by incorporating so-called moderator variables into equation 6 (4,29,65,67,68). A moderator variable is a characteristic of a study that is included in the data set for a meta-analysis (4). Examples include categorical variables such as cultivar and continuous variables such as environment. In many cases, it may be more interesting to determine which moderator variables are affecting effect sizes than it is to simply estimate the expected effect size (40–44).

The individual studies in a meta-analysis may consist of several treatments, such as different fungicides or cultivars. If one is actually interested in several effect sizes, one could actually conduct several separate meta-analyses (one for each treatment effect), as suggested by Borenstein (4). However, it may be more efficient to analyze all the effects simultaneously using multivariate mixed models or other expansions of equation 6 (1,2,31, 32,40,42,46–48,65,68). These expanded approaches are especially useful if one explicitly wants to compare different treatments with each other, and not just with the control. The multivariate approaches may also result in lower standard errors of estimated expected effect sizes because they account for the correlations of the treatment effects (47,48).

CONCLUSIONS

Meta-analysis has become quite common in many fields, and has even become the standard approach to research synthesis in some disciplines (4,26,27,62). As stated by Sutton and Higgins (62), "as the need for … research and clinical practice to be based on the totality of relevant and sound evidence has been increasingly recognized, the impact of meta-analysis has grown enormously." There are now so many meta-analyses performed in some fields (especially in medicine) that there are now meta-analyses of the meta-analyses (meta-meta-analyses) (13,60). Plant pathologists are now slowly utilizing the methodologies (literature review in literature citation 37), and considerable progress can certainly be made in the coming years in utilizing this approach for synthesizing the available information on a particular topic.

This article summarized the classical methods of meta-analysis, focusing on fitting the random-effect meta-analytical model with likelihood methods for the purpose of inference and prediction. Among other things, we showed how meta-analysis can lead to a high level of statistical power and provided methods to evaluate the potential impact of publication bias on the results. For those wanting a more thorough explanation of the practice of meta-analysis, Borenstein et al. (4) and Lipsey and Wilson (29) are excellent starting points.

APPENDIX

Moment-based estimation. A statistical problem in meta-analysis is that the estimate of ζ depends on σ^2, and the estimate of σ^2 depends on ζ. The moment-based approach gets around this problem in a clever manner (68). It is first assumed that $\sigma^2 = 0$ in equation 11 (i.e., a fixed-effect model is fitted). This makes the weight (equation 11) $w_i = s_i^{-2}$; using this weight, the estimate of ζ is direct with equation 10 (but not a very good estimate if there really is heterogeneity in the true effect sizes). We call this estimate $\hat{\zeta}_{FIX}$. A test of nonzero σ^2 is obtained from the statistic

$$Q = \sum s_i^{-2}\left(z_i - \hat{\zeta}_{FIX}\right)^2 \qquad (A1)$$

Under the null hypothesis that $\sigma^2 = 0$, Q has a chi-squared distribution with $df = K - 1$. Large values of Q (relative to df) is an indication that $\sigma^2 > 0$. It is common to use this test (3,4, 24,28,29,67), but it is well known to lack power when K is small (i.e., even when $\sigma^2 > 0$, the probability of finding a significant result is low when K is not large).

Based on the theoretical expectation of Q for any value of σ^2, DerSimonian and Laird (12) showed how to use Q calculated from equation A1 and the sampling variances for individual studies to directly estimate σ^2. This approach is not part of general-purpose statistical programs with mixed models, but specialized meta-analysis programs typically make this calcu-

lation (4,29). Page 4323 of Mittlböck and Heinzl (36) and page 73 of Borenstein et al. (4) succinctly show how to perform the calculations. The moment estimate is given simply as

$$\hat{\sigma}^2 = \frac{Q-(K-1)}{c} \quad \text{(A2)}$$

where c equals

$$c = \sum s_i^{-2} - \frac{\sum s_i^{-4}}{\sum s_i^{-2}}$$

One can then substitute the moment-based variance estimate into equation 11 in order to calculate the random-effects estimate of the expected effect size. Whenever the calculated Q is less than K–1, one uses 0 for the estimated among-study variance. More recently, DerSimonian and Kacker (11) presented some variations of the method of moment approach that can be used in meta-analysis.

