Article

MetaAnaly: The platform-independent computational tool for meta-analysis in the paradigm of new statistics

WenJun Zhang
School of Life Sciences, Sun Yat-sen University, Guangzhou, China
E-mail: zhwj@mail.sysu.edu.cn, wjzhang@iaees.org

Received 16 July 2023; *Accepted 15 December 2023*; *Published online 1 January 2024*; *Published 1 June 2024*

Abstract
Meta-analysis is a statistical method used in systematic review to quantitatively integrate the results of multiple related studies to obtain a pooled result that can represent these studies. Meta-analysis overcomes the limitations of traditional reviews that only conduct qualitative research. In present study, I developed a platform-independent computational tool for meta-analysis. It is a comprehensive tool consisting of a full set of meta-analysis methodology, including the methods for fixed-effects model and random-effects model and the methods for heterogenetity testing in which the effect size tests based heterogeneity testing were proposed. Effect size tests of difference significance for post meta-analysis were also presented. The computational tool is a web browser based meta-analyzer that includes both online and offline versions and can be used on various computing devices (PCs, iPads, smartphones, etc.), operating systems (Windows, Mac, Android, Harmony, etc.) and web browsers (Chrome, Firefox, etc). It can be used in various sciences as medicine, biology, ecology, psychology, sociology, economy, physics and chemistry etc.

Keywords meta-analysis; computational tool; effect size; estimator; averaged estimator; fixed-effects model; random-effects model; heterogeneity.

Network Biology
ISSN 2220-8879
URL: http://www.iaees.org/publications/journals/nb/online-version.asp
RSS: http://www.iaees.org/publications/journals/nb/rss.xml
E-mail: networkbiology@iaees.org
Editor-in-Chief: WenJun Zhang
Publisher: International Academy of Ecology and Environmental Sciences

1 Introduction

The systematic review is a method of systematically and quantitatively summarizing and integrating literature and making scientific inferences around specific research issues. It is an advanced review. The so-called "systematic" refers specifically to the comprehensiveness of the collection of original literature, the reliability and uniformity of operating methods, and the use of meta-analyses to quantify and to integrate the results (Tang and Yang, 2015). As a method of summarizing and integrating evidence, the systematic review can be used in any field. The systematic review has been widely used in education, psychology, behavior, social sciences and other fields, and has yielded a significant impact. Actually researchers have been developing

various methods for pooling or integrating the results from studies that provide commensurable evidence about a measurable effect (Bangert-Drowns, 1986). Promoted by the information explosion in the scientific literature (Adair and Vohra, 2003), meta-analyses emerged from 1970s. Meta-analyses are a statistical method used in systematic reviews to quantitatively pool the results of multiple related studies to obtain average results that represent these studies. It is an important (but not necessary) part of systematic reviews. Meta-analysis overcomes the limitations of traditional reviews that only conduct qualitative research, and is proposed to use quantitative methods to synthesize the results of different studies, which is a more systematic and standardized review method.

So far many meta-analysis methods have been proposed and integrated to determine unbiased and ideal estimates of the effects (Viechtbauer, 2005). Petropoulou and Mavridis (2017) assessed twenty estimators in pooling results from studies. In a meta-analysis study, Langan et al. (2019) compared nine estimators using simulated data. Veroniki et al. (2016) classified sixteen estimators into two categories, non-iterative estimators, including the DerSimonian and Laird (DL; DerSimonian and Laird, 1986) and the Hedges and Olkin (HO; Hedges and Olkin, 1985), and iterative estimators, including the Paule and Mandel (PM; Paule and Mandel, 1982, 1989), maximum likelihood (ML; Hardy and Thompson, 1996), and restricted maximum likelihood (REML; Raudenbush, 2009; Viechtbauer, 2007). The DL, HO, PM, ML, and REML estimators are all frequentist methods in which the pooled effect size is estimated as a weighted average. Unfortunately, the conclusions on performance and recommendations of these estimators from these studies are often inconsistent or even contradictory (Huang, 2023). Huang (2023) argued that the averaged estimator performs better than individual estimators in terms of bias and efficiency.

As an example of meta-analyses, the quality of the evidence is crucial in evidence-based medicine, and meta-analysis of randomized controlled clinical trials of good quality is considered the highest level of evidence (European Medicines Agency, 2006; Röver et al., 2015). Integrating and analyzing the results of multiple studies to provide more reliable and comprehensive evidence, meta-analyses are particularly important in this field due to the lack of large trials and the problem of small numbers of available studies (Korn et al., 2013; Gagne et al, 2014). As the top of the evidence-based medicine pyramid, meta-analyses have an unshakable important status and value (Riaz et al., 2018; Celeng et al., 2019; Pearce et al., 2022; Kim et al., 2023; Zhu et al., 2023). In 2023 alone, totally 32,476 meta-analysis papers were published around the world.

Meta-analyses are generally a computation-spending approach. Fortunately software to compute the different estimates is provided e.g., the metafor and metaR packages in R (Viechtbauer, 2010; Schwarzer, 2014), the NIST Consensus Builder (Koepke et al., 2017).

In present study, I propose a platform-independent computational tool for meta-analyses. It is a comprehensive tool that includes a full set of methodology used in meta-analyses, which is a web browser based meta-analyzer. Considering the *p*-value based statistical significance paradigm has been identified as one of the sources of false conclusions and research reproducibility crisis (Sellke et al., 2001; Ioannidis, 2005; Tong, 2019; Wasserstein et al., 2019; Huang, 2020; Li, 2021; Zhang, 2022a-c, 2023; Zhang and Qi, 2024), the available *p*-values are greatly enhanced in this tool. The computational tool can be used in various sciences as medicine, biology, ecology, psychology, sociology, economy, physics and chemistry etc.

2 Methodology of Meta-analysis

Generally a meta-analysis mainly follow several procedures: (1) Selecting functions based on variable types and effect indicators; (2) setting the specific effect indicators; (3) setting various methods for weighted averaging of different studies; (4) setting estimation methods for between-study variations, and (5) calculate

effect size and confidence interval. Meta-analyses must follow these important assumptions or principles: (1) The problems explored by the original studies to be synthesized must be the same, so they come from the same population, and the results are similar; (2) all relevant studies must be included when pooling the results, not only some selected studies, in order to reduce selection bias; (3) it is assumed that all included studies are free of bias, and the difference in results is entirely caused by sampling error, and (4) use the weighted averaging methods to quantitatively estimate the true effect, which is the most important in meta-analyses. However it itself does not guarantee the realization of the first three assumptions or principles. Therefore, when conducting a meta-analysis, more measures must be taken to ensure the realization of the first three assumptions or principles, in order to control bias and ensure the reliability of meta-analysis.

2.1 Heterogeneity testing

In most cases, the results of the same type of study on the same question are different. These variations can be caused by three different factors: probabilistic factors, clinical factors, and methodological factors (Rücker et al., 2008). Variation caused by probabilistic factors refers to the between-study variation caused by sampling, that is, the variation caused by sampling error. Due to the existence of sampling error, there must be variation in the results of different studies. Variation caused by probabilistic factors always exists, and it is part or all of the overall variation. In meta-analyses, the variation caused by non-probabilistic factors including clinical factors and methodological factors is generally called heterogeneity, which are called clinical heterogeneity and methodological heterogeneity respectively. Clinical heterogeneity refers to the between-study variation due to clinical factors such as patients, interventions, control treatments, outcome indicators, and intervention settings. Methodological heterogeneity refers to the variation caused by study designs, bias control, and statistics. For example, differences in study design, such as the difference between cohort studies and randomized controlled trials, and differences in bias control measures, such as random grouping and group concealment used in clinical trials, etc. If there is no heterogeneity, it means that the between-study variations are mainly caused by probabilistic factors, and the studies are said to be homogeneous (Tang and Mao, 2015; Mao, 2019).

For the variation caused by probabilistic factors, by pooling data from multiple different studies, the sample size can be effectively increased, the sampling error can be reduced, and the accuracy of the overall results can be improved, thereby reducing the impact of probabilistic factors on the overall results. Methodological heterogeneity and clinical heterogeneity have important practical significance. For methodological heterogeneity, if the results of high-quality studies are different from those of low-quality studies, the pooled results and conclusions should be mainly or even completely based on high-quality studies. For clinical heterogeneity, the existence of heterogeneity is similar to the existence of interaction or effect modification, which has medical decision-making significance.

2.1.1 Methodology of heterogeneity testing

If the variation between studies is entirely caused by probabilistic factors, the size of the variation is limited. Therefore, with a certain number of studies, there is sufficient confidence (such as 99.9%) that the observed overall variation should be less than a certain upper limit. If the actual variation observed is greater than this upper limit, it indicates that important clinical and/or methodological heterogeneity may exist, and further analysis of the causes of heterogeneity is required. Otherwise, there is no sufficient reason to consider clinical and methodological heterogeneity exists, and it is believed that the variation is mainly caused by probabilistic factors. It would be reasonable to use meta-analyses to pool data. The significance testing used to measure the size of the heterogeneity of a set of studies and to estimate whether it is entirely due to probability or not are called heterogeneity testing, including Q test, I^2 test, and effect size tests, etc. (Tang and Mao, 2015; Tang and Yang, 2015; Mao, 2019; Zhang, 2023).

(1) *Q* test

In the *Q* test, the measured total variation of a set of studies is

$$Q = \sum_{i=1}^{k} w_i(\theta_i - \theta)^2 \quad (1)$$

where θ is the pooled effect size, θ_i is the effect size of the *i*th study, and w_i is the weight of the *i*th study, *i*=1, 2,…,*k*. In heterogeneity testing, the inverse-variance (I-V) method is used to calculate w_i.

Q follows the χ^2 distribution with *k*-1 degrees of freedom, where *k* is the number of studies included. Therefore, if the actual total variation is entirely caused by probabilistic factors, that is, there is no clinical and methodological heterogeneity, then there is 100(1-α)% certainty that the actual total variation $Q \leq \chi^2_\alpha(k\text{-}1)$. Conversely, if $Q > \chi^2_\alpha(k\text{-}1)$, the clinical and/or methodological heterogeneity is likely to exist.

In meta-analyses, the number of studies included is usually small. In this case, the power of *Q* test is low and false negative errors are prone to occur, that is, the actual heterogeneity is missed. In order to improve the testing power, it is generally that we set α to 0.10 instead of the commonly used 0.05. If $Q > \chi^2_\alpha(k\text{-}1)$, then $P \leq \alpha$, indicating possible clinical and/or methodological heterogeneity (but the cause of heterogeneity may not be found); if $Q \leq \chi^2_\alpha(k\text{-}1)$, then $P > \alpha$, indicating that there may not be (or there is not enough evidence to show that there is) important clinical and/or methodological heterogeneity (but the possibility of true heterogeneity cannot be ruled out). The larger the *Q* value, the smaller the corresponding *P* value, indicating the greater the possibility of clinical and/or methodological heterogeneity between studies. The precision of the *Q* test depends on the sample size *n* of the *i*th individual study and the number of studies *k*.

(2) I^2 test

The size of heterogeneity can also be expressed by I^2 (heterogeneity index, signal content index, or relative amount of heterogeneity). I^2 is the percentage of heterogeneity caused by non-probabilistic factors that accounts for the actual total variation. It can be simply regarded as the the proportion of variation that cannot be explained by the sampling error.

