Plant Disease • 2024 • 108:1937-1945 • https://doi.org/10.1094/PDIS-11-23-2519-SR

Special Report

MSE FINDR: A Shiny R Application to Estimate Mean Square Error Using Treatment Means and Post Hoc Test Results

Vinicius C. Garnica,[1] Denis A. Shah,[2] Paul D. Esker,[3] and Peter S. Ojiambo[1,†]

[1] Center for Integrated Fungal Research, Department of Entomology and Plant Pathology, North Carolina State University, Raleigh, NC 27695
[2] Department of Plant Pathology, Kansas State University, Manhattan, KS 66506
[3] Department of Plant Pathology and Environmental Microbiology, The Pennsylvania State University, University Park, PA 16802

Abstract

Research synthesis methods such as meta-analysis rely primarily on appropriate summary statistics (i.e., means and variance) of a response of interest for implementation to draw general conclusions from a body of research. A commonly encountered problem arises when a measure of variability of a response across a study is not explicitly provided in the summary statistics of primary studies. Typically, these otherwise credible studies are omitted in research synthesis, leading to potential small-study effects and loss of statistical power. We present MSE FINDR, a user-friendly Shiny R application for estimating the mean square error (i.e., within-study residual variance, $\widehat{\sigma}^2$) for continuous outcomes from analysis of variance (ANOVA)-type studies, with specific experimental designs and treatment structures (Latin square, completely randomized, randomized complete block, two-way factorial, and split-plot designs). MSE FINDR accomplishes this by using commonly reported information on treatment means, significance level (α), number of replicates, and post hoc mean separation tests (Fisher's least significant difference [LSD], Tukey's honest significant difference [HSD], Bonferroni, Šidák, and Scheffé). Users upload a CSV file containing the relevant information reported in the study and specify the experimental design and post hoc test that was applied in the analysis of the underlying data. MSE FINDR then proceeds to recover $\widehat{\sigma}^2$ based on user-provided study information. The recovered within-study variance can be downloaded and exported as a CSV file. Simulations of trials with a variable number of treatments and treatment effects showed that the MSE FINDR-recovered $\widehat{\sigma}^2$ was an accurate predictor of the actual ANOVA $\widehat{\sigma}^2$ for one-way experimental designs when summary statistics (i.e., means, variance, and post hoc results) were available for the single factor. Similarly, $\widehat{\sigma}^2$ recovered by the application accurately predicted the actual $\widehat{\sigma}^2$ for two-way experimental designs when summary statistics were available for both factors and the sub-plot factor in split-plot designs, irrespective of the post hoc mean separation test. The MSE FINDR Shiny application, documentation, and an accompanying tutorial are hosted at https://garnica.shinyapps.io/MSE_FindR/ and https://github.com/vcgarnica/MSE_FindR/. With this tool, researchers can now easily estimate the within-study variance absent in published reports that nonetheless provide appropriate summary statistics, thus enabling the inclusion of such studies that would have otherwise been excluded in meta-analyses involving estimates of effect sizes based on a continuous response.

Keywords: meta-analysis, missing summary statistics, R Shiny, residual variance recovery, unreported variability

Scientific progress depends on our ability to reach broad generalizations from knowledge generated across a related body of research using results that may reside in many disparate sources (Hunter and Schmidt 2004; Madden and Paul 2011; Scheiner and Gurevitch 2001). Meta-analysis, a statistical synthesis methodology, has been fundamental in combining the results of separate, independent studies to reach an overall understanding of a research problem (Borenstein et al. 2021; Gurevitch et al. 2018; Lipsey and Wilson 2001). When studies of interest have met the criteria for inclusion in a meta-analysis following a systematic review, they are statistically combined to estimate the magnitude and direction of the overall effect size (Borenstein et al. 2021). In the present study, we focus on meta-analyses that utilize means and variances to estimate effect sizes based on a continuous response in an analysis of variance (ANOVA) setting.

Meta-analysis relies on the availability of either the actual individual-study raw data or study-level summary metrics such as the sample size, treatment mean, and variability of the data, i.e., variance or standard deviation (SD). Variance metrics are required because effect sizes in moment- and likelihood-based meta-analytic methods are commonly weighted by the inverse of their variances, whereby studies with low residual variances are given more weight than those with larger residual variances (Borenstein et al. 2021). However, plant disease (Madden and Paul 2011) and ecological (Gurevitch and Hedges 1999; Scheiner and Gurevitch 2001) research data are not always adequately reported in the literature, with many studies reporting mean effects but not any associated variability metrics. For example, Ngugi et al. (2011) observed that >97% of the reports on product efficacy published in *Fungicide & Nematicide Tests*, *Biological & Cultural Tests* and *Plant Disease Management Reports* did not directly provide the pooled sample variance or a related

†Corresponding author: P. S. Ojiambo; pojiamb@ncsu.edu

Data availability: MSE FINDR Shiny app code, tutorial, and example datasets are hosted at https://github.com/vcgarnica/MSE_FindR. The application is currently hosted at https://garnica.shinyapps.io/MSE_FindR/.

Funding: The study was funded in part by a grant from the USDA-AFRI program (grant number 2020-67013-31920), hatch funds from the North Carolina Agricultural Experiment Station for project NC02950, and the USDA National Institute of Food and Federal Appropriations under project PEN04836 and accession number 7005075.

e-Xtra: Supplementary material is available online.

The author(s) declare no conflict of interest.

Accepted for publication 1 February 2024.

statistic. Additionally, the underlying raw data that could be used to generate summary statistics for these reported studies are rarely available (Sparks et al. 2023). Variability metrics can be calculated algebraically from other parametric summary statistics such as *t* tests, *F*-value or *P* value, contained in the primary report (Batson and Burton 2016; Thiessen Philbrook et al. 2007). These calculations, however, assume the original data follow normality assumptions (Lipsey and Wilson 2001). Equations used in these calculations are also limited by the numerical precision of the reported summary statistics and become increasingly unreliable when too few digits are reported due to rounding (Acutis et al. 2022; Lajeunesse 2013). Consequently, studies in which the underlying raw data are not available or that lack basic summary statistics are usually omitted in projects on quantitative synthesis of research results. This can lead to imprecision and biases in meta-analysis results (Borenstein et al. 2009; Weir et al. 2018). As a result, there has been increasing interest in developing methods that will allow the inclusion of studies lacking the required summary statistics into the meta-analysis (Acutis et al. 2022; Adams et al. 1997; Chowdhry et al. 2016; Ngugi et al. 2011). The methods developed in these latter studies are increasingly being used by applied researchers in their research synthesis studies (Fohrafellner et al. 2023; Kong et al. 2023; Tadiello et al. 2023).

One method suggested to synthesize results from studies with no variability metrics is to conduct an unweighted meta-analysis using the log response ratio as the effect size (Adams et al. 1997). This approach requires only the mean but not the SD to compute bootstrapped confidence intervals (Gurevitch and Hedges 1999; Scheiner and Gurevitch 2001). However, the approach can be viewed as a very crude surrogate for a traditional meta-analysis and should be used when no other technique is available for research synthesis (Lajeunesse 2013). Imputation methods have also been proposed for estimating missing sample variances (Chowdhry et al. 2016; Furukawa et al. 2006). However, they assume that the individual-study information is missing at random and not because of reporting biases. In addition, the assumption that studies are missing at random is untestable (Higgins et al. 2008; Lajeunesse 2013). Imputation techniques are also appropriate only when a minority of studies to be included in a meta-analysis are missing variability metrics. To address the limitations of multiple imputation methods, Nakagawa et al. (2023) recently proposed using a weighted average coefficient of variation (CV) estimated from studies in the dataset that do report SDs. Their approach is limited to using the log response ratio as the effect size and can result in biased estimates of the effect size when CVs are different between studies and within-study sample size is relatively small (Nakagawa et al. 2023).

