Appraising Between-Study Homogeneity, Small-Study Effects, Moderators, and Confounders

12

Areti Angeliki Veroniki, Tania B. Huedo-Medina, and Kostas N. Fountoulakis

Abstract

Meta-analysis is the statistical synthesis of results from two or more clinical studies that address the same issue and compare two different interventions. Although the combination of results of several studies in a meta-analysis can increase power and improve precision, caution is needed in the presence of between-study heterogeneity and selection bias. These two factors can importantly impact meta-analysis conclusions and hence influence decision-making. Several methods have been developed to appraise the between-study variation and the tendency of small studies to yield larger intervention effects compared to larger studies. This chapter presents an overall review of methods presented in the meta-analysis literature along with their properties.

12.1 Introduction

Systematic reviews and meta-analyses of well-conducted randomized controlled trials (RCTs) that address the same clinical question(s) can provide the highest level of evidence for decision-making on interventions and are vital in the practice of evidence-based medicine. Although meta-analysis constitutes a valuable tool to

A.A. Veroniki, PhD (✉)
Li Ka Shing Knowledge Institute, Faculty of Medicine, St. Michael's Hospital and University of Toronto, 30 Bond St, Toronto, ON M5B 1W8, Canada
e-mail: veronikia@smh.ca

T.B. Huedo-Medina
Department of Allied Health Sciences, University of Connecticut, Storrs, CT, US

K.N. Fountoulakis
Department of Psychiatry, Division of Neurosciences School of Medicine, Aristotle University of Thessaloniki, Pournari Pylaia, Thessaloniki, Greece

G. Biondi-Zoccai (ed.), *Umbrella Reviews: Evidence Synthesis with Overviews of Reviews and Meta-Epidemiologic Studies*, DOI 10.1007/978-3-319-25655-9_12

summarize study-specific results and may reduce both bias and uncertainty from individual studies, it widely depends on the quality, homogeneity, and freedom from bias of the available studies. The main two threats of the meta-analysis validity are:

1. The between-study variability beyond random error, termed heterogeneity
2. The phenomenon that small RCTs suggest different, often larger, intervention effects than large RCTs, termed "small-study effects" [1–3]

A certain degree of variability in study-specific intervention effects is almost always present due to chance, but additional variability might occur due to many reasons. These might include differences in the way studies are conducted and how the intervention effect estimates are measured. There are three different types of heterogeneity:

1. Clinical heterogeneity, which is referred to as the variability in the participants, interventions, and outcomes
2. Methodological heterogeneity, which reflects the variability in study design and risk of bias
3. Statistical heterogeneity, which is referred to as the variability in the intervention effects

Statistical heterogeneity is usually a consequence of clinical or methodological variability, or both, among trials, and is often called "heterogeneity" omitting the term "statistical." The estimation of heterogeneity is an additional aim in meta-analysis as it improves interpretation of results and can provide insights on the summary intervention effect predictions. One of the most widely statistical methods used in meta-analysis is the inverse-variance method; it uses the reciprocal of the within-study variances as study weights. The presence of heterogeneity affects the estimation of study weights and hence the estimated uncertainty of the summary intervention effect.

A commonly encountered association in meta-analysis is the one between the estimated study-specific intervention effects and the size of studies; it can be caused by several reasons. One possible explanation is that small studies with non-significant results are less likely to be published, because journals and authors may tend to publish and submit small studies with significant results. Other explanations may include selective outcome reporting (e.g., reporting outcomes with statistically significant results), heterogeneity between small and large studies (e.g., small studies recruit patients of high baseline risk that would largely benefit from the intervention), mathematical artifact between the two factors, or simply coincidence.

Several approaches have been proposed to estimate the between-study heterogeneity and small-study effects as a result of selection bias (including publication bias, language bias, citation bias, and reporting bias) [4–6]. This chapter includes a review of the graphical methods, statistical tests, and statistical measures used in pairwise meta-analysis to evaluate homogeneity and selection bias.