(a) I^2 statistics-I

$$I^2 = (Q - (k\text{-}1))/Q \times 100\% \quad (2)$$

where *Q* is the measured variation, and *k* is the total number of studies included. The I^2 statistic-I is corrected by the degree of freedom and is not affected by the number of studies included, and is suitable for meta-analyses with different numbers of studies.

When $Q < k\text{-}1$, that is, $I^2 < 0$, then let $I^2 = 0$. If $I^2 = 0$, it means that the total variation of the actual measurement is mainly caused by probabilistic errors, and there may be no heterogeneity caused by non-probabilistic factors; the larger I^2, indicates the greater heterogeneity caused by non-probabilistic factors, and the greater possibility of clinical and/or methodological heterogeneity. When the I^2 statistic-I is approximately 25%, 50% or 75%, respectively, it indicates low, moderate or high heterogeneity respectively. If the sample size is larger, I^2 will be closer to 100%.

The testing power or sensitivity of I^2 is higher than that of *Q* test, and its sampling error will be affected by the sample size *n*. I^2 test is most commonly used in psychology and medicine.

(b) I^2 statistics-II (Higgins and Thompson, 2002; Higgins et al., 2003; Huang, 2020)

$$I^2 = \tau^2 / (\tau^2 + \sum_{i=1}^{k} s_{\theta_i}^2 / k) \quad (3)$$

where τ^2 is the between-study variance of effect sizes, and $s_{\theta_i}^2$ is the variance of effect size for *i*th study. When the I^2 statistic-II is approximately 25%, 50% or 75%, respectively, it indicates low, moderate or high heterogeneity respectively.

(3) CH test

The heterogeneity indices I^2 above represent the percentage of variation across studies due to heterogeneity (Higgins and Thompson, 2002; Higgins et al., 2003). Nevertheless, they are not the absolute measure of heterogeneity (Borenstein et al., 2017), nor are they a measure of the strength of heterogeneity (Huang, 2023). The strength of heterogeneity, relative to the pooled effect (θ), can be measured by coefficient of heterogeneity (CH; or between-study coefficient of variation) (Takkouche et al., 1999; Huang, 2023)

$$CH = \tau^2 / |\theta| \quad (4)$$

Approximately $CH<1$ means no significant heterogeneity, and $CH>1$ means there is the significant heterogeneity. The larger *CH* means the stronger heterogeneity.

(4) Effect size tests

(a) Cohen's *d* (Cohen, 1988, 2008; Zhang, 2023)

$$d_{ij} = |\theta^{(j)} - \theta^{(i)}| / \sqrt{(s_{\theta^{(i)}}^2 + s_{\theta^{(j)}}^2)/2} \quad (5)$$

$i, j=1,2,\ldots,k;\ i \neq j$

where $\theta^{(i)}$ and $\theta^{(j)}$ are the effect size of study *i* and study *j* respectively, and $s_{\theta^{(i)}}^2$ and $s_{\theta^{(j)}}^2$ are the variances of the effect size of study *i* and study *j*, respectively.

$d_{ij}<0.1$, no heterogeneity (effect difference) between studies *i* and *j*; $0.1 \leq d_{ij}<0.3$, small heterogeneity (effect difference) between studies *i* and *j*; $0.3 \leq d_{ij}<0.5$, intermediate heterogeneity (effect difference) between studies *i* and *j*; $d_{ij} \geq 0.5$, large heterogeneity (effect difference) between studies *i* and *j* (Zhang, 2023; Zhang and Qi, 2024).

Calculate the average d_{ij} and maximum d_{ij}: d = max $\{d_{ij}\}$, d = average $\{d_{ij}\}$. $d<0.1$, no between-study heterogeneity; $0.1 \leq d<0.3$, small heterogeneity; $0.3 \leq d<0.5$, intermediate heterogeneity; $d \geq 0.5$, large heterogeneity.

(b) Hedges' *g*

When the sample size is small, such as when the overall sample is less than 20 or samples of each study is less than 10, Cohen's *d* will have a large deviation. For this, Hedges and Olkin (1985) present a method to calculate Cohen's *d* based on small samples (Zhang, 2023):

$$g_{ij} = (|\theta^{(j)} - \theta^{(i)}|/s_{ij})*(1 - 3/(4*(n_i + n_j - 2) - 1)) \quad (6)$$

$i, j=1,2,\ldots,k;\ i \neq j$

where

$$s_{ij} = \sqrt{((n_i - 1)\, s_{\theta^{(i)}}^2 + (n_j - 1)\, s_{\theta^{(j)}}^2)/(n_i + n_j - 2)}$$

and n_i: the sample size of the ith study, i=1,2,…,k.

g_{ij}<0.1, no heterogeneity (effect difference) between studies i and j; 0.1≤g_{ij}<0.3, small heterogeneity (effect difference) between studies i and j; 0.3≤g_{ij}<0.5, intermediate heterogeneity (effect difference) between studies i and j; g_{ij}≥0.5, large heterogeneity (effect difference) between studies i and j (Zhang, 2023; Zhang and Qi, 2024).

Calculate the average g_{ij} and maximum g_{ij}: g = max {g_{ij}}, g = average {g_{ij}}. g<0.1, no between-study heterogeneity; 0.1≤g<0.3, small heterogeneity; 0.3≤g<0.5, intermediate heterogeneity; g≥0.5, large heterogeneity.

It should be noted that the effect size tests above can also be used to detect the difference significance of between pooled effect sizes (calculated from the following estimators).

2.1.2 Processing heterogeneity

After conducting heterogeneity testing, if the significant between-study heterogeneity is found, the heterogeneity can be dealt with from several aspects (Tang and Mao, 2015).

(1) Correct data errors

Sometimes, heterogeneity may be caused by errors in data extraction or intermediate calculations. For example, for continuous variables, if standard errors are used as standard deviations, the confidence intervals for the effect size of each study included can become very narrow, resulting in little overlap in the confidence intervals between studies, creating the false illusion of heterogeneity.

(2) Change effect measures

The choice of effect measures is closely related to heterogeneity, for example, for continuous variables, when different studies use different outcomes or different measures of the same outcome in effect measurement. For example, choosing the mean difference instead of the standardized mean difference as the effect measure may mistakenly cause great heterogeneity. For binary variables, the chance of heterogeneity when using odds ratios and rate ratios is much smaller than the rate differences.

(3) Analyze the sources of heterogeneity

After taking the above procedures, if heterogeneity still exists, further procedures need to be used to analyze the source of heterogeneity. As mentioned above, the essence of heterogeneity is interaction or effect modification, so the analysis strategy is similar. Analysis methods mainly include subgroup analysis, meta-regression, and sensitivity analysis. At the same time, it is generally required to conduct analysis only around pre-set factors that may affect the effect size, rather than performing post hoc analysis after knowing the study results, because post hoc analysis may produce false positive results. .

(I) Subgroup analysis

According to the properties of studies, the studies can be divided into different groups. For studies within the same group, the meta-analysis is performed to estimate the overall pooled results and compare whether there are differences in the pooled results of different groups. This method is called subgroup analysis. Subgroup analysis should not be data-driven, but should be performed on the premise of knowing which subgroups may be different and making assumptions in advance. It can be considered that subgroup analysis is a kind of ANOVA under meta-analyses. Simply speaking, it uses the fixed-effects model to compare between-subgroup effect sizes (subgroups can be named in advance based on data or actual conditions). The random-effects model is still used within the subgroup. Therefore, this type of method is called the mixed-effects model. Zhang and Qi (2024) has developed the ANOVA methodology and computational tools in the paradigm of new statistics (http://www.iaees.org/publications/journals/ces/articles/2024-14(1)/ANOVA-nSTAT.htm), which can be used in subgroup analysis.

In subgroup analysis, the key is the factors used to grouping the groups. Grouping factors are the reasons

that may cause heterogeneity between studies. The causes for between-study heterogeneity are diverse, mainly including clinical factors and methodological factors. Methodological factors include study type (such as randomized controlled trial or cohort study) and bias control measures (such as clinical trial outcome measurement method, grouping method, group concealment, blinding, follow-up rate, intention-to-treat analysis, etc). Clinical factors are mainly related to PICOS, such as the patient's gender, age, severity of illness, etc., as well as the route of administration, dosage, total course of treatment, etc., as well as the selection of outcome measures and the quality of treatment conditions, etc. Researchers must analyze many possible causes, propose one or several factors that are most likely to cause heterogeneity, and then conduct subgroup analysis only for one or several predetermined factors. Subgroup analysis of unplanned, unpurposeful, untargeted and all possible factors based after data collection should be avoided as much as possible, because non-prespecified, blind subgroup analysis is likely to lead to false positive results, especially when analyzing many factors (Tang and Mao, 2015).

Subgroup analysis is generally based on subgroups produced by the properties of the original studies. For example, the studies are divided into two groups according to the average age of the study subjects or the proportion of males, and the efficacy of each group is estimated separately, and the results are compared. Or, subgroup analysis is based on the results of the subgroup analyses reported in the same original studies. For example, many of the original studies reported separately the true effect of the treatment in the male subgroup and the true effect in the female subgroup, and the treatment effect in the male and female subgroups can be pooled separately and can be compared.

The subgroup analysis based on the internal subgroups of the original studies is called paired comparison or direct comparison; the subgroup analysis based on the overall properties of the studies is called unpaired comparison or indirect comparison; the subgroup analysis based on mixed data is called mixed subgroup analysis. Unpaired subgroup analysis cannot eliminate the impact of differences in factors other than grouping factors between studies on subgroup comparisons, while the advantage of paired subgroup analysis is precisely to eliminate the impact of these factors, using data from direct comparisons of studies. Therefore, paired subgroup analysis is better than unpaired subgroup analysis. Factors other than grouping factors can be divided into two types. One is various factors that affect the treatment effect between studies, such as patient characteristics and diagnosis and treatment conditions; the other is differences in patient characteristics within the study. Even in paired subgroup analysis, the impact of differences in factors other than the grouping factors within the study on subgroup comparisons cannot be ruled out, because the grouping factors are not randomly produced.

When using subgroup analysis to determine whether heterogeneity exists, the more of the following are confirmed, the greater the chance of heterogeneity (Tang and Mao, 2015)

(a) Subgroup analysis is proposed when formulating the meta-analysis plan rather than during the analysis process.

(b) The differences between subgroups are one of the few factors tested (the greater the number of hypotheses tested, the greater the probability of finding differences between subgroups due to chance).

(c) The differences between subgroups are found based on direct comparisons rather than indirect comparisons.

(d) The differences between subgroups are large enough.

(e) The differences between subgroups are statistically significant.

(f) The differences between subgroups are consistent among different studies (if multiple original studies find differences between subgroups and the sizes are consistent, it means that the differences between subgroups are more consistent across multiple studies, which can increase the results' credibility).

(g) The differences between subgroups are supported by external evidences.

The disadvantage of subgroup analysis is that it cannot effectively rule out the possibility of other factors (i.e., non-grouping factors) causing differences between subgroups (i.e., confounding effects).

Borenstein et al. (2011) suggested that the minimum unit for subgroup analysis is 10 articles per subgroup. If there are less than 10 articles, subgroup analysis is meaningless.

(II) Meta-regression

Subgroup analysis is ANOVA in the context of meta-analyses, and ANOVA is essentially a special form of OLS regression. Therefore, meta-regression can be simply understood as extending subgroup analysis to the regression scenario, i.e., using regression methods to predict effect size.