In an ANOVA setting, the within-study residual variance, $\widehat{\sigma}^2$, is equivalent to the mean square error (MSE) and is an estimate of the true variance (σ^2), given the assumption of homogenous (pooled) variances among treatment groups. In ANOVA, post hoc test procedures (mean separation or multiple comparisons) are commonly conducted and are based on a test statistic indicating whether a given pairwise treatment mean difference is statistically different from zero (Montgomery 2001). Post hoc test procedures are based partly on $\widehat{\sigma}^2$, and formulas have been presented to recover MSE from these procedures (Acutis et al. 2022; Lipsey and Wilson 2001; Ma et al. 2008; Ngugi et al. 2011). Ngugi et al. (2011) proposed a method for recovering $\widehat{\sigma}^2$ based on the premise that the actual least significant difference (LSD) between two means lies somewhere between the largest non-significant difference (lower limit) and the smallest significant pairwise difference (upper limit) given by a post hoc test. These bounds are assumed to contain the Fisher LSD bounds because Fisher's LSD test is the most liberal among the post hoc tests. The estimated LSD (ELSD) is then obtained by averaging the upper and lower LSD limits. Fisher's LSD formula is then applied to recover $\widehat{\sigma}^2$ using ELSD as the plug-in estimate of the LSD. This procedure works only for studies with at least one significant mean separation. The accuracy of ELSD decreases when either the number of treatments or the number of non-significant or significant treatment differences decreases. In estimating $\widehat{\sigma}^2$ using ELSD, Ngugi et al. (2011) also use a conservative 97.5th percentile point of the standard normal distribution instead of percentile point of t-distribution with its associated degrees of freedom, which may result in less accurate values of $\widehat{\sigma}^2$. Acutis et al. (2022) recently developed the EXTRACT tool that is coded in the Microsoft Excel environment to recover the pooled SD from multiple comparison tests following ANOVA. EX-TRACT allows users to enter summary statistic metrics for a variety of post hoc methods and experimental designs to recover $\widehat{\sigma}^2$.

Building on the above concepts and ideas, we present MSE FINDR, a Shiny R application for recovering $\widehat{\sigma}^2$ from ANOVA-type studies, where the MSE is missing in studies that otherwise provide treatment means, significance level, number of replicates, and results of post hoc tests. Specifically, MSE FINDR extends on the concepts used by Ngugi et al. (2011) for LSD by applying the correct post hoc test distribution and associated degrees of freedom to recover $\widehat{\sigma}^2$. The MSE FINDR application handles some additional experimental designs and post hoc tests not covered in EXTRACT. The web-based design feature of the Shiny application that may be easier to use than an R package, should appeal to a broad range of user audiences interested in estimating $\widehat{\sigma}^2$ from published studies reporting a continuous response in an ANOVA setting.

Materials and Methods

Software development and workflow

The source code for MSE FINDR, simulations, and datasets used in this study are available at https://github.com/vcgarnica/MSE_FindR. MSE FINDR (version 1.0.1) is written in the R programming language version 4.2.2 (R Core Team 2022). The application supports a variety of one- and two-way experimental designs (Table 1), which are common in the agricultural, ecological, natural resources, engineering, physical, and chemical sciences (Montgomery 2001; Scheiner and Gurevitch 2001). The post hoc mean comparison tests supported by the application are Fisher's LSD, Tukey's honest significant difference (HSD), Bonferroni and Šidák correction for multiple comparisons, and the Scheffé test. To help users navigate the tool, the Shiny application is organized into three main modules: Documentation, File upload, and Estimator (see details below). A user-generated CSV input file containing trial-specific information is required to use the application. Below, we briefly describe how to assemble the input CSV file. Detailed guidelines for users are provided in the GitHub tutorial at https://github.com/vcgarnica/MSE_FindR.

Documentation. This module contains a walk-through tutorial and downloadable example files that are available at https://github.com/vcgarnica/MSE_FindR. Users compile reports with the same experimental configuration, i.e., *experimental design*, *post hoc test*, *significance level*, and *experimental structure* (for two-way designs), in a designated folder (Fig. 1). For each folder, the user will generate a CSV input file comprising the trial information organized into columns: (i) trial identification number, (ii) factor level, (iii) factor level means, (iv) number of replicates or blocks (if applicable based on the experimental design), and (v) corresponding post hoc test letter results (Fig. 1). By specifying a trial-specific identification number column in the CSV input file, many trials with the same configuration can be processed simultaneously.

MSE FINDR handles various one- and two-way experimental designs. In one-way designs, a single column is used to designate the factor (herein referred to as factor *A*) for which $\widehat{\sigma}^2$ is to be recovered. In two-way designs, there are two separate columns in the input CSV file to account for the two factors (*A* and *B*) that are present in the designs. This structure must be maintained regardless of whether the factors are listed completely (means and post hoc test results available for both factors, *A* and *B*) or partially (means and post hoc test results are available for either main effect *A* or *B* but not both) in the studies. While this seems counterintuitive, the latter scenario is

common when the ANOVA interaction effect is not statistically significant, and authors opt to include means and post hoc tables for factors *A* and *B* (main effects) individually. In this case, users must still create an additional column denoting the *number of levels for the omitted factor*, in addition to the present main effect column, for which $\widehat{\sigma}^2$ will be recovered. This step is crucial for the tool to accurately calculate the appropriate degrees of freedom for the extraction of $\widehat{\sigma}^2$. Incorrect $\widehat{\sigma}^2$ values can arise if incomplete or incorrect trial information is included in the CSV input file. Familiarity with the example files and tutorial guidelines is necessary to properly collate trial information in the CSV input file before using the application.

File upload. In this module, users upload a CSV input file that contains trial-specific information organized as per the tutorial's guidelines as briefly described above. The application supports standard CSV formats. The default format uses commas as separators. Default settings can also be overridden to align with the format of your CSV data file.

Estimator. Once the CSV input file has been uploaded, this module details the estimation and extraction of $\widehat{\sigma}^2$. Users specify the underlying trial configuration (*experimental design*, *post hoc test*, and *significance level*) applied to all trials in the designated folder and the CSV input file using the design box (Fig. 1). The selection fields in the column assignment box are dynamically updated as the experimental designs are changed in the design box after the user clicks on the "estimate" button. Users must match columns in the CSV input file to the respective selection fields in the column assignment box. This is a critical step in the proper recovery of $\widehat{\sigma}^2$. A drop-in download button appears after recovering $\widehat{\sigma}^2$, which enables the results to be exported as a CSV file. MSE FINDR output includes all previously added trial-specific information along with the recovered $\widehat{\sigma}^2$ and its respective degrees of freedom.

The MSE FINDR algorithms compute the largest non-significant and the smallest significant difference for all mean pairwise comparisons within the specified post hoc test. The mean of these two values is defined as the ELSD, as described previously by Ngugi et al. (2011). However, unlike Ngugi et al. (2011), estimates of $\widehat{\sigma}^2$ are subsequently calculated by using ELSD as the plug-in for the LSD in the specific post hoc test (Milliken and Johnson 2009).