12.2 Approaches for Assessing the Between-Study Heterogeneity

A key aim in meta-analysis is to make inferences about the between-study heterogeneity as its presence can have a considerable impact on the meta-analysis conclusions. There are multiple approaches available to evaluate heterogeneity in meta-analysis, including graphical methods and statistical tests to assess its presence, statistical measures to quantify heterogeneity, and methods to estimate its magnitude. This section discusses several alternatives to appraise between-study heterogeneity in meta-analysis.

12.2.1 Graphical Representation of the Between-Study Heterogeneity

A visual inspection of graphical representations is commonly the first approach researchers select to assess the variation between study-specific effects due to heterogeneity, beyond what is expected by chance. This is an informal approach but a very useful way to indicate outlier studies, as well as those that might be responsible for the between-study heterogeneity. In the next subsections, we present the graphical displays that have most commonly been used in the meta-analysis literature [7, 8].

12.2.1.1 Forest Plot

Forest plots (Fig. 12.1) are the most popular plots in meta-analysis; they display the study-specific effect estimates along with their confidence intervals, and at the bottom of the plot, the meta-analysis result is provided [10–12]. The effect measure (e.g., odds ratio) is usually presented on the horizontal axis allowing detailed study data to be plotted alongside the results, such as the number of events and sample size for each study arm. However, some authors argue that the effect measure should be presented on the vertical axis as dependent variables are commonly plotted in statistics [13]. The size of the plotting symbol used to represent the intervention effect is usually selected to be proportional to the inverse of the variance of the

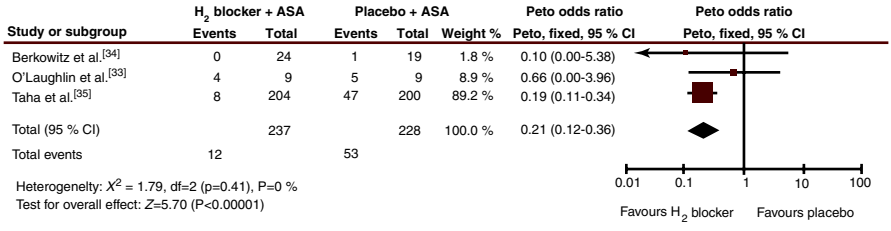

Fig. 12.1 Forest plot. Meta-analysis of three randomized controlled trials of histamine H_2 receptor antagonists (H_2 blockers) in conjunction with acetylsalicylic acid (*ASA*) therapy for outcome of peptic ulcer (Reproduced with permission [9])

study effect estimate. Therefore, more precise estimates (i.e., with smaller variance) are represented by larger plot symbols, highlighting also the amount of information that they contribute to the meta-analysis.

A greater variation in the study-specific intervention effects, more than it would be expected by chance alone, suggests there is evidence for between-study heterogeneity. In a forest plot, this is usually inspected by the poor overlap of the intervention effects' confidence intervals.

12.2.1.2 Galbraith Plots

Galbraith (or radial) plots (Fig. 12.2) are often used to present the results of studies in a meta-analysis and to informally assess between-study heterogeneity [15, 16]. The plot is a scatter plot of the standardized study-specific intervention effects, i.e., the estimated effect measures (e.g., log-odds ratio) divided by their standard errors (SE) (or equivalently the z-score) on the *y*-axis, against their inversed SEs on the *x*-axis. Each study is represented by a single point, and a regression line is drawn corresponding to the pooled fixed-effect meta-analysis estimate. Therefore, the slope of the regression is as an estimate of the intervention effect, when there are no small-study effects. In addition, the 95 % confidence region of the through-the-origin regression line is depicted by the area between the two lines drawn at a vertical distance of ± 2 above and below the regression line. Under the assumption that all studies estimate a common (fixed) intervention effect, we expect that the majority (95 %) of study points lie within this confidence region.