Meta-regression is the weighted regression analysis based on collective data. In the meta-regression model, the dependent variable is the point estimate of the study's effect size, the independent variables are factors used to explain heterogeneity, and the weight is the weight given to each study in the meta-analysis. .

If the regression model does not contain independent variables, the weight is the weight of the inverse-variance method in the fixed-effects model. The constant value of the linear regression equation is equal to the pooled effect of the meta-analysis. The standard error of the constant and 100(1-α)% confidence interval are the same as results of the meta-analysis.

If there is statistically significant heterogeneity between studies, the weight of the random-effects model should be used. If the source of heterogeneity needs to be analyzed, the suspected factors causing the heterogeneity can be used as independent variables and included in the meta-regression equation. In the meta-regression with independent variables, random-effects models are generally used, and only the inverse-variance method can be used. Same as general linear regression analysis, the effect size of the dependent variable should conform to or be close to the normal distribution, and each study should be independent of each other and have homogeneous variances. Otherwise, it will increase the risk of statistical test error I. The random-effects model only requires that the effect size of the dependent variable is or is close to a normal distribution.

If k studies are included in a meta-regression and n suspected heterogeneity factors are explored, the general linear regression equations of the fixed-effects model and the random-effects model are respectively

$$\theta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2\mathrm{i}} + \cdots + \beta_n x_{ni} + \varepsilon_i$$
$$\theta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2\mathrm{i}} + \cdots + \beta_n x_{ni} + \varepsilon_i + \tau_i \qquad (7)$$

where θ_i is the estimated effect size of the ith study, x_1, …, x_n are suspected heterogeneity factors, β_0 is a constant, β_1, …, β_n are the partial regression coefficients of each suspected heterogeneity factor, and τ_i is the between-study error of the ith study, ε_i is the within-study error of the ith study, i=1,2,…,k.. The β of the meta-regression is tested for significance using the t-test. The overall goodness of fit of the model is tested using R^2.

In addition to general linear regression, meta-regression can use the stepwise regression method (Qi et al., 2016; Zhang, 2016), which can be called meta-stepwise regression.

Subgroup analysis can handle one heterogeneous factor at a time, while meta-regression can handle multiple heterogeneous factors at the same time, and can control the confounding effects of other factors included in the equation. It is worth noting that the number of studies included in the regression analysis cannot too few, preferably no less than 10. Since the number of studies included in a meta-analysis is generally small, the heterogeneity factors included in the equation should generally not exceed 3 to 5.

(III) Sensitivity analysis

In meta-analyses, sensitivity analysis refers to exclude certain studies one by one and observe the changes of pooled effect size. After deleting a certain study or a few studies, the heterogeneity among the remaining studies is greatly reduced, indicating that the deleted studies are most probably the sources of heterogeneity.

(4) Use random-effects model for pooling studies

(5) Abandon the meta-analysis

If the number of studies is small and the heterogeneity between studies is large, such as the direction of effects is obviously inconsistent, or the confidence intervals do not overlap, and there are important differences in PICOS between studies but the heterogeneity cannot be explained by subgroup analysis or regression analysis, retc., then it is better to abandon the meta-analysis and describe different studies separately.

2.2 Fixed-Effects Model and Random-Effects Model

When pooling effect sizes in a meta-analysis, there are two models we can use, the fixed-effects model and the random-effects model (Borenstein et al., 2011). Compared with fixed-effects models, random-effects models have the disadvantage of being more concerned about small studies with bias issues (Schwarzer, Carpenter, and Rücker, 2015). Some researchers advocate the preferred fixed-effects model (Poole and Greenland, 1999; Furukawa et al., 2003). It has been recommended to only resort to the random-effects model in clinical psychology and the health sciences (Cuijpers et al., 2016).

In my view, it should be advocated to make a choice between the random-effects model and the fixed-effect model based on the actual situation (refers to theory-driven, rather than simply judging the model used based on a single heterogeneity test). When it is determined that all studies are homogeneous, that is, all studies come from the same population, the fixed-effects model can be used for data analysis, otherwise the random-effects model should be used. Although most researchers advocate the use of theory-driven methods to choose between the fixed-effect model and the random-effects model, for novices, a common data-driven approach is to conduct heterogeneity analysis and assure that the studies in heterogeneity analysis have multiple subgroups, that is, they are derived from different populations. In this case, a random-effects model needs to be used.

There is a growing consensus that the random and mixed-effects models should be preferred over the decidedly more simple fixed-effects model (Erez et al., 1996; Hunter and Schmidt, 2000).

2.2.1 Fixed-effects model

The fixed-effects model can be described by

$$\theta_i = \theta + \epsilon_i,\ i=1,2,\ldots,k \qquad (8)$$

where

$$\epsilon_i \sim N(0, s_{\theta_i}^2),\ \theta_i \sim N(\theta, s_{\theta_i}^2)$$

The fixed-effects model is suitable for pooling studies that are homogeneous. If there is no heterogeneity between stdies and the differences between studies are caused only by sampling error and fluctuate randomly around the true value, the fixed-effects model can be used. The fixed-effects model requires that the pooled studies have the same true value, that is, they all come from the same population. The weighted averaging can well reflect the true value.

The weighted averaging methods commonly used in fixed-effects models, which include the sample size weighting method, the inverse-variance weighted averaging method (I-V method), the Mantel-Haenszel

method (M-H method), and the Peto method. Commonly used methods include the I-V method, etc., which can be used for all statistical measures that can estimate standard error; M-H method is mostly used for effect measures based on dichotomous variables, including odds ratios, rate ratios and rate differences; Peto method can only be used for pooling Peto odds ratios based on dichotomous variables and risk ratios in survival data, which requires the numbers of subjects in the comparison groups are close and the true odds ratio is close to 1. Strictly speaking, the sample size weighting method is also a general method, but it will only be considered when any of the above methods is inappropriate. When individual studies with too small sample sizes and the number of outcomes in the group is equal to or close to zero, both the Mantel-Haenszel method and the Peto method are better than the inverse-variance method (Tang and Mao, 2015; Tang and Yang, 2015; Mao, 2019). In addition, an inverse-variance - sample size hybrid weighted averaging method is proposed in this study.

(1) Sample size weighted averaging

If all studies are from the same population and there is no bias, the estimated effect sizes of different studies are different, and the average of the results of large-sample studies is closer to the true value than the results of small-sample studies. Different effects can be assigned weights according to the sample size of the study. The larger the sample size, the greater the weight given, and vice versa. The average calculated in this way is the weighted averaging. The weighted averaging is directly affected by the sample size and reflects more the results of large sample studies. In other word, the method using sample size weighting is called the sample size weighted averaging.

(2) Inverse-variance weighted averaging

If all studies come from the same population and there is no bias, the difference between the studies and the true value is directly inversely proportional to its variance, and indirectly proportional to the sample size through the variance. In fact, studies with small sample size exhibit more heterogeneity than that with large sample sizes (IntHout et al., 2015). Therefore, a common method of meta-analyses is to conduct it based on the variance of the effect. This weighting method is called the inverse-variance weighted averaging. This method can be used for any outcome or effect measure. It is the most commonly used method in meta-analyses and is called the universal weighted averaging. The inverse-variance method can be applied to pool effect sizes of a variety of data types study designs. In addition to the aforementioned dichotomous variables and continuous variables, it can also be used to pool standardized mortality ratios, hazard ratios, diagnostic tests, and effect measures in crossover trials and group randomized trials. It can also be used for pooling single group rates and means.

In the inverse-variance weighted averaging (inverse-variance method, or I-V method), the inverse of the variance of the effect size is the weight of the corresponding study. Let θ_i be the effect size of the *i*th study, which can be the logarithm of rate ratio or odds ratio, or rate difference, mean difference or standardized mean difference, etc., then the pooled effect size is (Tang and Mao, 2015; Tang and Yang, 2015; Mao, 2019)

$$\theta_{\mathrm{IV}} = \sum_{i=1}^{k} w_i\theta_i / \sum_{i=1}^{k} w_i \qquad (9)$$

where w_i is the weight of *i*th study, i.e., the inverse of the squared standard error of the effect estimate

$$w_i = 1/s_{\theta_i}^2 \qquad (10)$$

Since the standard error is inversely proportional to the sample size, compared to studies with a smaller sample size, studies with a larger sample size tend to have larger weights when pooled because of smaller standard errors.

The standard error of the pooled effect size is calculated as follows

$$s_{\theta_{\mathrm{IV}}} = 1/\sqrt{\sum_{i=1}^{k} w_i} \quad (11)$$

The 100(1-α)% confidence interval (confidence interval, CI) is

$$\theta_{\mathrm{IV}} \pm z_{\alpha} s_{\theta_{\mathrm{IV}}} \quad (12)$$

where *z*-value rather than *t*-value is used due to the limitation of *t*-tests (Huang, 2018; Zhang, 2022b; the same as the following statements).

As a general meta-analysis method, the inverse-variance method can be used for all outcomes and effect measures. For the calculation of the standard errors (s_{θ_i}) of other commonly used effect measures, just see Table 1.

Table 1 Calculation methods for standard error (s_{θ_i}) of commonly used effect measures (Tang and Mao, 2015).

Effect measure	Standard error (s_{θ_i})
Mean Difference (MD)	$s_{MD_i} = \sqrt{SD_{1\mathrm{i}}^2/n_{1\mathrm{i}} + SD_{2\mathrm{i}}^2/n_{2i}}$
Odd Ratio (OR)	$s_{\ln(OR_i)} = \sqrt{1/a_{\mathrm{i}} + 1/b_{\mathrm{i}} + 1/c_{\mathrm{i}} + 1/d_{\mathrm{i}}}$
Risk Ratio (RR)	$s_{\ln(RR_i)} = \sqrt{1/a_{\mathrm{i}} + 1/c_{\mathrm{i}} + 1/n_{1\mathrm{i}} + 1/n_{2\mathrm{i}}}$
Risk Difference (RD)	$s_{RD_i} = \sqrt{a_{\mathrm{i}} b_{\mathrm{i}}/n_{1i}^3 + c_{\mathrm{i}} d_{\mathrm{i}}/n_{2i}^3}$

SD_1 and SD_2 are the standard deviations of the experimental group and the control group respectively; n_1 and n_2 are the sample sizes of the experimental group and the control group respectively; a_i and c_i are the number of outcome events in the experimental group and the control group respectively; $b_i$and d_i are the number of non-outcome events in the experimental group and the control group respectively.

(3) Inverse-variance - sample size hybrid weighted averaging

Although in general, variance is inversely proportional to sample size (Zhang, 2024). However, when the number of studies included is small, the inverse relationship between variance and sample size will be weaker. In addition, some studies may show disconnection in results reliability and variance-sample size relationship. For example, some studies have large sample sizes and reliable results, but the variances may also be large, and the inverse-variance weighted averaging can easily weaken the results of these studies; while some studies have small sample sizes and unreliable results, but the variances are also small. The inverse-variance weighted averaging can easily strengthen the results of these studies. Considering that the number of studies included in meta-analyses is generally small (Turner et al., 2012; Kontopantelis et al., 2013), the weakening of the inverse relationship and the disconnection of results may significantly affect the pooled effect size. At this time, both the variance and the sample size should be taken into account. To this end, I propose here the inverse-variance - sample size hybrid weighted averaging (IVSS). The principle of this method is to arithmetically combine the standardized inverse-variance weighted averaging and the standardized sample size weighted averaging.

$$\theta_{\mathrm{IS}} = \sum_{i=1}^{k} w_i \theta_i / \sum_{i=1}^{k} w_i \quad (13)$$

where

$$w_i = (1/s_{\theta_i}^2)/\sum_{i=1}^{k}(1/s_{\theta_i}^2) + n_i/\sum_{i=1}^{k} n_i \quad (14)$$

and θ_i is the effect size of the *i*th study, and n_i is the sample size of *i*th study. The standard error of the pooled effect size θ_{IS}, and the confidence interval are the same to eqs (10) to (12).