For Fisher's LSD, $\widehat{\sigma}^2$ is given as:

$$\widehat{\sigma}^2 = 0.5 \cdot n \cdot \left(\frac{\text{ELSD}}{qt(1-\alpha/2, df)}\right)^2 \quad (1)$$

where, n is number of replications or blocks per trial, qt is the quantile value from a Student t-distribution based on the significance level of the posthost test (α) and the error degrees of freedom (df).

For Tukey's HSD, the estimate is calculated as:

$$\widehat{\sigma}^2 = n \cdot \left(\frac{\text{ELSD}}{qtukey(1-\alpha/2, nlevels, df)}\right)^2 \quad (2)$$

where $qtukey$ is the quantile value from the q (or studentized range) distribution based on the significance level of the post hoc test (α), the number of levels of a factor ($nlevels$), the error degrees of freedom (df), and all other variables are as described above.

For the Bonferroni correction, the estimate was calculated as:

$$\widehat{\sigma}^2 = 0.5 \cdot n \cdot \left(\frac{\text{ELSD}}{qt(1-\alpha/2m, df)}\right)^2 \quad (3)$$

in which m is the total number of pairwise comparisons per trial, and all other variables are as described above.

For the Šidák correction, the estimate is calculated as:

$$\widehat{\sigma}^2 = 0.5 \cdot n \cdot \left(\frac{\text{ELSD}}{qt(1-(1-\alpha)^{1/m}, df)}\right)^2 \quad (4)$$

where all variables are as defined above.

For Scheffé's test, the estimate is computed as:

$$\widehat{\sigma}^2 = \frac{n \cdot \text{ELSD}^2}{2(nlevels-1) \cdot qf(1-\alpha, nlevels-1, df)} \quad (5)$$

in which qf is the quantile value from the F-distribution, and all other variables are as defined above. Equations 1, 2, 3, 4, and 5 are for the calculation of $\widehat{\sigma}^2$ from ELSD based on one-way designs. Appropriate adjustments to these equations for two-way

Table 1. Experimental designs and treatment structures used in a simulation study to assess the accuracy of mean square error estimates generated by the MSE FINDR application

Design	Description and treatment structure[v]	Number of factors	Factors and treatment levels[w]	Factor omitted[x]	Source of variation[y]	Post hoc mean comparison information reported in study[z]
1	Latin square	1	*A* (4–8)	–	*A*	Between a single effect
2	Complete randomized design (CRD)	1	*A* (4–20)	–	*A*	Between a single effect
3	Randomized complete block design (RCBD)	1	*A* (4–20)	–	*A*	Between a single effect
4a	Two-way factorial CRD	2	*A* (4–7), *B* (4–7)	–	*A* × *B*	Between interaction effect
4b	Two-way factorial CRD	2	*A* (4–7), *B* (4–7)	*B*	*A*	Between one main effect
5a	Two-way factorial RCBD	2	*A* (4–7), *B* (4–7)	–	*A* × *B*	Between interaction effect
5b	Two-way factorial RCBD	2	*A* (4–7), *B* (4–7)	*B*	*A*	Between one main effect
6a	Split-plot CRD	2	*A* (4–7), *B* (4–7)	*A*	*B*	Between one sub-plot effect
6b	Split-plot CRD	2	*A* (4–7), *B* (4–7)	*B*	*A*	Between one main-plot effect
6c	Split-plot CRD	2	*A* (4–7), *B* (4–7)	–	*B* within *A*	Between interaction effect of sub-plot within main-plot
7a	Split-plot RCBD	2	*A* (4–7), *B* (4–7)	*A*	*B*	Between one sub-plot effect
7b	Split-plot RCBD	2	*A* (4–7), *B* (4–7)	*B*	*A*	Between one main-plot effect
7c	Split-plot RCBD	2	*A* (4–7), *B* (4–7)	–	*B* within *A*	Between interaction effect of sub-plot within main-plot

[v] Experimental design and treatment structure are based on linear models as described in the simulation study.
[w] Values in parentheses are the number of levels for each factor in the experimental design considered in the simulation.
[x] Refers to cases where summary statistics (mean, variance, and post hoc test results) for one factor are missing (i.e., omitted factor) and not available in a report, and the goal is to recover the $\widehat{\sigma}^2$ for the second factor in a two-way design.
[y] Source of variation refers to the factor for which treatment means and post hoc test results are available in the report and for which $\widehat{\sigma}^2$ is to be recovered.
[z] In two-way factorial designs, *A* and *B* are interchangeable. For the split-plot design, factor *A* is the main-plot, while factor *B* is the sub-plot. Factors *A* and *B* are not interchangeable for the split-plot design due to its hierarchical structure.

designs in the calculation of $\hat{\sigma}^2$ are provided in the R code in the application.

Software testing using simulated data

Simulated trials. To assess the performance of the MSE FINDR algorithms in recovering $\hat{\sigma}^2$, randomized controlled trial datasets were simulated for each experimental design and treatment structure available in the application (Table 1). The simulation study was designed to generate data where the number of treatments, treatment effect sizes, and the number of replicates varied randomly for each simulated trial, mimicking a diverse range of experimental settings. For each simulated experimental design, 10,000 individual trial datasets were generated with treatment levels ranging from 4 to 20, depending on the design and treatment structure, and with three to five replications (Table 1).

Linear models. The response variable y (a continuous outcome) was assumed to be independent and normally distributed. We adopt the notation in Oehlert (2000) to describe the linear models used in our simulations. One-way experiments arranged in a completely randomized design (CRD) were simulated using the linear model:

$$y_{ij} = \mu + A_i + e_{ij} \quad (6)$$

where μ is the overall mean, A_i is the effect of the ith level of factor A, j is the replication within the level of A, and e_{ij} is the random error (Gaussian-distributed with a mean of zero and variance σ^2). The linear model for one-way experiments arranged in a randomized complete block design (RCBD) was:

$$y_{ik} = \mu + A_i + \delta_k + e_{ik} \quad (7)$$

where δ_k represents the kth block effect, e_{ik} is the error term, and A_i is as described above.

Data for the two-way factorial design with treatments arranged in a CRD were simulated using the model:

$$y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + e_{ijk} \quad (8)$$

where B_j is the effect due to the jth level of factor B, $(AB)_{ij}$ is the effect of the interaction between the ith level of A and the jth level of B, and e_{ijk} is the error term.