Using this graphical representation, studies outside this region contribute to between-study heterogeneity, and the imprecise (small 1/SE, or large SE, or small

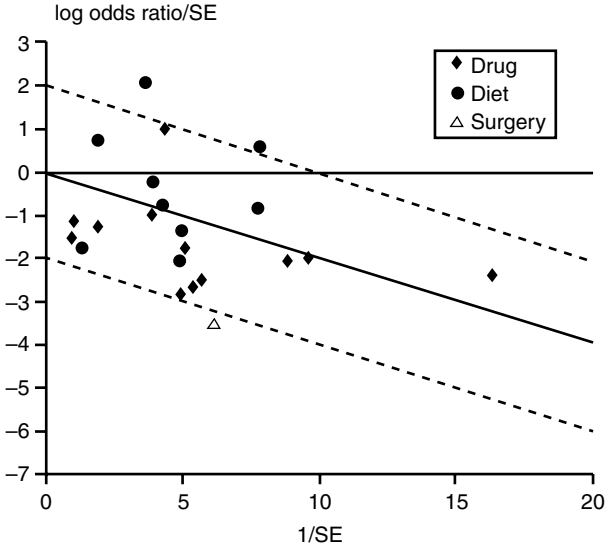

Fig. 12.2 Galbraith plot. Log-odds ratio for ischemic heart disease in trials of serum cholesterol reduction by type of intervention (Reproduced with permission [14])

studies) intervention effects lie close to the *y*-axis, whereas precise intervention effects will be situated further away.

12.2.1.3 L'Abbé Plot

L'Abbé plots (Fig. 12.3) facilitate the examination of whether the intervention effects across studies are homogeneous, but they can be used for dichotomous outcome data only [18]. This type of plot presents the risks (or odds) in the intervention group on the *y*-axis against those of the control group on the *x*-axis and often includes the diagonal line of equality and a regression line. The diagonal line of equality indicates that the risks in the control and intervention groups are equal within trials, and the regression line represents the risk ratio (or odds ratio), which is estimated by pooling the results in the meta-analysis. It is advisable that the study points are presented according to the precision of the intervention effect estimates (or study size) to make the plot more informative [7].

The plot can be used to infer the presence of heterogeneity, specifically where trials are widely spread around the regression line. In the absence of heterogeneity, study points should lie closely around the regression line.

12.2.1.4 Baujat Plot

Baujat plots (Fig. 12.4) are used to identify studies that influence the overall intervention effect and impact on the magnitude of the heterogeneity [19]. The rationale is that excluding an influential study will affect the meta-analytic estimate, and hence this plot assesses which studies cause the between-study heterogeneity and the greatest shifts in the overall intervention effect. The plot presents the contribution of each study to the Cochran *Q*-statistic (see Sect. 12.2.2.1) on the *x*-axis against

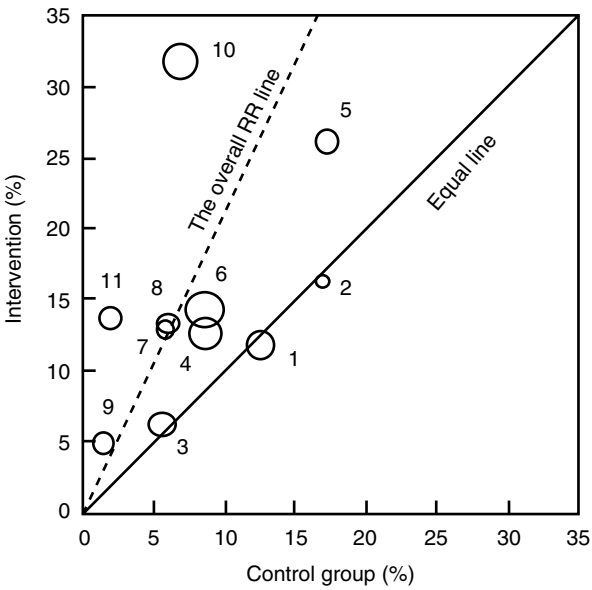

Fig. 12.3 L'Abbé plot. Rates of smoking cessation in the intervention and control group (Reproduced with permission [17])