(4) Mantel-Haenszel method

The Mantel-Haenszel method (M-H method) is a stratified analysis suitable for pooling data whose outcomes are categorical variables (Mantel and Haenszel, 1959). In this method, each independent study is equivalent to one layer. Estimate the effect of each layer and use the consistency test of M-H method to check the heterogeneity of the results of each layer. If the difference in the effect sizes of each layer is not statistically significant, the M-H method can be used to pool the results of each layer to obtain the overall effect size of the pooled layers.

Let θ_i be the effect size of the *i*th study, such as odds ratio, rate ratio and rate difference, and w_i be the weight of the study, then the pooled effect size of Mantel-Haenszel method is

$$\theta_{\mathrm{MH}} = \sum_{i=1}^{k} w_i\theta_i/\sum_{i=1}^{k} w_i \quad (15)$$

The weight of the M-H method and the standard error of the pooled effect size are calculated using different methods due to different effect measures, as shown in Table 2.

Table 2 Weights and standard errors of pooled effect values of the M-H method (Tang and Mao, 2015).

	OR	RR	RD
θ_{MH}	$OR_{\mathrm{MH}} = \sum_{i=1}^{k} w_i OR_i/\sum_{i=1}^{k} w_i$	$RR_{\mathrm{MH}} = \sum_{i=1}^{k} w_i RR_i/\sum_{i=1}^{k} w_i$	$RD_{\mathrm{MH}} = \sum_{i=1}^{k} w_i RD_i/\sum_{i=1}^{k} w_i$
$s_{\theta_{\mathrm{MH}}}$	$s_{\ln(OR_{\mathrm{MH}})} = \sqrt{\frac{1}{2}\left(\frac{E}{R^2} + \frac{F+G}{R\times S} + \frac{H}{S^2}\right)}$	$s_{\ln(RR_{\mathrm{MH}})} = \sqrt{\frac{P}{U\times V}}$	$s_{RD_{\mathrm{MH}}} = \sqrt{\frac{J}{K^2}}$
where	$w_i = \frac{b_i c_i}{N_i}$	$w_i = \frac{c_i n_{1i}}{N_i}$	$w_i = \frac{n_{1i} n_{2i}}{N_i}$
	$R = \sum \frac{a_i d_i}{N_i}, S = \sum \frac{b_i c_i}{N_i}$	$P = \sum \frac{n_{1i} n_{2i} m_{1i} - a_i c_i N_i}{N_i^2}$	$J = \sum \frac{a_i b_i n_{2i}^3 + c_i d_i n_{1i}^3}{n_{1i} n_{2i} N_i^2}$
	$E = \sum \frac{(a_i+d_i) a_i d_i}{N_i^2}$, $F = \sum \frac{(a_i+d_i) b_i c_i}{N_i^2}$	$U = \sum \frac{a_i n_{2i}}{N_i}$	$K = \sum \frac{n_{1i} n_{2i}}{N_i}$
	$G = \sum \frac{(b_i+c_i) a_i d_i}{N_i^2}$, $H = \sum \frac{(b_i+c_i) b_i c_i}{N_i^2}$	$V = \sum \frac{c_i n_{1i}}{N_i}$	-
CI of $\ln(\theta_{\mathrm{MH}})$	$\ln(OR_{\mathrm{MH}}) \pm z_\alpha s_{\ln(OR_{\mathrm{MH}})}$	$\ln(RR_{\mathrm{MH}}) \pm z_\alpha s_{\ln(RR_{\mathrm{MH}})}$	-
CI of θ_{MH}	$\exp(\ln(OR_{\mathrm{MH}}) \pm z_\alpha s_{\ln(OR_{\mathrm{MH}})})$	$\exp(\ln(RR_{\mathrm{MH}}) \pm z_\alpha s_{\ln(RR_{\mathrm{MH}})})$	$RD_{\mathrm{MH}} \pm z_\alpha s_{(RD_{\mathrm{MH}})}$

For *i*-th study, suppose there are two groups A and B, and the total cases in groups A and B are n_{1i} ($=a_i+b_i$) and n_{2i} ($=c_i+d_i$) respectively, and the total cases with and without the outcome event are m_{1i} ($=a_i+c_i$) and m_{2i} ($=b_i+d_i$) respectively, where a_i and c_i are the cases with the outcome event respectively, and b_i and d_i are the cases without the outcome event respectively. $N_i = n_{1i} + n_{2i} = m_{1i} + m_{2i}$. *z*-value rather than *t*-value is used for the limitation of *t*-tests (Huang, 2018; Zhang, 2022b), e.g., for α=0.05,z_α=1.96.

(5) Peto method

Like the M-H method, the Peto method is often used for stratified analysis of data (Brockhaus et al., 2014). The Peto method is generally used for pooling odds ratios, and can also be used for pooling hazard ratios (HR) in survival data. When using the Peto method, all calculations must be based on logarithms (Tang and Mao, 2015; Tang and Yang, 2015; Mao, 2019).

The pooled effect size, OR_{Peto}, of the Peto method, and the standard error are calculated as follows

$$\ln(OR_{\text{peto}}) = \frac{\sum O_i - \sum E_i}{\sum V_i} \quad (16)$$

$$s_{\ln(OR_{\text{peto}})} = 1/\sqrt{\sum v_i} \quad (17)$$

where

$$O_i = a_i$$
$$E_i = (a_i+b_i)(a_i+c_i)/N_i$$
$$V_i = (a_i+b_i)(a_i+c_i)(c_i+d_i)(b_i+d_i)/((N_i\text{-}1)N_i^2)$$

The CIs of $\ln(OR_{\text{peto}})$ and OR_{peto} are

$$\ln(OR_{\text{peto}}) \pm z_\alpha s_{\ln(OR_{\text{peto}})} \quad (18)$$

and

$$\exp(\ln(OR_{\text{peto}}) \pm z_\alpha s_{\ln(OR_{\text{peto}})}) \quad (19)$$

respectively.

Some research argued that Peto's odds ratio method may yield inconsistent results and has no advantage over classic odds ratios in meta-analysis (Xu et al., 2020; Schwarzer et al., 2021).

Recommended usage for the various weighted averaging methods mentioned above include

(1) When pooling RR and RD in binary variables, the I-V method, the IVSS method, and the M-H method can be used.

(2) When pooling OR in binary variables, the I-V method, the IVSS method, the M-H method and the Peto method can be used.

(3) When pooling WMD (Weighted Mean Difference) and SMD (Standardized Mean Difference) in continuous variables, the I-V method and the IVSS method can be used.

(4) When pooling HR values in survival data, the I-V method, the IVSS method and the Peto method can be used.

2.2.2 Random-effects model

The random-effects model assumes that the variation in effect size estimates, drawn from a set of studies, can be divided into two parts, between-study heterogeneity (τ^2) and sampling variance. The random-effects model can be described by

$$\theta_{ei} = \theta_i + \epsilon_i = \mu_\theta + \tau_i + \epsilon_i,\ i=1,2,\ldots,k \quad (20)$$

where

$$\tau_i \sim N(0,\tau^2),\ \epsilon_i \sim N(0,s_{\theta_i}^2),\ \theta_{ei} \sim N(\mu_\theta,\tau^2+s_{\theta_i}^2)$$

Due to differences in factors such as environmental locations, inclusion criteria, measurement methods, etc., there will be differences in the effect sizes between studies, and the true value of the effect size will be different (Viechtbauer, 2005; Korn et al., 2013; Gagne et al, 2014). The measure to quantify this variation is the between-study heterogeneity variance, or between-study variance (heterogeneity, between-study variance, or the variance of the distribution of true effect sizes) τ^2. Therefore, the meta-analysis should take the between-study variance τ^2 into account when calculating weights (Clinical Research and Medical Statistics, 2023)

$$w_i = 1/(s_{\theta_i}^2 + \tau^2),\ i=1,2,\ldots,k \qquad (21)$$

The model that takes into account between-study variation is the random-effects model. The random-effects model is suitable for pooling studies with heterogeneity (Tang and Mao, 2015).

When τ^2 is zero, the weight of the random-effects model is equal to the weight w_i in the inverse-variance method under the fixed-effect model. When τ^2 is not zero, the weight of the random-effects model will be smaller than the weight w_i of the fixed-effects model, and the standard error of the pooled result will be larger than that of the fixed-effects model. Therefore, the point estimate of a random-effects model is generally closer to that of a fixed-effects model, and the confidence interval is generally wider than that of a fixed-effects model. Since τ^2 is a fixed value that is assigned equally to all studies included, so it will reduce the relative difference in weights between studies, increase the relative weight of small sample studies, and reduce the relative weight of large sample studies. The larger τ^2, the smaller the relative difference in weights between studies. τ^2 is not very sensitive to the number of studies and sample size.

The random-effects model assumes that there is not only one true effect size, but also a distribution of true effect sizes. Therefore, the goal of random-effects models is not to estimate one true effect size across all studies, but rather the mean of the true effect's distribution. The random-effects model assumes that the true values of the pooled studies are different, they come from different populations, the results are heterogeneous, the differences between the results are caused by two factors, sampling error and true differences, and they fluctuate randomly around the true values.

The random-effects model increases the relative weight of studies with small sample sizes. Therefore, if there is bias in small sample studies, such as low methodological quality or selective publication of positive results, then the random-effects model will increase the impact of this bias. At this time, we can use subgroup analysis (see above), or conduct sensitivity analysis for small sample size studies to ensure that small sample sizes have little impact on the results of the random-effects model.

A common approach to inference within the random-effects model is to first estimate the heterogeneity variance τ^2, and subsequently estimate the effect (Röver et al., 2015).

Generally, for an estimator from the random-effects model, the pooled effect size is given by

$$\theta = \sum_{i=1}^{k} w_i\theta_i / \sum_{i=1}^{k} w_i \qquad (22)$$

where θ_i is the effect size of the *i*th study, and θ is the pooled effect size. The confidence (coverage) interval is given by

$$\theta \pm z_\alpha s_\theta \tag{23}$$

and the standard error s_θ is

$$s_\theta = 1/\sqrt{\sum w_i}$$

A general form for the standard error of θ is given by (Huang, 2023)

$$s_\theta = \sqrt{\sum_{i=1}^{k} w_i(\theta_i - \theta)^2 / \sum_{i=1}^{k} w_i}\,/(C_{4,k}\sqrt{k-1}) \tag{24}$$

where

$$C_{4,k} = \sqrt{\frac{2}{k-1}}\,\Gamma(\frac{k}{2})/\Gamma(\frac{k-1}{2})$$

Some commonly used estimators are given in the following (Hartung and Knapp, 2001a-b; Viechtbauer, 2005; Sidik and Jonkman, 2007; Panityakul et al., 2013; Röver et al., 2015; Veroniki et al., 2015)

(1) DerSimonian-Laird estimator

The DerSimonian-Laird method (DerSimonian and Laird, 1986) is a commonly used method for the random-effects model. Its weight is given by

$$w_i = 1/(s_{\theta_i}^2 + \tau^2)$$

where $s_{\theta_i}^2$ is the variance of *i*th study, and $\tau^2$is between-study variance, and

$$\tau^2 = \frac{\sum_{i=1}^{k}(\theta_i - \theta)^2/s_{\theta_i}^2 - (k-1)}{\sum_{i=1}^{k} 1/s_{\theta_i}^2 - \sum_{i=1}^{k} 1/s_{\theta_i}^4 / \sum_{i=1}^{k} 1/s_{\theta_i}^2} \tag{25}$$

The normal approximation based on τ^2 usually works well for many studies (large *k*) and small standard errors (small $s_{\theta_i}^2$), or negligible heterogeneity (small τ^2), but otherwise tends to be anticonservative (Hartung and Knapp, 2001a-b; Röver et al., 2015).