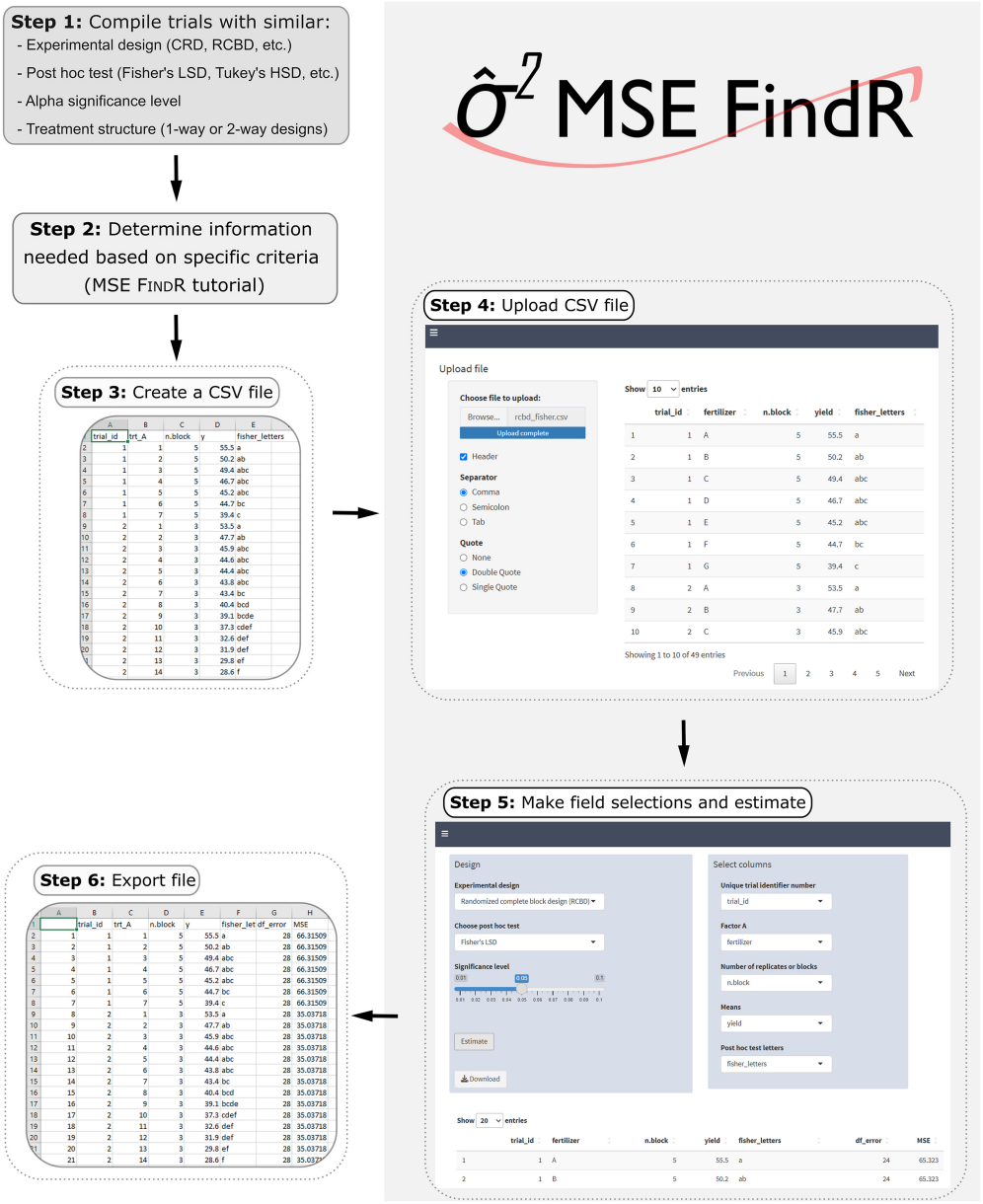

Fig. 1. Conceptual flow chart to illustrate the use of MSE FINDR highlighting the important steps of information input and the recovery of $\hat{\sigma}^2$ based on a CSV file containing multiple trials arranged in a randomized complete block design and means separated using Fisher's least significant difference (LSD) test with α set to 0.05.

The two-way factorial design with treatments arranged in a RCBD was specified as:

$$y_{ijk} = \mu + \delta_k + A_i + B_j + (AB)_{ij} + e_{ijk} \quad (9)$$

where δ_k is the block effect and e_{ijk} is the residual error term. All the other terms are as described for equation 8.

The model for a split-plot design with the whole-plot factor (A) arranged in a CRD was:

$$y_{ijk} = \mu + A_i + \varphi_{k(i)} + B_j + (AB)_{ij} + e_{k(ij)} \quad (10)$$

where $\varphi_{k(i)}$ is the whole-plot error component and $e_{k(ij)}$ is the split-plot level error. The subscript notation $k(i)$ conveys that the whole-plot error is nested within treatment i.

The linear model for a split-plot design with the whole-plot factor (A) arranged as a RCBD was:

$$y_{ijkl} = \mu + \delta_k + A_i + \varphi_{l(ik)} + B_j + (AB)_{ij} + e_{l(ijk)} \quad (11)$$

where δ_k is the blocking effect, $e_{l(ijk)}$ is the split-plot error, and all other terms are as described for equation 10.

For experiments with a basic Latin square design, y was generated using:

$$y_{imn} = \mu + A_i + R_m + C_n + e_{imn} \quad (12)$$

where μ and A_i are as described above, R_m is the effect of the mth row, C_n is the effect of the nth column, and e_{imn} is the random error term.

Treatment effect size and error specification. Factor effects were fixed within a trial but allowed to vary across trials. The overall mean μ was simulated from a *Gamma* (k, θ) distribution with shape parameter $k = 60.2$ and scale parameter $\theta = 1$, ensuring that μ was strictly positive. The effects of treatments, their interactions, block (δ_k), row (R_m), and column (C_n) were simulated from a Gaussian (Normal) distribution: A_i and $B_j \sim N(3, 2)$; $AB_{ij} \sim N(1, 1.5)$; and δ_k, R_m, and $C_n \sim N(1, 0.2)$. The error terms in equations 6, 7, and 12 were simulated from a $N(0, 5)$ distribution, the error terms in equations 8 and 9 from a $N(0, 7)$ distribution, and the error terms in equations 10 and 11 from a $N(0, 6)$ distribution. The effects $\varphi_{k(i)}$ and $\varphi_{l(ik)}$ were simulated from a $N(0, 3)$ distribution.

Data analysis

From the 10,000 datasets simulated for each experimental design, we randomly sampled 1,000 trials that yielded statistically significant ($P \leq 0.05$) ANOVA results for the factors of interest (A, B, or both) and extracted their respective $\widehat{\sigma}^2$ values (i.e., actual $\widehat{\sigma}^2$). For each of these trials, treatment means were estimated via least squares, and post hoc mean comparisons were performed at $\alpha = 0.05$ using the R packages "emmeans" version 1.9.0 (Lenth et al. 2023), "agricolae" version 1.3-7 (de Mendiburu 2023), and "multcomp" version 1.4-25 (Hothorn et al. 2023). The post hoc mean separation tests evaluated were Fisher's LSD, Tukey HSD, the Šidák and Bonferroni corrections, and the Scheffé test.

Information on trial design and treatment means and associated post hoc test results were then submitted to the MSE FINDR application to recover $\widehat{\sigma}^2$ (i.e., MSE FINDR $\widehat{\sigma}^2$). Incomplete data reporting is of particular concern (Gurevitch and Hedges 1999), so we also examined scenarios where summary statistics and post hoc test results are reported for one factor only in two-way factorial and split-plot designs (no reported results having been presented on the second factor). The goal in these cases is to recover $\widehat{\sigma}^2$ for the factor that was reported (Table 1).

Lin's concordance analysis (Lin 1989), implemented using the R package 'DescTools' version 0.99.50 (Signorell 2023), was used to assess the agreement between MSE FINDR $\widehat{\sigma}^2$ and the actual $\widehat{\sigma}^2$. Lin's concordance correlation coefficient (ρ_c) measures the variation of data from the line of concordance (a slope of one and zero intercept). The concordance coefficient is the product of Pearson's correlation coefficient (r), which is a measure of the precision (or variability) with which the MSE FINDR $\widehat{\sigma}^2$ values estimate the actual $\widehat{\sigma}^2$, and a coefficient of bias (C_b), which is a measure of the closeness of the best-fitting line to the concordance line. Values of ρ_c range from −1 to +1, with ρ_c values near +1 indicating strong concordance, while those near −1 indicating a strong discordance. Values of C_b range from 0 to +1, with C_b near +1 indicating a closeness of the best-fitting line to the concordance line (i.e., low bias).