ACKNOWLEDGMENTS

Salaries and research support were provided by state and federal funds to the Ohio Agricultural Research and Development Center. This investigation is based upon work supported, in part, by the U.S. Department of Agriculture (USDA) Agreement No. 59-0790-4-112 and by USDA-CSREES Special Grant 2008-34493-19444. This is a cooperative project with the U.S. Wheat & Barley Scab Initiative (USWBSI). Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the view of the U.S. Department of Agriculture.

LITERATURE CITED

1. Arends, L. R., Voko, Z., and Stijnen, T. 2003. Combining multiple outcome measures in a meta-analysis: An application. Stat. Med. 22:1335-1353.
2. Bagos, P. G. 2008. A unification of multivariate methods for meta-analysis of genetic association studies. Stat. Appl. Genet. Mol. Biol. 7(1):article 31. DOI:10.2202/1544-6115.1408.
3. Biggerstaff, B. J., and Jackson, D. 2008. The exact distribution of Cochran's heterogeneity statistic in one-way random effects meta-analysis. Stat. Med. 27:6093-6110.
4. Borenstein, M., Hedges, L. V., Higgins, J. P. T., and Rothstein, H. R. 2009. Introduction to Meta-Analysis. John Wiley & Sons, Chichester, U.K.
5. Chalmers, I., Hedges, L. V., and Cooper, H. 2002. A brief history of research synthesis. Evaluation and the Health Professions 25:12-37.
6. Chalmers, T. C., and Lau, J. 1993. Meta-analytic stimulus for changes in clinical trials. Stat. Meth. Med. Res. 2:161-172.
7. Cochran, W. G. 1954. The combination of estimates from different experiments. Biometrics 10:101-129.
8. Cooper, H., and Hedges, L. V. 2009. Research synthesis as a scientific process. Pages 3-16 in: The Handbook of Research Synthesis and Meta-Analysis, 2nd ed. H. Cooper, L. V. Hedges, and J. C. Valentine, eds. Sage Publications, Thousand Oaks, CA.
9. Cooper, H., Hedges, L. V., and Valentine, J. C., editors. 2009. The Handbook of Research Synthesis and Meta-Analysis, 2nd ed. Sage Publications, Thousand Oaks, CA.
10. Copas, J., and Jackson, D. 2004. A bound for publication bias based on the fraction of unpublished studies. Biometrics 60:146-153.
11. DerSimonian, R., and Kacker, R. 2007. Random-effects model for meta-analysis of clinical trials: An update. Contemp. Clin. Trials 28:105-111.
12. DerSimonian, R., and Laird, N. 1986. Meta-analysis in clinical trials. Contolled Clin. Trials 7:177-188.
13. Engels, E. A., Schmid, C. H., Terrin, N., Olkin, I., and Lau, J. 2000. Heterogeneity and statistical significance in meta-analysis: An empirical study of 125 meta-analyses. Stat. Med. 19:1707-1728.
14. Eysenck, H. J. 1952. The effects of psychotherapy: An evaluation. J. Consult. Psychol. 16:319-324.
15. Eysenck, H. J. 1978. An exercise in mega-silliness. Am. Psychol. 33:517-519.
16. Feinstein, A. R. 1995. Meta-analysis: Statistical alchemy for the 21st century. J. Clin. Epidemiol. 48:71-79.
17. Fisher, R. A. 1932. Statistical Methods for Research Workers, 4th ed. Oliver & Boyd, London.
18. Furukawa, T. A., Barbui, C., Cipriani, A., Brambilla, P., and Watanabe, N. 2006. Imputing missing standard deviations in meta-analyses can provide accurate results. J. Clin. Epidemiol. 59:7-10.
19. Glass, G. V. 1976. Primary, secondary, and meta-analysis. Educational Res. 5:3-8.
20. Glass, G. V., McGaw, B., and Smith, M. L. 1981. Meta-Analysis in Social Research. Sage Publications, Beverly Hills, CA.
21. Hartung, J., Knapp, G., and Sinha, B. K. 2008. Statistical Meta-Analysis with Applications. John Wiley & Sons, Hoboken, NJ.
22. Hedges, L. V., Gurevitch, J., and Curtis, P. S. 1999. The meta-analysis of response ratios in experimental ecology. Ecology 80:1150-1156.
23. Hedges, L. V., and Olkin, I. 1980. Vote-Counting methods in research synthesis. Psychol. Bull. 88:359-369.