DerSimonian and Laird estimator is commonly used in medicine and psychology, but also widely criticized (Böhning et al., 2002).

(2) Hunter-Schmidt estimator

The weight of Hunter-Schmidt estimator (Hunter and Schmidt, 1990) is given by

$$w_i = 1/(s_{\theta_i}^2 + \tau^2)$$

where

$$\tau^2 = (\sum_{i=1}^{k} w_i(\theta_i - \theta)^2 - \sum_{i=1}^{k} w_i\, s_{\theta_i}^2)/\sum_{i=1}^{k} w_i \tag{26}$$

(3) Maximum-likelihood estimator (ML)
For ML, the weight is given by (Huang, 2023)

$$w_i = 1/(s_{\theta_i}^2 + \tau^2)$$

where

$$\tau^2 = \frac{\left(\sum_{i=1}^{k} w_i^2[(\theta_i - \theta)^2 - s_{\theta_i}^2\right]}{\sum_{i=1}^{k} w_i^2} \quad (27)$$

(4) Restricted maximum-likelihood estimator (REML)
For REML (Viechtbauer, 2005; Raudenbush, 2009), the weight is given by (Huang, 2023)

$$w_i = 1/(s_{\theta_i}^2 + \tau^2)$$

where

$$\tau^2 = \frac{\left(\sum_{i=1}^{k} w_i^2[(\theta_i - \theta)^2 - s_{\theta_i}^2\right]}{\sum_{i=1}^{k} w_i^2} + 1/\sum_{i=1}^{k} w_i \quad (28)$$

(5) Inverse-RSE estimator
The weight of inverse-RSE (root-squared error) estimator is given by (Huang, 2023)

$$w_i = 1/\sqrt{s_{\theta_i}^2 + \left(\theta_i - \theta\right)^2} \quad (29)$$

(6) Hartung-Knapp-Sidik-Jonkman (HKSJ) estimator
Hartung and Knapp (2001a-b) and Sidik and Jonkman (2002) independently introduced an estimator in which the weight is given by

$$w_i = 1/(s_{\theta_i}^2 + \tau^2)$$

where τ^2 is calculated using eq (28) in present study.

The adjusted confidence interval is

$$\theta \pm z_\alpha s_{\theta'}\sqrt{q} \quad (30)$$

where

$$s_{\theta'} = 1/\sqrt{\sum w_i}$$
$$q = \sum_i^k w_i(\theta_i - \theta)^2/(k - 1),\ s_\theta = s_{\theta'}\sqrt{q}$$

(7) Modified Knapp-Hartung (mKH) estimator
For the modified Knapp-Hartung (mKH) estimator (Röver et al., 2015), the weight is given by

$$w_i = 1/(s_{\theta_i}^2 + \tau^2)$$

and the modified confidence interval is given by

$$\theta \pm z_\alpha s_{\theta'}\sqrt{\max\{1, q\}} \quad (31)$$

where

$$s_{\theta'} = 1/\sqrt{\sum w_i}$$
$$q = \sum_{i=1}^{k} w_i(\theta_i - \theta)^2/(k - 1), \; s_\theta = s_{\theta'}\sqrt{\max\{1, q\}}$$

(8) Paule-Mandel estimator

The weight of Paule-Mandel estimator (Paule and Mandel, 1982; Rukhin et al., 2000) is given by (Huang, 2023)

$$w_i = 1/(s_{\theta_i}^2 + \tau^2)$$

where

$$\sum_{i=1}^{k}(\theta_i - \theta)^2/(s_{\theta_i}^2 + \tau^2) = k - 1 \quad (32)$$

Recommended usage for the above methods includes

(1) For effect sizes based on continuous outcome event data, the restricted maximum likelihood estimator can be used as a first choice.

(2) For binary effect size data, if there are no extreme changes in sample size, the Paule-Mandel estimator is the first choice when the number of studies is small (Bakbergenuly et al., 2020). A recent simulation study produced similar results but found that the Paule-Mandel estimator may be suboptimal when the sample size of the study changes drastically.

(3) If τ^2 is very large, and if avoiding false positives has a very high priority, the Sidik-Jonkman estimator (Sidik and Jonkman, 2007) can be used.

(4) If you want to replicate your results as accurately as possible, the DerSimonian-Laird estimator is the preferred method.

Whichever method is used, it should be supplemented with other methods.

The choice of effect measure depends on the purpose of the study. Taking randomized controlled trials that evaluate intervention effects as an example, if the result is a dichotomous variable, the effect size is expressed in various forms such as relative risk reduction percentage, absolute risk reduction value, and number of people needing treatment (Table 3; Tang and Mao, 2015).

Table 3 Suggested criteria for selection of effect measures (Tang and Mao, 2015).

Research purposes	Variable types	Available effect measures
Casual inference	Binary variable	Rate ratio, odds ratio, rate difference, etc.
	Continuous variable	Correlation coefficient, regression coefficient, between-mean difference, standardized between-mean difference, etc.
	Survival time	Risk ratio, survival rate ratio, survival time difference, etc.
Estimate averaged trend	Binary variable	Prevalence rate, incidence rate, survival rate, etc.
	Continuous variable	Mean, median, etc.
Assess diagnostic accuracy	Binary variable	Sensitivity, specificity, odds ratio, etc.
	Continuous variable	Area under ROC curve
	Nominal variable	Likelihood ratio for diagnosis

On causal inferences on different types of variables, see Zhang (2021a-c), and Antonelli and Cefalu (2020).

If the effect measure is incidence rate, and the incidence rate is too low or too high, it is recommended to consider making certain conversions. Usually the corresponding methods are (Clinical Research and Medical Statistics, 2023)

(1) Log conversion
(2) Squared root conversion
(3) Freeman-Tukey double inverse sine conversion
(4) No conversion

If the effect measure is a proportion, especially if the proportion is too low or too high, it is recommended to consider making certain conversions. Usually the corresponding methods are (Clinical Research and Medical Statistics, 2023)

(1) Logit arcsine transformation
(2) Arcsine conversion
(3) Freeman-Tukey double inverse sine conversion
(4) Log conversion
(5) No conversion

2.2.3 Averaged estimator

As noted by Huang (2023), all estimators above must be valid because they are developed from or supported by statistical principles. Nevertheless the heterogeneity variance given by different estimators for the same dataset may often differ significantly. Therefore, the choice of different estimators will affect conclusion of the meta-analysis. Antonelli and Cefalu (2020) argued that using as many estimators as possible into the list of candidate estimators is ideal, which helps to reduce the impact that some bad estimators can have, while increasing the efficiency of the averaged estimator. However there is no universally accepted metric for

determining which estimator is optimal among a set of candidate estimators. Naturally, a method called "estimator averaging" may be used to solve the estimator selection problem, which is aimed to provide a new and estimator that linearly combined of a set of individual estimators in order to provide a better estimate (Lavancier and Rochet, 2015, 2017; Mitra et al., 2019; Antonelli and Cefalu, 2020; Huang, 2023). Huang (2023) firstly proposed the estimator averaging methodology used in meta-analyses. He proved that the averaged estimator performs better than individual estimators in terms of bias and efficiency. Here I use the estimator-averaging method of Huang (2023) to pool the effects from individual estimators. First we assume that the biases of individual estimators are uniformly distributed about zero.

The averaged estimator for pooled effect sizes is the weighted average of *N* individual estimators, which is given by

$$\theta_A = \sum_{i=1}^{N} w_i \theta^{(i)} / \sum_{i=1}^{N} w_i \quad (33)$$

where w_i is the weight associated with the *i*th estimator $\theta^{(i)}$. The weights are given by

$$w_i = 1/s^2_{\theta^{(i)}}$$

where $s_{\theta^{(i)}}$ is the standard error for *i*th estimator $\theta^{(i)}$.

The standard error of pooled effect size is given by (Huang, 2023)

$$s_{\theta_A} = \sum_{i=1}^{N} (1/s_{\theta^{(i)}}) / \sum_{i=1}^{N} (1/s^2_{\theta^{(i)}}) \quad (34)$$

The 100(1-α)% confidence (coverage) interval is given by

$$\theta_A \pm z_\alpha s_{\theta_A}$$

If the estimators are heterogeneous, the between-estimator variance is given by

$$\tau_A^2 = \sum_{i=1}^{N} (1/s^2_{\theta^{(i)}}) \tau_i^2 / \sum_{i=1}^{N} (1/s^2_{\theta^{(i)}}) \quad (35)$$

and τ_i^2 is the between-study variance of *i*th estimator. And the 100(1-α)% confidence (coverage) interval is given by

$$\theta_A \pm z_\alpha s_A$$

where

$$s_A = \sqrt{\tau_A^2 + s^2_{\theta_A}} \quad (36)$$

3 Computaional Tools

In present study, I delveloped the computational tool, MetaAnaly, for meta-analysis. It includes both online (http://www.iaees.org/publications/journals/nb/articles/2024-14(2)/MetaAnaly.htm) and offline versions, and can be used for various computing devices (PCs, iPads, smartphones, etc.), operating systems (Windows, Mac, Android, Harmony, etc.) and web browsers (Chrome, Firefox, Sougo, 360, etc)(Fig. 1). It can be used in

experimental sciences such as medicine, biology, ecology, psychology, sociology, economy, physics and chemistry etc.

Both user manual guide and offline tool can be found at:

http://www.iaees.org/publications/journals/nb/articles/2024-14(2)/e-suppl/MetaAnaly.rar

Double-click the offline tool, it will be opened in the default web browser.

4 Reporting Criteria of Meta-analyses

After determining the research topic, if a comprehensive literature search is conducted, but no relevant studies are found, or the number of studies found is very small, or the study results are very different, then the meta-analysis may not be necessary. Without comprehensive literature collection and reliable operational methods, meta-analyses will be unreliable and cannot be correctly interpreted and utilized.

For meta-analyses, in order to ensure the scientificity, standardization and transparency of research reports, some unified report writing specifications have been formulated and widely adopted internationally, such as PRISMA statement, MOOSE statement, Cochrane Systematic Review specifications, etc. Referring to the reporting standards when writing a paper can effectively improve the practicality and quality of the scientific article. MOOSE (Meta-analyses Of Observational Studies in Epidemiology) is funded by the U.S. Centers for Disease Control and Prevention which convened experts to discuss and formulate the criteria for systematic review reporting specifications in observational studies in epidemiology (Stroup et al., 2000; Clinical Research and Medical Statistics, 2023). Here, we take Elsevier's MOOSE statement as an example to illustrate the reporting specifications of meta-analysis. Elsevier's MOOSE statement requires that the reporting content of observational studies include 7 Parts, the specific contents are as follows (Elsevier, 2023)

(1) Reporting of Background: Problem definition; Hypothesis statement; Description of Study Outcome(s); Type of exposure or intervention used; Type of study design used; Study population.