Results

Concordance of $\widehat{\sigma}^2$ estimates

The extent of agreement (ρ_c) between $\widehat{\sigma}^2$ recovered by MSE FINDR and the actual $\widehat{\sigma}^2$ depended on the experimental design, treatment structure, and the post hoc test used to compare treatment means.

For one-way designs (Latin square, CRD, and RCBD), values of ρ_c were high (>0.89) across the post hoc mean separation tests with values of 0.92, 0.92, and 0.94 for Designs 1, 2, and 3, respectively (Table 2). Similarly, values of ρ_c were high (>0.82) for two-way designs when summary statistics were present for both factors A and B (Designs 4a, 5a, 6c, and 7c) (Table 2). In this case, the mean of

Table 2. Lin's concordance correlation coefficient (ρ_c) and correlation coefficient (r) for the agreement between actual $\widehat{\sigma}^2$ obtained from analysis of variance of simulated trials and $\widehat{\sigma}^2$ recovered by MSE FINDR

	Concordance correlation coefficient (ρ_c)					Correlation coefficient (r)					Mean[w]	
Design[x]	Fisher's LSD[y]	Tukey HSD[y]	Šidák	Bonferroni	Scheffé	Fisher's LSD[y]	Tukey HSD[y]	Šidák	Bonferroni	Scheffé	ρ_c	r
1	0.89	0.92	0.93	0.93	0.92	0.89	0.92	0.92	0.92	0.92	0.92	0.91
2	0.93	0.93	0.92	0.92	0.90	0.93	0.93	0.92	0.92	0.90	0.92	0.92
3	0.93	0.94	0.94	0.94	0.93	0.93	0.94	0.94	0.94	0.94	0.94	0.94
4a	0.99	0.97	0.96	0.96	0.82	0.99	0.97	0.96	0.96	0.84	0.94	0.94
4b	0.41	0.53	0.55	0.55	0.57	0.50	0.58	0.59	0.60	0.64	0.52	0.58
5a	0.99	0.97	0.96	0.96	0.91	0.99	0.97	0.96	0.96	0.92	0.96	0.96
5b	0.45	0.54	0.54	0.54	0.52	0.46	0.62	0.59	0.59	0.59	0.52	0.57
6a	0.39	0.59	0.56	0.56	0.56	0.53	0.66	0.67	0.67	0.65	0.53	0.64
6b	0.87	0.89	0.90	0.90	0.88	0.88	0.90	0.89	0.90	0.91	0.89	0.90
6c	0.95	0.89	0.88	0.87	0.83	0.95	0.90	0.90	0.90	0.87	0.88	0.90
7a	0.44	0.52	0.52	0.52	0.52	0.54	0.58	0.57	0.58	0.56	0.50	0.57
7b	0.81	0.89	0.89	0.88	0.89	0.82	0.90	0.89	0.89	0.89	0.87	0.88
7c	0.93	0.89	0.88	0.88	0.84	0.94	0.90	0.89	0.89	0.86	0.88	0.90
Mean[z]	0.76	0.81	0.80	0.80	0.78	0.79	0.83	0.82	0.82	0.80	0.79	0.82

[w] Values are means of ρ_c and r summarized by experimental design across post hoc mean separation tests.
[x] Design description is as in Table 1.
[y] LSD is the least significant difference, while HSD is the honest significant difference.
[z] Means of ρ_c and r summarized by post hoc mean separation test across experimental designs.

ρ_c across post hoc mean separation tests for the two-way factorial CRD (Design 4a) and RCBD (Design 5a) were 0.94 and 0.96, respectively, while the mean of ρ_c for the two-way split-plot CRD (Design 6c) and RCBD (Design 7c) was 0.88. Across all one- and two-way designs (when both factors are present), ρ_c were high and relatively similar for all post hoc mean separation tests (values 0.92 to 0.94) except for the Scheffé test, where ρ_c was 0.88 (Table 2).

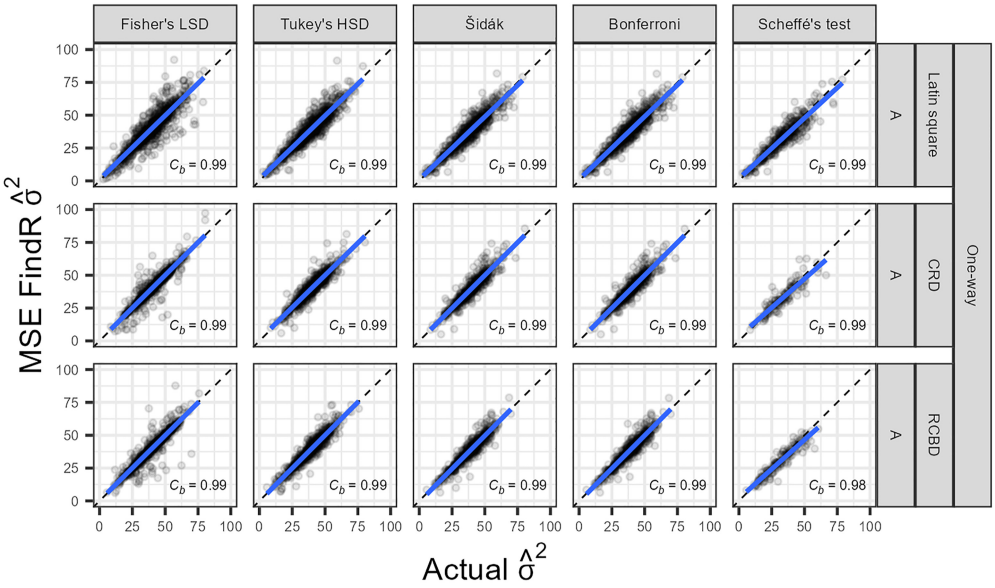

Fig. 2. Relationship between MSE FINDR recovered $\hat{\sigma}^2$ and actual $\hat{\sigma}^2$ in simulated data from one-way experimental designs: Latin square, completely randomized design (CRD), and randomized completed block design (RCBD). Actual $\hat{\sigma}^2$ values were obtained from an analysis of variance (ANOVA) of simulated data, while MSE FINDR $\hat{\sigma}^2$ values were extracted using treatment means, post hoc test results stemming from the ANOVA, and other basic trial information. The open circles are the $\hat{\sigma}^2$ estimates, the dashed line represents the concordance line, indicating perfect alignment between actual and estimated $\hat{\sigma}^2$, and the blue line represents the best-fitting linear regression line to the data. The bias correlation factor, C_b, is a measure of the closeness of the best fitting line to the concordance line.