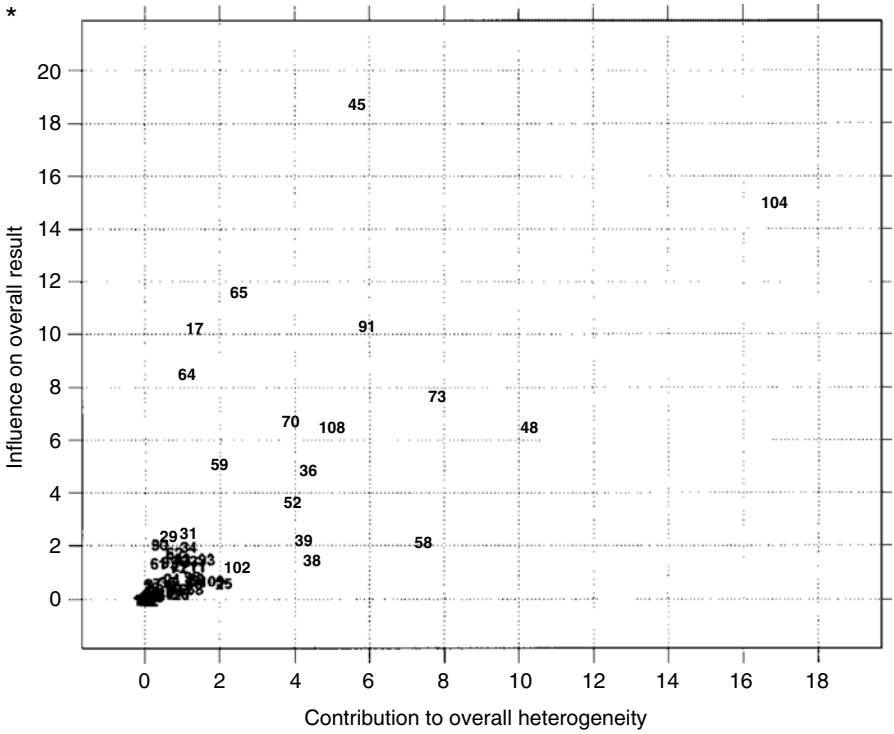

Fig. 12.4 Baujat plot for a meta-analysis of chemotherapy in head and neck cancer (Reproduced with permission [19])

the influence of each study. The influence of each study is defined as the standardized squared difference between the overall intervention effects with and without the *i*th study under the fixed-effect model, on the *y*-axis. Studies lying on the upper right corner of the plot are the most influential with the highest contribution to the total heterogeneity.

12.2.2 Statistical Tests for the Evaluation of the Between-Study Variance

The most commonly used method to assess the homogeneity assumption in meta-analysis is to carry out a statistical test. Several tests for this evaluation have been suggested in the literature, including the "generalized Cochran *Q*," Wald, likelihood ratio, and score tests [20, 21]. A popular choice for the between-study homogeneity assessment in meta-analyses is the Cochran *Q*-statistic (see Sect. 12.2.2.1) [22]. It has been suggested that among the aforementioned homogeneity tests, the Cochran *Q*-statistic performs best in terms of type I error for meta-analyses with large studies

(e.g., with arm size greater than 640) [21]. The Cochran Q-statistic belongs to the "generalized Cochran between-study variance statistics" family [23], with

$$Q_a = \sum a_i \left(y_i - \mu_a \right)^2,$$

where y_i is the observed intervention effect (e.g., log-odds ratio), index i refers to the ith study with $i = 1,\ldots k$, a_i the weight assigned to each study, and $\mu_a = \left(\sum a_i y_i\right) / \sum a_i$ the overall intervention effect. Jackson showed that Q_a has a χ^2_{k-1} distribution as a linear combination of independent central χ_1^2 random variables [24].

12.2.2.1 Cochran *Q*-Statistic

The standard test widely used in meta-analysis, is the Cochran Q-statistic testing the hypothesis that all studies share a common true effect (μ) or equivalently that the between-study variance (τ^2) is zero [22]. The Cochran Q-statistic is a special form of the "generalized Cochran between-study variance statistic" for $a_i = 1/v_i$, with vi the within-study variance in study $i = 1,\ldots k$. Hence, the Q-statistic is the weighted sum of squared differences between the observed study-specific effects and the overall effect across studies derived under the fixed-effect model. Under the null hypothesis, $H_0: \tau^2 = 0$, the Q-statistic follows approximately a χ^2-distribution with $k-1$ degrees of freedom and a critical region $Q > \chi^2_{k-1,\,1-(a/2)}$, where a is the confidence level. Several efforts have been done to define the distribution of the Q-statistic, including Biggerstaff and Tweedie approximating Q with a gamma distribution, and Biggerstaff and Jackson deriving the exact distribution, when $\tau^2 \neq 0$ [25, 26].