(2) Reporting of Search Strategy: Qualifications of searchers (eg, librarians and investigators); Search strategy, including time period included in the synthesis and keywords; Effort to include all available studies, including contact with authors; Databases and registries searched; Search software used, name and version, including special features used (e.g., explosion); Use of hand searching (eg, reference lists of obtained articles); List of citations located and those excluded, including justification; Method for addressing articles published in languages other than English; Method of handling abstracts and unpublished studies; Description of any contact with authors.

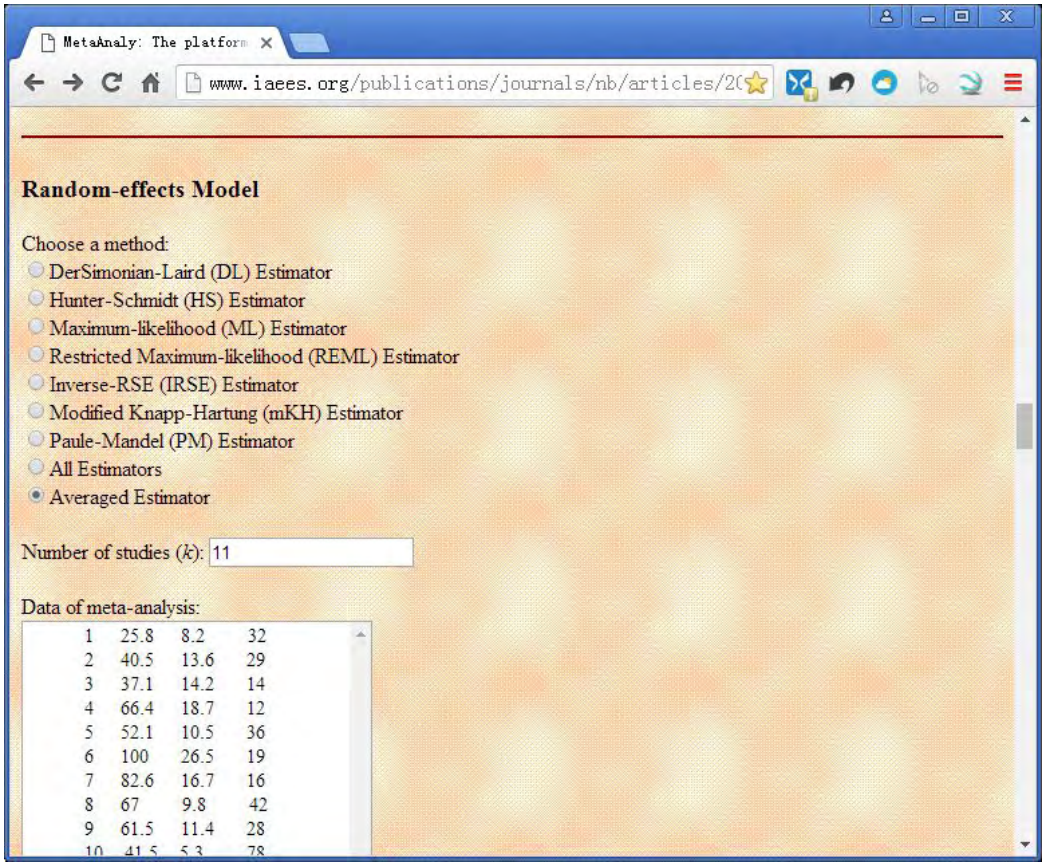
(3) Reporting of Methods: Description of relevance or appropriateness of studies assembled for assessing the hypothesis to be tested; Rationale for the selection and coding of data (e.g., sound clinical principles or convenience); Documentation of how data were classified and coded (e.g., multiple raters, blinding, and interrater reliability); Assessment of confounding (e.g., comparability of cases and controls in studies where appropriate.

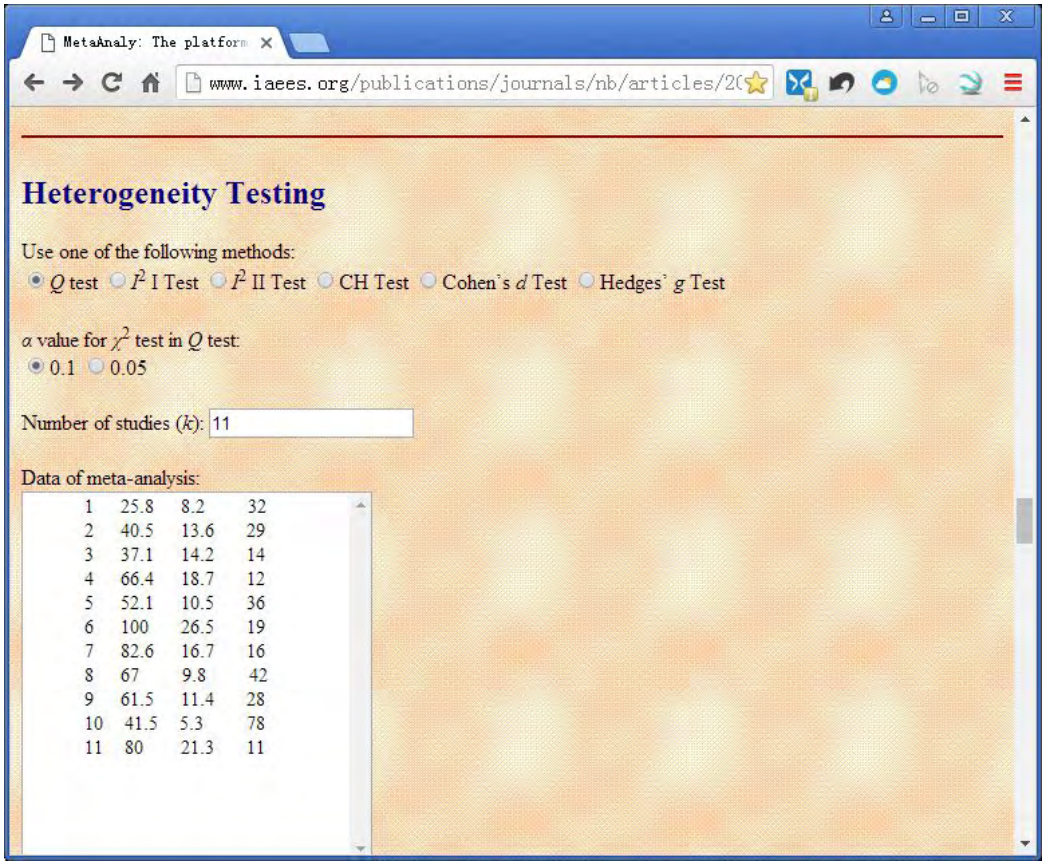
(4) Reporting Criteria: Assessment of study quality, including blinding of quality assessors, stratification or regression on possible predictors of study results; Assessment of heterogeneity; Description of statistical methods (e.g., complete description of fixed or random effects models, justification of whether the chosen models account for predictors of study results, dose-response models, or cumulative meta-analysis) in sufficient detail to be replicated; Provision of appropriate tables and graphics.

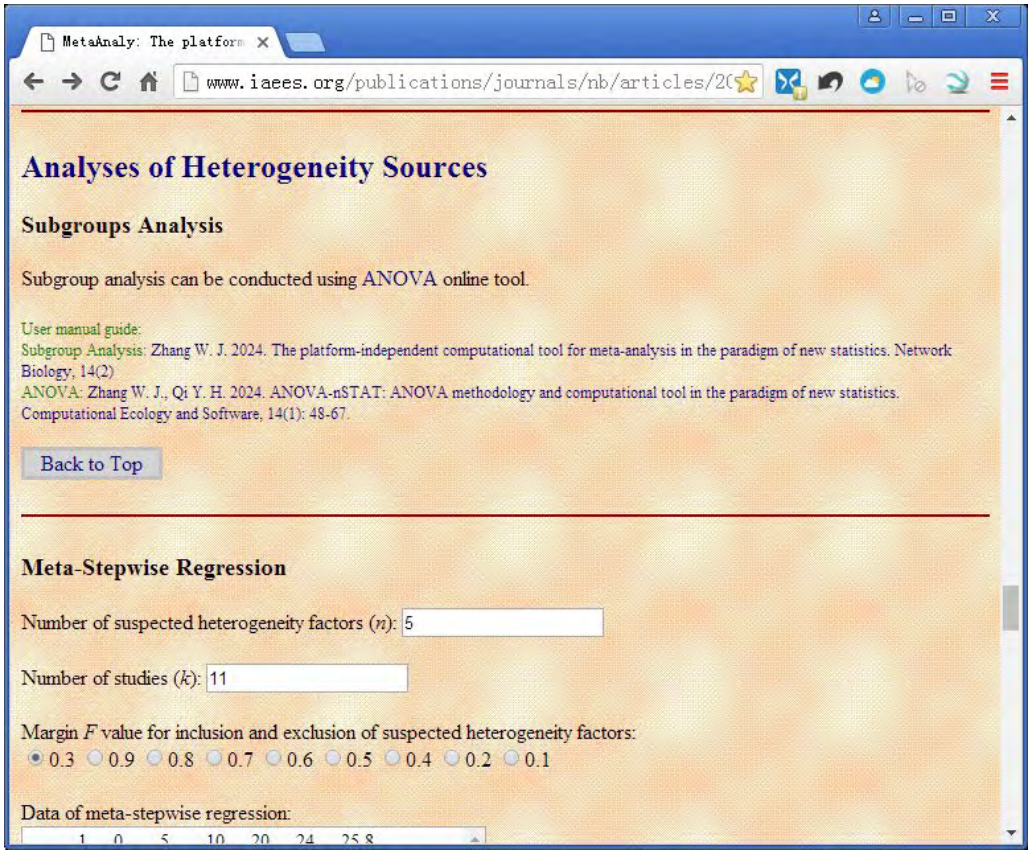
(5) Reporting of Results: Table giving descriptive information for each study included; Results of sensitivity testing (e.g., subgroup analysis); Indication of statistical uncertainty of findings.

(6) Reporting of Discussion: Quantitative assessment of bias (e.g., publication bias); Justification for exclusion (eg, exclusion of non–English-language citations); Assessment of quality of included studies.

(7) Reporting of Conclusions: Consideration of alternative explanations for observed results; Generalization of the conclusions (i.e., appropriate for the data presented and within the domain of the literature review); Guidelines for future research; Disclosure of funding source.