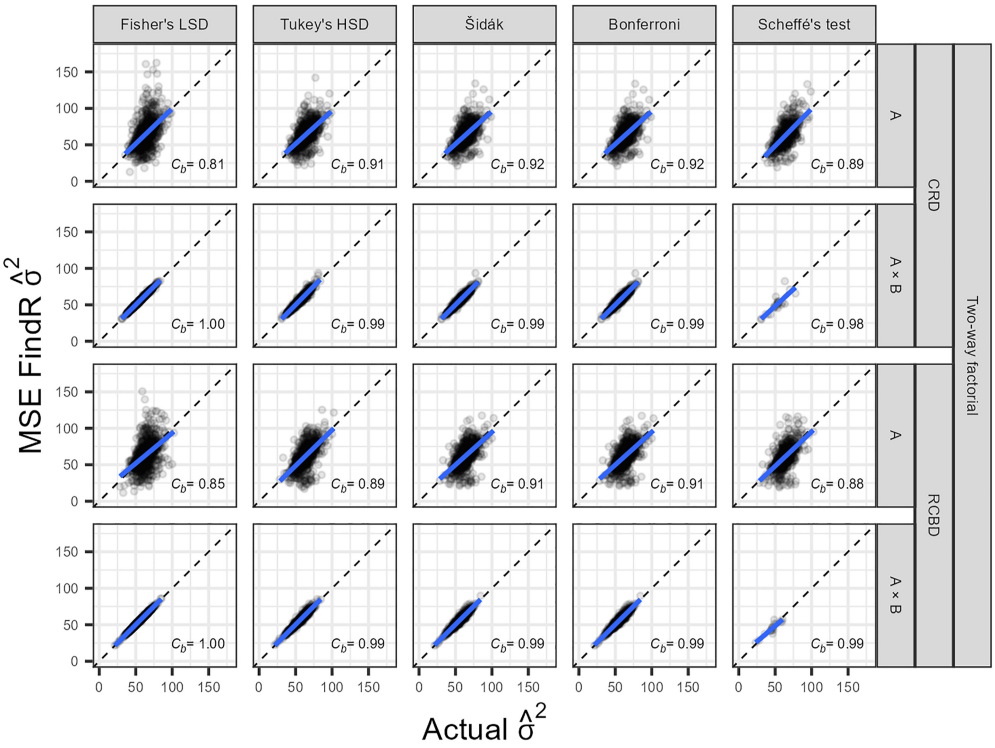

Fig. 3. Relationship between MSE FINDR $\hat{\sigma}^2$ and actual $\hat{\sigma}^2$ for simulated data from two-way experimental designs: completely randomized design (CRD) and randomized completed block design (RCBD). For both CRD and RCBD designs, two variations of the treatment structure are considered: (i) two factors, *A* and *B*, and their interaction (*A* × *B*) are assessed and (ii) one factor (herein, *B*, and hence the interaction) is omitted, and the goal is to recover $\hat{\sigma}^2$. Actual $\hat{\sigma}^2$ values were obtained from an analysis of variance (ANOVA) of simulated data, while MSE FINDR $\hat{\sigma}^2$ values were generated using the treatment means, post hoc test results stemming from the ANOVA, and other basic trial information. The open circles are the $\hat{\sigma}^2$ estimates, the dashed line represents the concordance line, indicating perfect alignment between actual and estimated $\hat{\sigma}^2$, and the blue line represents the best-fitting linear regression line to the data. The bias correlation factor, C_b, is a measure of the closeness of the best fitting line to the concordance line.

The impact of the absence of summary statistics on ρ_c for one factor in the two-way designs was dependent on the experimental structure. For the two-way factorial designs of CRD and RCBD, absence of summary statistics of one factor resulted in a 30 to 59% decrease in ρ_c, with a reduction of 45% across post hoc tests. For the two-way split-plot designs of CRD and RCBD, the impact of the absence of summary statistics on ρ_c depended on whether the missing statistics were of the main- or sub-plot factor. The mean value of ρ_c across post hoc tests was reduced by 40 to 43% when data of the main-plot factor were missing, while there was no reduction in ρ_c when data for the sub-plot factor were missing. Values of ρ_c for the latter were similar to those where summary statistics of both factors were present. Across the two-way designs where summary statistics of one factor (factorial design) or the main-plot factor (in split-plot design) were missing, ρ_c was relatively lower for Fisher's LSD (ρ_c = 0.42) than for the more conservative post hoc mean separation tests (ρ_c = 0.54). Similarly, ρ_c was relatively lower across experimental designs and treatment structures for Fisher's LSD (ρ_c = 0.76) than for more conservative post hoc mean separation tests examined (ρ_c = 0.78 to 0.81) (Table 2).

Bias and precision of $\hat{\sigma}^2$ estimates

Bias was generally low (C_b values ranging from 0.97 to 1.0; Figs. 2, 3, and 4) across all post hoc mean separation tests and design structures. The main exception was for the Fisher's LSD test and for design structures where summary statistics (especially the main-plot factor in the split-plot design) were missing, for which bias was relatively high, with C_b values ranging from 0.77 to 0.85 (Figs. 3 and 4).

In contrast to bias, the precision (r) of MSE FINDR estimates of $\hat{\sigma}^2$ was strongly influenced by the absence of reported summary statistics (i.e., means and post hoc test results). For example, estimates of $\hat{\sigma}^2$ were precise (r ranging from 0.90 to 0.96; Table 1) with one-way designs when summary statistics were available, regardless of the post hoc mean separation test, and similarly for the two-way factorial design when summary statistics for both factors were present (r ranging from 0.92 to 0.99), except when Scheffé's test was used