24. Hedges, L. V., and Olkin, I. 1985. Statistical Methods for Meta-Analysis. Academic Press, Orlando, FL.
25. Higgins, J. P. T., and Thompson, S. G. 2002. Quantifying heterogeneity in a meta-analysis. Stat. Med. 21:1539-1558.
26. Higgins, J. P. T., Thompson, S. G., and Spiegelhalter, D. J. 2009. A re-evaluation of random-effects meta-analysis. J. R. Stat. Soc. A 172:137-159.
27. Hunter, J. E., and Schmidt, F. L. 2004. Methods of Meta-Analysis: Correcting Error and Bias in Research Findings. 2nd ed. Sage Publications Inc., Thousand Oaks, CA.
28. Jackson, D. 2006. The power of the standard test for the presence of heterogeneity in meta-analysis. Stat. Med. 25:2688-2699.
29. Lipsey, M. W., and Wilson, D. B. 2001. Practical Meta-Analysis. Sage Publications Inc., Thousand Oaks, CA.
30. Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., and Schabenberger, O. 2006. SAS for Mixed Models, 2nd ed. SAS Institute, Cary, NC.
31. Lu, G., and Ades, A. E. 2004. Combination of direct and indirect evidence in mixed treatment comparisons. Stat. Med. 23:3105-3124.
32. Lu, G., and Ades, A. E. 2006. Assessing evidence consistency in mixed treatment comparisons. J. Am. Stat. Assoc. 101:447-459.
33. Madden, L. V., Hughes, G., and van den Bosch, F. 2007. The Study of Plant Disease Epidemics. American Phytopathological Society, St. Paul, MN.
34. Madden, L. V., and Paul, P. A. 2009. Assessing heterogeneity in the relationship between wheat yield and Fusarium head blight intensity using random-coefficient mixed models. Phytopathology 99:850-860.
35. Mila, A. L., and Ngugi, H. K. 2011. A Bayesian approach to meta-analysis of plant pathology studies. Phytopathology 101:42-51.
36. Mittlböck, M., and Heinzl, H. 2006. A simulation study comparing properties of heterogeneity measures in meta-analyses. Stat. Med. 25:4321-4333.
37. Ngugi, H. K., Esker, P. D., and Scherm, H. 2011. Meta-analysis to determine the effects of plant disease management measures: Review and case studies on soybean and apple. Phytopathology 101:31-41.
38. O'Brien, R. G., and Castelloe, J. 2007. Sample-size analysis for traditional hypothesis testing: Concepts and issues. Pages 237-271 in: Pharmaceutical Statistics Using SAS: A Practical Guide. A. Dmitrienko, C. Chuang-Stein, and R. D'Agostino, eds. SAS Institute, Cary, NC.
39. Ojiambo, P. S., and Scherm, H. 2006. Biological and application-oriented factors influencing plant disease suppression by biological control: A meta-analytical review. Phytopathology 96:1168-1174.
40. Paul, P. A., Hershman, D. E., McMullen, M. P., and Madden, L. V. 2010. Meta-analysis of the effects of triazole-based fungicides on wheat yield and test weight as influenced by Fusarium head blight intensity. Phytopathology 100:160-171.
41. Paul, P. A., Lipps, P. E., Hershman, D. E., McMullen, M. P., Draper, M. A., and Madden, L. V. 2007. A quantitative review of tebuconazole effect on Fusarium head blight and deoxynivalenol content in wheat. Phytopathology 97:211-220.
42. Paul, P. A., Lipps, P. E., Hershman, D. E., McMullen, M. P., Draper, M. A., and Madden, L. V. 2008. Efficacy of triazole-based fungicides for Fusarium head blight and deoxynivalenol control in wheat: A multivariate meta-analysis. Phytopathology 98:999-1011.
43. Paul, P. A., Lipps, P. E., and Madden, L. V. 2005. Relationship between visual estimates of Fusarium head blight intensity and deoxynivalenol accumulation in harvested wheat grain: A meta-analysis. Phytopathology 95:1225-1236.
44. Paul, P. A., Lipps, P. E., and Madden, L. V. 2006. Meta-analysis of regression coefficients for the relationship between Fusarium head blight and deoxynivalenol content of wheat. Phytopathology 96:951-961.
45. Pearson, K. 1904. Report on certain enteric fever inoculation statistics. Brit. Med. J. 2:1243-1246.