It has been shown that the power of the test to detect heterogeneity depends on the number and size of studies, as well as the magnitude of the true between-study variance [21]. Simulation studies suggest that the test has low power when the total information available in the meta-analysis is low (e.g., sparse data, small size and number of studies), and hence a nonsignificant result might erroneously be interpreted as absence of between-study heterogeneity [21, 27]. It is therefore recommended that reviewers use 0.10 as a cutoff level of significance instead of the usual 0.05 [28, 29]. However, a higher cutoff value increases type I error and the risk of drawing false-positive results. The Q-statistic may suggest significant heterogeneity when many studies are included in the meta-analysis and particularly when their sample sizes are very large (see, e.g., Barbui et al. that included over 15,000 participants from 135 studies) [30]. The power of the test may also be limited when the study sizes differ substantially or a single study is a lot larger when compared with the others in the analysis [27].

12.2.2.2 Generalized *Q*-Statistic

Similarly to Cochran Q, the generalized Q-statistic (Q_{gen}) is a special form of the "generalized Cochran" between-study variance statistic for $a_i = 1/\left(v_i + \tau^2\right)$. The Q_{gen}-statistic is the weighted sum of squared differences between the observed study-specific effects and the overall effect derived under the random-effects model. Under the null hypothesis that the true between-study variance is equal to a certain

amount ($\tau_0^2 \geq 0$), Q_{gen} follows a χ^2-distribution with $k-1$ degrees of freedom and a critical region: $Q_{\text{gen}} > \chi^2_{k-1,\,1-(a/2)}$.

To the best of our knowledge, the properties of the test have not been examined, providing an avenue for further work.

12.2.2.3 Cochran *Q*-Statistic Adjusted for Small-Study Effects

Rücker et al. extended Cochran Q-statistic by adjusting for small-study effects [31]. We call "small-study effects" the tendency of small studies to show larger intervention effects compared to the larger studies (see also Sect. 12.4). This can be derived by

$$Q_a^{\text{Adj}} = \sum a_i \left(y_i - \mu_a^{\text{Adj}} - \frac{\hat{s}_a}{\sqrt{a_i}} \right)^2,$$

where μ_a^{Adj} is the summary intervention effect adjusted for small-study effects with $a_i = 1/v_i$ and $\hat{s}_a$ represents a potential small-study effect. The $Q_a{}^{\text{Adj}}$ measures the residual variation with respect to a fixed-effect model allowing for small-study effects, and compared to the Cochran's Q-statistic, it holds that $Q_a^{\text{Adj}} \leq Q$. Under the null hypothesis of no between-study heterogeneity, $Q_a{}^{\text{Adj}}$ follows a χ^2-distribution with $k-2$ degrees of freedom and a critical region: $Q_a^{\text{Adj}} > \chi^2_{k-2,\,1-(a/2)}$.

In the presence of small-study effects, it is suggested to use $Q_a{}^{\text{Adj}}$ to assess the remaining between-study heterogeneity [31]. The main limitation of the Cochran's Q-statistic adjusted for small-study effects is that it depends on the choice of the estimation method for τ^2 (see Sect. 12.2.4).

12.2.3 Statistical Measures to Quantify Between-Study Variance

The statistical tests discussed in Sect. 12.2.2 are only useful for testing the existence of heterogeneity, but do not quantify the extent of heterogeneity. To date, several statistical measures have been suggested for the quantification of the degree of variability in a meta-analysis that is explained by between-study differences rather than by random error [32–34]. As for every point estimate, apart from quantifying between-study heterogeneity using a statistical measure, it is important to quantify its corresponding uncertainty too. Confidence intervals provide information on the precision and the range of values that reflect the statistical measure for heterogeneity. Methods for constructing the confidence intervals include the variance estimates recovery method [35, 36], the method using the distribution of Q_a-statistic [24–26, 32], the method based on the statistical significance of Q [32], the method based on the between-study variance estimator (see Sect. 12.2.4) [5, 32, 37], and the method using a nonparametric bootstrap approach [32].