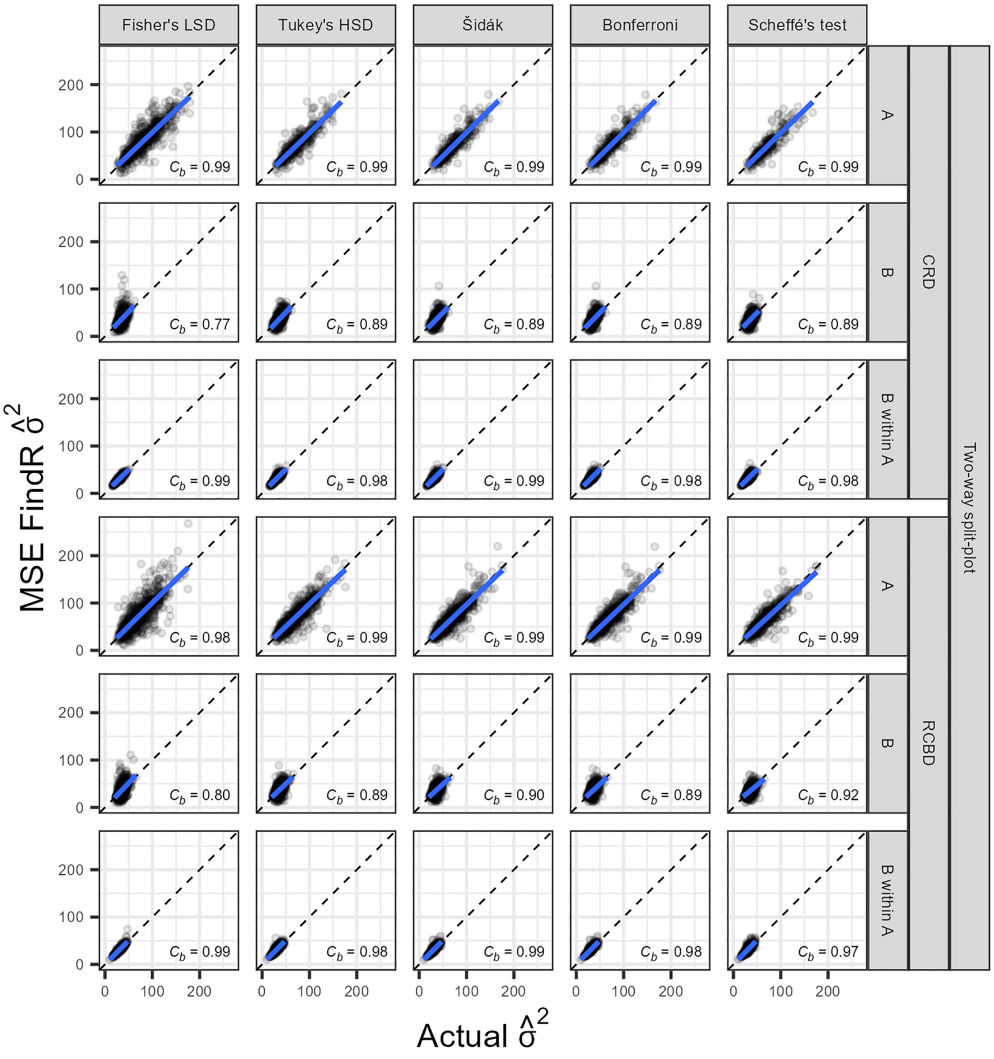

Fig. 4. Relationship between MSE FINDR $\hat{\sigma}^2$ and actual $\hat{\sigma}^2$ for simulated data from split-plot designs with treatments arranged in a completely randomized design (CRD) and randomized completed block design (RCBD). For both CRD and RCBD designs, three variations of treatment structures are considered: (i) one factor (*A*) is omitted, *B* is present, and the goal is to recover $\hat{\sigma}^2$ of *B*; (ii) one factor (*B*) is omitted, *A* is present, and the goal is to recover the $\hat{\sigma}^2$ of *A*; and (iii) both factors are present (i.e., *B* within *A*), wherein factor *A* is the main-plot, *B* is the sub-plot, and the goal is to recover $\hat{\sigma}^2$ for *A*. Actual $\hat{\sigma}^2$ values were obtained based on an analysis of variance (ANOVA) of simulated data, while MSE FINDR $\hat{\sigma}^2$ values were generated using treatment means, post hoc test results stemming from the ANOVA, and other basic trial information. The open circles are the $\hat{\sigma}^2$ estimates, the dashed line represents the concordance line, indicating perfect alignment between actual and estimated $\hat{\sigma}^2$, and the blue line represents the best-fitting linear regression line to the data. The bias correlation factor, C_b, is a measure of the closeness of the best fitting line to the concordance line.

($r = 0.84$; Design 4a) (Table 1). However, missing summary statistics for one factor (e.g., *B*) in the two-way factorial design resulted in a much lower precision of $\widehat{\sigma}^2$ estimates for the other factor (*A*), with *r* values ranging from 0.54 to 0.64 across post hoc tests. Likewise, the precision of $\widehat{\sigma}^2$ estimates for the split-plot design was high when summary statistics were available for both factors (*r* values from 0.88 to 0.95) but considerably lower when summary statistics for either the sub-plot factor ($0.53 \geq r \geq 0.91$) or the main-plot factor ($0.53 \geq r \geq 0.67$) were missing (Table 1).

Discussion

Meta-analysis combines information from multiple studies to draw generalized inferences from a body of research. However, incomplete reporting of variances in the primary reports can lead to biased estimates of the effect size if such reports are excluded because of the missing variances data (Weir et al. 2018). Thus, there has been an interest in developing techniques that still allow meta-analysis to include studies where the basic summary statistics are lacking. In this study, we present and test MSE FINDR, a user-friendly web-based interface that enables the recovery of $\widehat{\sigma}^2$ from reports that have not included the within-study variance but do contain the treatment means, post hoc test results, and a description of the experimental protocol (statistical design and number of replications). The ability of MSE FINDR to recover $\widehat{\sigma}^2$ was analyzed using simulated data representing a variety of experimental designs and treatment structures commonly found across a range of scientific fields (Montgomery 2001; Scheiner and Gurevitch 2001). Actual values of $\widehat{\sigma}^2$ from simulated trials were compared with estimates of $\widehat{\sigma}^2$ recovered by MSE FINDR given an experiment's summarized results (treatment means and post hoc test results but no measure of variance). Our results indicated that MSE FINDR is accurate in estimating $\widehat{\sigma}^2$, but the accuracy does depend on the experimental design and treatment structure of the trial.

Simulations suggested that MSE FINDR performs well with both one-way (Latin square, CRD, and RCBD) and two-way (factorial and split-plot) designs when means and post hoc test results are available for all factors involved, irrespective of the post hoc test used. The accuracy of MSE FINDR is reduced for two-way split-plot designs when the whole-plot factor summaries are missing (no means or post hoc test results given) with the goal of recovering $\widehat{\sigma}^2$ for the sub-plot factor. Our simulations also indicate that for two-way factorial designs, the absence of information on one factor reduces the accuracy of MSE FINDR in recovering $\widehat{\sigma}^2$ for the second factor. In the latter case, users should evaluate whether to use the recovered $\widehat{\sigma}^2$ in their quantitative synthesis and how this could impact the interpretation of the results of their meta-analysis.

Recently, Acutis et al. (2022) developed EX-TRACT, a Microsoft Excel-based tool that recovers $\widehat{\sigma}^2$ for a variety of experimental designs and post hoc methods, some of which overlap with those implemented in MSE FINDR. The MSE FINDR application includes three aspects that are not covered by EX-TRACT. Firstly, MSE FINDR implements the recovery of $\widehat{\sigma}^2$ for the one-way Latin square design, which is not included in EX-TRACT. Secondly, MSE FINDR recovers $\widehat{\sigma}^2$ for studies where Bonferroni and Šidák corrections and the Scheffé test are the post hoc tests, and these multiple comparison tests are not implemented in EX-TRACT. Thirdly, MSE FINDR distinguishes between two-way split-plot designs in which the main plot is in a CRD or in an RCBD, whereas EX-TRACT handles main plots arranged in an RCBD only. Several trials with the same configurations can also be processed simultaneously by MSE FINDR compared with EX-TRACT. The implementation of MSE FINDR and its Shiny-based web interface within the R environment permits new algorithms to be easily integrated into the application. By providing this Shiny application in R, a smoother integration in the programming environments is enabled within which researchers can manipulate datasets and perform other analyses while taking advantage of the rich universe of existing R packages and graphics capabilities (Title et al. 2022). EX-TRACT was validated by comparing its output to a known standard deviation value generated from simulated experiments (Acutis et al. 2022). The accuracy of MSE FINDR was determined using Lin's concordance analysis (Lin 1989) over repeatedly simulated data. We compared the performance of MSE FINDR with that of EX-TRACT using a subset of our simulated data. Estimates of MSE FINDR $\widehat{\sigma}^2$ were very similar to the mean values of $\widehat{\sigma}^2$ extracted using EX-TRACT for comparable experimental designs and posthost test results (Supplementary Tables S1 to S3).