MetaAnaly: The platform
www.iaees.org/publications/journals/nb/articles/2(
Random-effects Model
Choose a method:
DerSimonian-Laird (DL) Estimator
Hunter-Schmidt (HS) Estimator
Maximum-likelihood (ML) Estimator
Restricted Maximum-likelihood (REML) Estimator
Inverse-RSE (IRSE) Estimator
Modified Knapp-Hartung (mKH) Estimator
Paule-Mandel (PM) Estimator
All Estimators
Averaged Estimator
Number of studies (k): 11
Data of meta-analysis:
1 25.8 8.2 32
2 40.5 13.6 29
3 37.1 14.2 14
4 66.4 18.7 12
5 52.1 10.5 36
6 100 26.5 19
7 82.6 16.7 16
8 67 9.8 42
9 61.5 11.4 28


MetaAnaly: The platform
www.iaees.org/publications/journals/nb/articles/2(
Heterogeneity Testing
Use one of the following methods:
Q test I² I Test I² II Test CH Test Cohen's d Test Hedges' g Test
α value for χ² test in Q test:
0.1 0.05
Number of studies (k): 11
Data of meta-analysis:
1 25.8 8.2 32
2 40.5 13.6 29
3 37.1 14.2 14
4 66.4 18.7 12
5 52.1 10.5 36
6 100 26.5 19
7 82.6 16.7 16
8 67 9.8 42
9 61.5 11.4 28
10 41.5 5.3 78
11 80 21.3 11



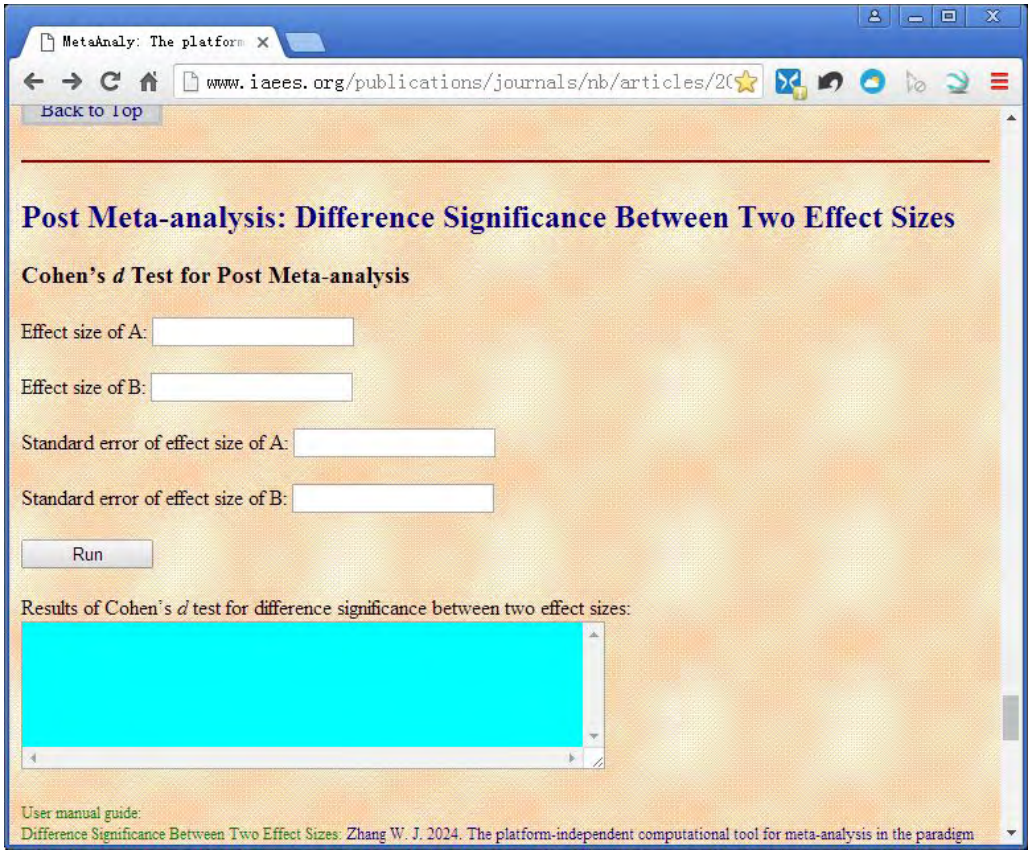


Fig. 1 Online computational tool, MetaAnaly, for meta-analysis.

Acknowledgment

I am thankful to the support of Research on New Technologies for Tannery Wastewater Treatment (2020.9-2024.9), from Zhongmeng Environmental Construction Co., Ltd., China.

References

Adair JG, Vohra N. 2003. The explosion of knowledge, references, and citations. Psychology's unique response to a crisis. American Psychologist, 58(1): 15-23. https://doi.org/10.1037/0003-066x.58.1.15

Antonelli J, Cefalu M. 2020. Averaging causal estimators in high dimensions. Journal of Causal Inference, 8(1): 92-107. https://doi.org/10.1515/jci-2019-0017

Bakbergenuly I, Hoaglin DC, Kulinskaya E. 2019. Estimation in meta-analyses of mean difference and standardized mean difference. Statistics in Medicine, 39(2): 171-191. https://doi.org/10.1002/sim.8422

Bangert-Drowns RL. 1986. Review of developments in meta-analytic method. Psychological Bulletin, 99(3): 388-399. https://doi.org/10.1037/0033-2909.99.3.388

Böhning D, Malzahn U, Dietz E, Schlattmann P, Viwatwongkasem C, Biggeri A. 2002. Some general points in estimating heterogeneity variance with the DerSimonian-Laird estimator. Biostatistics, 3(4): 445-457. https://doi.org/10.1093/biostatistics/3.4.445

Borenstein M. 2011. Pooling Effect Sizes. In: Doing Meta-Analysis with R: A Hands-On Guide (Harrer M, et al., eds). Chapman and Hall/CRC Press, Boca Raton, FL and London, USA. https://bookdown.org/MathiasHarrer/Doing_Meta_Analysis_in_R/pooling-es.html

Borenstein M, Higgins JPT, Larry V, Hedges LV, Rothstein HR. 2017. Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. Research Synthesis Methods, 8(1): 5-18. https://doi.org/10.1002/jrsm.1230

Brockhaus AC, Bender R, Skipka G. 2014. The Peto odds ratio viewed as a new effect measure. Statistics in Medicine, 33(28): 4861-4874. https://doi.org/10.1002/sim.6301

Celeng C, Leiner T, Maurovich-Horvat P, et al. 2019. Anatomical and functional computed tomography for diagnosing hemodynamically significant coronary artery disease: A meta-analysis. JACC: Cardiovascular Imaging, 12(7 Pt 2): 1316-1325. doi: 10.1016/j.j https://doi.org/2018.07.022

Cohen BH. 2008. Explaining Psychological Statistics (3rd ed). John Wiley and Sons, New York, USA. https://www.amazon.com/Explaining-Psychological-Statistics-Barry-Cohen/dp/0470007184

Cohen J. 1988. Statistical Power Analysis for the Behavioral Science. Lawrence Erlbaum Associates, Hillsdale, New Jersey, USA. https://www.amazon.com/Statistical-Power-Analysis-Behavioral-Sciences/dp/0805802835

Cuijpers P, Cristea IA, Karyotaki E, et al. 2016. How effective are cognitive behavior therapies for major depression and anxiety disorders? A meta-analytic update of the evidence. World Psychiatry, 15(3): 245-258. https://doi.org/10.1002/wps.20346

DerSimonian R, Laird N. 1986. Meta-analysis in clinical trials. Controlled Clinical Trials, 7(3): 177-188. https://doi.org/10.1016/0197-2456(86)90046-2

Elsevier. 2023. MOOSE (Meta-analyses Of Observational Studies in Epidemiology) Checklist. https://legacyfileshare.elsevier.com/promis_misc/ISSM_MOOSE_Checklist.pdf

Erez A, Bloom MC, Wells MT. 1996. Using random rather than fixed effects models in meta-analysis: Implications for situational specificity and validity generalization. Personnel Psychology, 49(2): 275-306. https://doi.org/10.1111/J.1744-6570.1996.tb01801.x

European Medicines Agency (EMEA). 2006. Guideline on clinical trials in small populations. CHMP/EWP/83561/2005.

http //www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003615.pdf

Furukawa T, McGuire H, Barbui C. 2003. Low dosage tricyclic antidepressants for depression. Cochrane Database Systematic Review, 2003(3): CD003197. https://doi.org/10.1002/14651858.CD003197

Gagne JJ, Thompson L, O'Keefe K, Kesselheim AS. 2014. Innovative research methods for studying treatments for rare diseases: Methodological review. BMJ, 349: 6802. https://doi.org/10.1136/bmj.g6802

Hardy RJ, Thompson SG. 1996. A likelihood approach to metaanalysis with random effects. Statistics in Medicine, 15(6): 619-629. https://pubmed.ncbi.nlm.nih.gov/8731004/

Harrer M, Cuijpers P, Furukawa TA, Ebert DD. 2021. Doing Meta-Analysis with R: A Hands-On Guide. Chapman & Hall/CRC Press, Boca Raton, FL and London, USA. https://www.amazon.com/Doing-Meta-Analysis-R-Hands-Guide/dp/0367610078

Hartung J, Knapp G. 2001a. On tests of the overall treatment effect in meta-analysis with normally distributed responses. Statistics in Medicine, 20(12): 1771-1782. https://doi.org/10.1002/sim.791

Hartung J, Knapp G. 2001b. A refined method for the meta-analysis of controlled clinical trials with binary outcome. Statistics in Medicine, 20(24): 3875-3889. https://doi.org/10.1002/sim.1009

Hedges LV, Olkin I. 1985. Statistical Methods for Meta-Analysis. Academic Press, New York, USA. https://idostatistics.com/hedges-olkin-1985-statistical-methods-for-meta-analysis/

Higgins JP, Thompson SG. 2002. Quantifying heterogeneity in a meta-analysis. Statistics in Medicine, 21(11): 1539-1558. https://doi.org/10.1002/sim.1186

Higgins JP, Thompson SG, Deeks JJ, Altman DG. 2003. Measuring inconsistency in meta-analyses. BMJ, 327(7414): 557-560. https://doi.org/10.1136/bmj.327.7414.557

Huang HN. 2018. Uncertainty estimation with a small number of measurements, part I: new insights on the *t*-interval method and its limitations. Measurement Science and Technology, 29: 015004. https://doi.org/10.1088/1361-6501/aa96c7

Huang HN. 2020. Signal content index (SCI): a measure of the effectiveness of measurements and an alternative to *p*-value for comparing two means. Measurement Science and Technology, 31(4): 045008. https://doi.org/10.1088/1361-6501/ab46fd

Huang HN. 2023. Combining estimators in interlaboratory studies and meta-analyses. Research Synthesis Methods, 14(3): 526-543. https://doi.org/10.1002/jrsm.1633

Hunter J, Schmidt F. 2000. Fixed effects vs. random effects meta‐analysis models: Implications for cumulative research knowledge. International Journal of Selection and Assessment, 8: 275-292. https://doi.org/10.1111/1468-2389.00156

IntHout J, Ioannidis JPA, Borm GF, Goeman JJ. 2015. Small studies are more heterogeneous than large ones a meta-meta-analysis. Journal of Clinical Epidemiology, 68(8): 860-869. https://doi.org/10.1016/j.jclinepi.2015.03.017

Ioannidis JPA. 2005. Why most published research findings are false. Plos Medicine, https://doi.org/10.1371/journal.pmed.0020124

Kafdar K. 2021. Editorial: Statistical significance, *p*-values, and replicability. The Annals of Applied Statistics. https://doi.org/10.1214/21-AOAS1500

Kim KA, Kim NJ, Choo EH. 2023. The effect of fibrates on lowering low-density lipoprotein cholesterol and cardiovascular risk reduction: A systemic review and meta-analysis. European Journal of Preventive Cardiology, zwad331. https://doi.org/10.1093/eurjpc/zwad331