46. Raudenbush, S. W., Becker, B. J., and Kalaian, H. 1988. Modeling multivariate effect sizes. Psychol. Bull. 103:111-120.
47. Riley, R. D., Abrams, K. R., Lambert, P. C., Sutton, A. J., and Thompson, J. R. 2007. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. Stat. Med. 26:78-97.
48. Riley, R. D., Abrams, K. R., Sutton, A. J., Lambert, P. C., and Thompson, J. R. 2007. Bivariate random-effects meta-analysis and the estimation of between-study correlation. BMC Med. Res. Methodol. 7:3. DOI:10.1186/1471-2288-7-3.
49. Rosenberg, M. S., Garrett, K. A., Su, Z., and Bowden, R. L. 2004. Meta-analysis in plant pathology: Synthesizing research results. Phytopathology 94:1013-1017.
50. Rosenthal, R. 1979. The "file-drawer problem" and tolerance for null results. Psychol. Bull. 86:638-641.
51. Rosenthal, R. 1984. Meta-Analytic Procedures for Social Research. 1st ed. Sage Publication, Beverly Hills, CA.
52. Rosenthal, R., and Rubin, D. B. 1978. Interpersonal expectancy effects: The first 345 studies. Behavioral Brain Sci. 3:377-386.
53. Rothstein, H. R., Sutton, A. J., and Borenstein, M. E. 2005. Publication Bias in Meta-Analysis—Prevention, Assessment and Adjustments. John Wiley & Sons, Chichester, UK.
54. Rubin, D. B. 1996. Multiple imputation after 18+ years. J. Am. Stat. Assoc. 91:473-489.
55. SAS/STAT 9.2 User's Guide: Power Analysis (Book Excerpt). 2009. SAS Publishing, SAS Institute, Inc., Cary, NC.
56. Schmidt, F. L., and Hunter, J. E. 1977. Development of a general solution to the problem of validity generalization. J. Appl. Psychol. 62:529-540.
57. Shah, D. A., and Dillard, H. R. 2006. Yield loss in sweet corn caused by *Puccinia sorghi*: A meta-analysis. Plant Dis. 90:1413-1418.
58. Sidik, K., and Jonkman, J. N. 2007. A comparison of heterogeneity variance estimators in combining results of studies. Stat. Med. 26:1964-1981.
59. Smith, M. L., and Glass, G. V. 1977. Meta-analysis of psychotherapy outcome studies. Am. Psychol. 32:752-760.
60. Sterne, J. A. C., Jüni, P., Schulz, K. F., Altman, D. G., Bartlett, C., and Egger, M. 2002. Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. Stat. Med. 21:1513-1524.
61. Stroup, W. W. 2002. Power analysis based on spatial effects models: A tool for comparing design and analysis strategies in the presence of spatial variability. J. Agric. Biol. Environ. Stat. 7:491-511.
62. Sutton, A. J., and Higgins, J. P. T. 2008. Recent developments in meta-analysis. Stat. Med. 27:625-650.
63. Tippett, L. H. C. 1931. The Methods of Statistics. Williams & Norgate, London.
64. van Houwelingen, H. C. 1997. The future of biostatistics: Expecting the unexpected. Stat. Med. 16:2773-2784.
65. van Houwelingen, H. C., Arends, L. R., and Stijnen, T. 2002. Advanced methods in meta-analysis: Multivariate approach and meta-regression. Stat. Med. 21:589-624.
66. Verbeke, G., and Molenberghs, G. 1997. Linear Mixed Models in Practice: A SAS-Oriented Approach. Springer, New York.
67. Viechtbauer, W. 2007. Confidence intervals for the amount of heterogeneity in meta-analysis. Stat. Med. 26:37-52.
68. Whitehead, A. 2002. Meta-Analysis of Controlled Clinical Trials. John Wiley & Sons, West Sussex, England.
69. Whitehead, A., and Whitehead, J. 1991. A general parametric approach to the meta-analysis of randomized clinical trials. Stat. Med. 10:1665-1677.
70. Yates, F., and Cochran, W. G. 1938. The analysis of groups of experiments. J. Agric. Sci. 28:556-580.