In summary, MSE FINDR is an additional tool that users can use to estimate $\widehat{\sigma}^2$ from reports that lack information on within-study variance but provide means, post hoc test results, and other basic experimental design information. This should allow users to easily estimate $\widehat{\sigma}^2$ from a variety of studies with pertinent information to calculate the required effect size for inclusion in meta-analyses of estimates of effect size and variances based on a continuous response in an ANOVA setting. Complementary tools that compute a range of effect sizes (Lipsey and Wilson 2001) have been implemented in a web-based Practical Meta-Analysis Effect Size Calculator (https://www.campbellcollaboration.org/research-resources/effect-size-calculator.html). These web-based meta-analytical tools, in combination with MSE FINDR should expand the array of studies and reports for inclusion in quantitative research synthesis that would have otherwise been omitted due to the lack of information on within-study variability or related metrics. In its current version, MSE FINDR is not designed to recover $\widehat{\sigma}^2$ of the main-plot factor in split-plot designs if only summary statistics of the sub-plot factor have been reported. Thus, subsequent improvements are needed when the interaction is significant and when one is interested in recovering $\widehat{\sigma}^2$ for both factors. Additional experimental designs (e.g., split-split-plot and three-way factorial designs) should be implemented in future versions of MSE FINDR, and the tool should also expand on the estimation method to return not only $\widehat{\sigma}^2$ but the upper and lower bounds of the estimate too. This study highlights the need for published reports to provide, at a minimum, some basic summary statistics (means and variance) where raw datasets are not available to readers to facilitate quantitative syntheses of results for broad generalization across a body of research.

Acknowledgments

The authors thank Felipe Dalla Lana (Louisiana State University), Mladen Cucak and Olanrewaju Shittu (The Pennsylvania State University), and Wanderson Bucker Moraes (The Ohio State University) for helpful comments on the documentation and development of the application.

Literature Cited

Acutis, M., Tadiello, T., Perego, A., Di Guardo, A., Schillaci, C., and Valkama, E. 2022. EX-TRACT: An excel tool for the estimation of standard deviations from published articles. Environ. Model. Softw. 147:105236.

Adams, D. C., Gurevitch, J., and Rosenberg, M. S. 1997. Resampling tests for meta-analysis of ecological data. Ecology 78:1277-1283.

Batson, S., and Burton, H. 2016. A systematic review of methods for handling missing variance data in meta-analyses of interventions in type 2 diabetes mellitus. PLoS One 11:e0164827.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., and Rothstein, H. R. 2009. Introduction to Meta-Analysis. John Wiley & Sons Ltd., New York, NY.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., and Rothstein, H. R. 2021. Introduction to Meta-Analysis, 2nd ed. John Wiley & Sons, New York, NY.

Chowdhry, A. K., Dworkin, R. H., and McDermott, M. P. 2016. Meta-analysis with missing study-level sample variance data. Stat. Med. 35:3021-3032.

de Mendiburu, F. 2023. agricolae: Statistical Procedures for Agricultural Research. R package version 1.3-7. https://cran.r-project.org/package=agricolae

Fohrafellner, J., Zechmeister-Boltenstern, S., Murugan, R., Keiblinger, K., Spiegel, H., and Valkama, E. 2023. Meta-analysis protocol on the effects of cover crops on pool specific soil organic carbon. MethodsX 11:102411.

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.

Gurevitch, J., and Hedges, L. V. 1999. Statistical issues in ecological meta-analyses. Ecology 80:1142-1149.

Gurevitch, J., Koricheva, J., Nakagawa, S., and Stewart, G. 2018. Meta-analysis and the science of research synthesis. Nature 555:175-182.

Higgins, J. P. T., Deeks, J. J., and Altman, D. G. 2008. Special topics in statistics. Pages 481-529 in: Cochrane Handbook for Systematic Reviews of Interventions: Cochrane Book Series. J. P. T. Higgins and S. Green, eds. Wiley, Chichester, England.

Hothorn, T., Bretz, F., Westfall, P., Heiberger, R. M., Schuetzenmeister, A., and Scheibe, S. 2023. multcomp: Simultaneous Inference in General Parametric Models. R package version 1.4-25. https://cran.r-project.org/package=multcomp

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.

Kong, F., Li, Q., Yang, Z., and Chen, Y. 2023. Does the application of biogas slurry reduce soil N_2O emissions and increase crop yield? – A systematic review. J. Environ. Manag. 342:118339.

Lajeunesse, M. J. 2013. Recovering missing or partial data from studies: A survey of conversions and imputations for meta-analysis. Pages 195-206 in: Handbook of Meta-Analysis in Ecology and Evolution. J. Koricheva, J. Gurevitch, and K. Mengersen, eds. Princeton University Press, Princeton, NJ.

Lenth, R. V., Bolker, B., Buerkner, P., Giné-Vázquez, I., Herve, M., Jung, M., Love, J., Miguez, F., Piaskowski, J., Riebl, H., and Singmann, H. 2023. emmeans: Estimated Marginal Means. R package version 1.9.0. https://github.com/rvlenth/emmeans

Lin, L. I.-K. 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics 45:255-268.

Lipsey, M. W., and Wilson, D. B. 2001. Practical Meta-Analysis. Sage Publications, Inc., Thousand Oaks, CA.

Ma, J., Liu, W., Hunter, A., and Zhang, W. 2008. Performing meta-analysis with incomplete statistical information in clinical trials. BMC Med. Res. Methodol. 8:56.

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

Milliken, G. A., and Johnson, D. E. 2009. Analysis of Messy Data Volume 1: Designed Experiments, 2nd ed. CRC Press, Boca Raton, FL.

Montgomery, D. C. 2001. Design and Analysis of Experiments, 5th ed. John Wiley & Sons, New York, NY.

Nakagawa, S., Noble, D. W. A., Lagisz, M., Spake, R., Viechtbauer, W., and Senior, A. M. 2023. A robust and readily implementable method for the meta-analysis of response ratios with and without missing standard deviations. Ecol. Lett. 26:232-244.

Ngugi, H. K., Lehman, B. L., and Madden, L. V. 2011. Multiple treatment meta-analysis of products evaluated for control of fire blight in the eastern United States. Phytopathology 101:512-522.

Oehlert, G. W. 2000. A First Course in Design and Analysis of Experiments. W. H. Freeman and Company, New York City, NY.

R Core Team. 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Scheiner, S. M., and Gurevitch, J. 2001. Design and Analysis of Ecological Experiments, 2nd ed. Oxford University Press, New York City, NY.

Signorell, A. 2023. DescTools: Tools for Descriptive Statistics. R package version 0.99.52. https://andrisignorell.github.io/DescTools/

Sparks, A. H., Del Ponte, E. M., Alves, K. S., Foster, Z. S. L., and Grünwald, N. J. 2023. Openness and computational reproducibility in plant pathology: Where we stand and a way forward. Phytopathology 113:1159-1170.

Tadiello, T., Acutis, M., Perego, A., Schillaci, C., and Valkama, E. 2023. Soil organic carbon under conservation agriculture in Mediterranean and humid subtropical climates: Global meta-analysis. Eur. J. Soil Sci. 74:e13338.

Thiessen Philbrook, H., Barrowman, N., and Garg, A. X. 2007. Imputing variance estimates do not alter the conclusions of a meta-analysis with continuous outcomes: A case study of changes in renal function after living kidney donation. J. Clin. Epidemiol. 60:228-240.

Title, P. O., Swiderski, D. L., and Zelditch, M. L. 2022. ECOPHYLOMAPPER: An R package for integrating geographical ranges, phylogeny and morphology. Methods Ecol. Evol. 13:1912-1922.

Weir, C. J., Butcher, I., Assi, V., Lewis, S. C., Murray, G. D., Langhorne, P., and Brady, M. C. 2018. Dealing with missing standard deviation and mean values in meta-analysis of continuous outcomes: A systematic review. BMC Med. Res. Methodol. 18:25.