Koepke A, Lafarge T, Toman B, Possolo A. 2017. NIST Consensus Builder—User's Manual. National Institute of Standards and Technology, Washington DC, USA. https://consensus.nist.gov

Kontopantelis E, Springate DA, Reeves D. 2013. A re-analysis of the Cochrane Library data: The dangers of unobserved heterogeneity in meta-analyses. PLoS ONE, 8(7): 69930. https://doi.org/10.1371/journal.pone.0069930

Korn EL, McShane LM, Freidlin B. 2013. Statistical challenges in the evaluation of treatments for small patient populations. Science Translational Medicine, 2178. https://doi.org/10.1126/scitranslmed.3004018

Langan D, Higgins JPT, Dan JacksonD, et al. 2019. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. Research Synthesis Methods, 10(1): 83-98. https://doi.org/10.1002/jrsm.1316

Lavancier F, Rochet P. 2016. A general procedure to combine estimators. Computational Statistics and Data Analysis, 94: 175-192. doi:10.1016/j.csda.2015.08.001

Li HH. 2021. p-values are too sensitive, and the effect size is long and good to save. Science Net. http //blog.sciencenet.cn/blog-2619783-1286084.html

Clinical Research and Medical Statistics. 2023. Let's talk about the statistical nature of meta-analysis technology. https://mp.weixin.qq.com/s/Q4TY_st-oSVCiNr-rZDFQw

Mantel N, Haenszel W. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22(4): 719-748. https://pubmed.ncbi.nlm.nih.gov/13655060/

Mao Z. 2019. Statistical analysis in systematic review: Purpose of analysis and principles of meta-analysis. Clinical Research and Evidence-Based Medicine.https://mp.weixin.qq.com/s/ayAOuoFqk14qsmxz2T0p-g

Mitra P, Lian H, Mitra R, Liang H, Xie M. 2019. A general framework for frequentist model averaging. Science China Mathematics, 62(2): 205-226. doi:10.1007/s11425-018-9403-x

Paule RC, Mandel J. 1982. Consensus values and weighting factors. Journal of Research of the National Bureau of Standards, 87(5): 377-385. https://doi.org/10.6028/jres.087.022

Paule RC, Mandel J. 1989. Consensus values, regressions, and weighting factors. Journal of Research of the National Institute of Standards and Technology, 94(3): 197-203. https://doi.org/10.6028/jres.094.020

Pearce M, Garcia L, Abbas A, et al. 2022. Association between physical activity and risk of depression: A systematic review and meta-analysis. JAMA psychiatry, 79(6): 550-559. https://doi.org/10.1001/jamapsychiatry.2022.0609

Petropoulou M, Mavridis DA. 2017. Comparison of 20 heterogeneity variance estimators in statistical synthesis of results from studies: a simulation study. Statistics in Medicine, 36(27): 4266-4280. doi:10.1002/sim.7431

Poole C, S Greenland S. 1999. Random-effects meta-analyses are not always conservative. American Journal of Epidemiology, 150(5): 469-475. https://doi.org/10.1093/oxfordjournals.aje.a010035

Qi YH, Liu GH, Zhang WJ. 2016. A Matlab program for stepwise regression. Network Pharmacology, 1(1): 36-40.
http://www.iaees.org/publications/journals/np/articles/2016-1(1)/Matlab-program-for-stepwise-regression.pdf

Raudenbush SW. 2009. Analyzing effect sizes: random-effects models. In: The Handbook of Research Synthesis and Meta-Analysis (Cooper H, Hedges LV, Valentine JC, eds). 295-315, Russell Sage Foundation, New York, USA.
https://www.russellsage.org/publications/handbook-research-synthesis-and-meta-analysis-second-edition

Riaz H, Khan MS, Siddiqi TJ, et al., 2018. Association Between Obesity and Cardiovascular Outcomes A Systematic Review and Meta-analysis of Mendelian Randomization Studies, JAMA Netw Open, 1(7): e183788. https://doi.org/10.1001/jamanetworkopen.2018.3788

Röver C, Knapp G, Friede T. 2015. Hartung-Knapp-Sidik-Jonkman approach and its modification for

random-effects meta-analysis with few studies. BMC Medical Research Methodology, 15: 99. https://doi.org/10.1186/s12874-015-0091-1

Rukhin AL, Biggerstaff BJ, Vangel MG. 2000. Restricted maximum-likelihood estimation of a common mean and the Mandel-Paule algorithm. Journal of Statistical Planning and Inference, 83(2): 319-330 https://doi.org/10.1016/S0378-3758(99)00098-1

Schwarzer G. 2014. Meta-analysis with R. R package version 3.7-1. http //CRAN.R-project.org/package=meta

Schwarzer G, Efthimiou O, Rücker G. 2021. Inconsistent results for Peto odds ratios in multi-arm studies, network meta-analysis and indirect comparisons. Research Synthesis Methods, 12(6): 849-854. https://doi.org/10.1002/jrsm.1503

Sellke T, Bayarri MJ, Berger JO. 2001. Calibration of *p* values for testing precise null hypotheses. The American Statistician, 55(1): 62-71. https://doi.org/10.1198/000313001300339950

Sidik K, Jonkman JN. 2002. A simple confidence interval for meta-analysis. Statistics in Medicine, 21(21): 3153-3159. https://doi.org/10.1002/sim.1262

Sidik K, Jonkman JN. 2007. A comparison of heterogeneity variance estimators in combining results of studies. Statistics in Medicine, 26(9): 1964-1981. https://onlinelibrary.wiley.com/doi/10.1002/sim.2688

Stroup DF, Berlin JA, Morton SC, Olkin I, et al. 2000. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. JAMA, 19: 283(15): 2008-2012. https://10.1001/jama.283.15.2008

Takkouche B, Cadarso-Suarez C, Spiegelman D. 1999. Evaluation of old and new tests of heterogeneity in epidemiologic meta-analysis. American Journal of Epidemiology, 150(2): 206-215. https://doi.org/10.1093/oxfordjournals.aje.a009981

Tang JL, Mao Z. 2015. Statistical analysis in systematic review (Chapter 31). In: Epidemiology Volume 1 (3rd ed) (Li LM, ed). People's Medical Publishing House, Beijing, China. https://www.taobao.com/list/item/753734635956.htm?spm=a21wu.10013406.taglist-content.39.43006f04W2LLpM

Tang JL, Yang ZY. 2015. Systematic review and meta-analysis (Chapter 14). In: Epidemiology Volume 1 (3rd ed) (Li LM, ed). People's Medical Publishing House, Beijing, China. https://www.taobao.com/list/item/753734635956.htm?spm=a21wu.10013406.taglist-content.39.43006f04W2LLpM

Tong C. 2019. Statistical Inference Enables Bad Science; Statistical Thinking Enables Good Science. The American Statistician, 73(S1): 246-261. https://doi.org/10.1080/00031305.2018.1518264

Turner RM, Davey J, Clarke MJ, Thompson SG, Higgins JPT. 2012. Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane Database of Systematic Reviews. International Journal of Epidemiology, 41(3): 818-827. https://doi.org/10.1093/ije/dys041

Veroniki AA, Jackson D, Viechtbauer W, et al. 2016. Methods to estimate the between-study variance and its uncertainty in meta-analysis. Research Synthesis Methods, 7(1): 55-79. doi:10.1002/jrsm.1164

Viechtbauer W. 2005. Bias and efficiency of meta-analytic variance estimators in the random-effects model. Journal of Educational and Behavioral Statistics, 30(3): 261-293. https://doi.org/10.3102/10769986030003261

Viechtbauer W. 2007. Confidence intervals for the amount of heterogeneity in meta-analysis. Statistics in Medicine, 26(1): 37-52. https://doi.org/10.1002/sim.2514

Viechtbauer W. 2010. Conducting meta-analyses in R with the metafor package. Journal of Statistical Software, 36(3). https://doi.org/10.18637/jss.v036.i03

Wasserstein RL, Schirm AL, Lazar NA, 2019. Editorial: Moving to a world beyond "p<0.05". The American Statistician, 79: 1-19. https://doi.org/10.1080/00031305.2019.1583913

Xu C, Furuya-Kanamori L, Lin LF, Doi SA. 2020. Peto odds ratios demonstrate no advantage over classic odds ratios in meta-analysis of binary rare outcomes. MedRxiv, https://doi.org/10.1101/2020.10.13.20212290

Zhang WJ. 2016. Screening node attributes that significantly influence node centrality in the network. Selforganizology, 3(3): 75-86.
http://www.iaees.org/publications/journals/selforganizology/articles/2016-3(3)/screening-node-attributes-that-influence-node-centrality.pdf

Zhang WJ. 2021a. A statistical simulation method for causality inference of Boolean variables. Network Biology, 11(4): 263-273.
http://www.iaees.org/publications/journals/nb/articles/2021-11(4)/a-method-for-causality-inference-of-Boolean-variables.pdf

Zhang WJ. 2021b. Causality inference of linearly correlated variables: The statistical simulation and regression method. Computational Ecology and Software, 11(4): 154-161
http://www.iaees.org/publications/journals/ces/articles/2021-11(4)/causality-inference-of-linearly-correlated-variables.pdf

Zhang WJ. 2021c. Causality inference of nominal variables: A statistical simulation method. Computational Ecology and Software, 11(4): 142-153
http://www.iaees.org/publications/journals/ces/articles/2021-11(4)/causality-inference-of-nominal-variables-with-statistical-simulation-method.pdf

Zhang WJ. 2022a. Confidence intervals Concepts, fallacies, criticisms, solutions and beyond. Network Biology, 12(3): 97-115.
http://www.iaees.org/publications/journals/nb/articles/2022-12(3)/confidence-intervals-fallacies-criticisms-solutions.pdf

Zhang WJ. 2022b. Dilemma of *t*-tests: Retaining or discarding choice and solutions. Computational Ecology and Software, 12(4): 181-194.
http://www.iaees.org/publications/journals/ces/articles/2022-12(4)/dilemma-of-t-tests.pdf

Zhang WJ. 2022c. *p*-value based statistical significance tests: Concepts, misuses, critiques, solutions and beyond. Computational Ecology and Software, 12(3): 80-122.
http://www.iaees.org/publications/journals/ces/articles/2022-12(3)/p-value-based-statistical-significance-tests.pdf

Zhang WJ. 2023. A desktop calculator for effect sizes: Towards the new statistics. Computational Ecology and Software, 13(4): 136-181.
http://www.iaees.org/publications/journals/ces/articles/2023-13(4)/4-Zhang-Abstract.asp

Zhang W.J. 2024. SampSizeCal: The platform-independent computational tool for sample sizes in the paradigm of new statistics. Network Biology, 14(2): 100-155.
http://www.iaees.org/publications/journals/nb/articles/2024-14(2)/5-Zhang-Abstract.asp

Zhang WJ, Qi YH. 2024. ANOVA-nSTAT: ANOVA methodology and computational tools in the paradigm of new statistics. Computational Ecology and Software, 14(1): 48-67.
http://www.iaees.org/publications/journals/ces/articles/2024-14(1)/4-Zhang-Abstract.asp

Zhu YW, Liu K, Zhu H, et al. 2023. Comparative efficacy and safety of novel immuno-chemotherapy for extensive-stage small-cell lung cancer: a network meta-analysis of randomized controlled trial. Therapeutic Advances in Medical Oncology, 15: 17588359231206147. https://doi.org/10.1177/17588359231206147