12.2.3.1 H^2 Index

H^2 index (also known as Birge ratio) [38] has been presented by Higgins and Thompson [32] and shows the excess of the observed Q over its expected value,

$E(Q) = k - 1$. The measure reflects the relationship of between- and within-study variance and can be obtained by

$$H^2 = \frac{Q}{k-1} = \frac{\hat{\tau}^2_{\mathrm{DL}} + \sigma^2}{\sigma^2}$$

where $\hat{\tau}^2_{\mathrm{DL}}$ is the estimated between-study variance using the DerSimonian and Laird [39] estimator and σ^2 is the "typical" within-study variance:

$$\sigma^2 = \frac{\sum \frac{1}{v_i}(k-1)}{\left(\sum \frac{1}{v_i}\right)^2 - \sum\left(\frac{1}{v_i}\right)^2}.$$

The statistic takes values within the range (1, ∞), and in the absence of between-study heterogeneity, it equals 1. Higgins and Thompson [32] suggest that there is no universal rule to define thresholds for 'low,' 'moderate,' and 'high' heterogeneity for H^2. However, they suggest that values greater than 1.5 may show considerable heterogeneity, and values lower than 1.2 may show moderate to low heterogeneity.

12.2.3.2 I^2 Index

The I^2 index reflects the percentage of the total variability in a set of effect measures that is due to between-study variability beyond what is expected by within-study error. The "generalized I^2 statistics" family [37] can be expressed as

$$I^2 = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \sigma^2}$$

where $\hat{\tau}^2$ is the estimated between-study variance using one of the methods suggested in the literature (see Sect. 12.2.4) [5]. The I^2 index can be expressed as a percentage ranging from 0 to 100 %, where a value of 0 % indicates no observed heterogeneity. The Cochrane Handbook advises avoiding the use of specific thresholds for the interpretation of the I^2 statistic as they may be misleading. A general guideline to its interpretation is the following [3]:

- From 0 to 40 %, may not be important.
- From 30 to 60 %, may represent moderate heterogeneity.
- From 50 to 90 % may represent substantial heterogeneity.
- From 75 to 100 %, may represent considerable heterogeneity.
- Note that should these guidelines be used with caution, and always interpret the I^2 index along with its confidence interval.

I^2 Index Based on Cochran *Q*-Statistic

The I^2 based on Cochran Q-statistic is the most popular statistic and is usually the default method to quantify heterogeneity in most meta-analysis software.

The method is a special form of the "generalized I^2 statistics" using the DerSimonian and Laird approach [39] (see Sect. 12.2.4.1):

$$I^2_{\mathrm{DL}} = \frac{\hat{\tau}^2_{\mathrm{DL}}}{\hat{\tau}^2_{\mathrm{DL}} + \sigma^2}.$$

Alternatively, the method can be presented as

$$I^2_{\mathrm{DL}} = \frac{H^2 - 1}{H^2} = \frac{Q - (k-1)}{Q}$$

in terms of either H^2 or Cochran's Q-statistic and its degrees of freedom $(k-1)$. The I^2 statistic should be interpreted with caution when the number and size of studies in the meta-analysis are small (e.g., for fewer than ten studies in the meta-analysis and studies with fewer than 100 participants) [34, 40, 41]. Simulation studies have shown that I^2_{DL} increases with increasing study size [40, 41] and that it is associated with low power when a small number of studies are included in the meta-analysis [34]. Empirical evidence suggests care is also needed with the interpretation of I^2_{DL} when a meta-analysis includes roughly fewer than 500 events and that 95 % confidence intervals for I^2_{DL} have on average a good coverage [42].

I^2 Index Based on Generalized Q-Statistic

The I^2 based on generalized Q-statistic is a special form of the "generalized I^2 statistics" expressed as [37]

$$I^2_{\mathrm{PM}} = \frac{\hat{\tau}^2_{\mathrm{PM}}}{\hat{\tau}^2_{\mathrm{PM}} + \sigma^2}$$

where $\hat{\tau}^2_{\mathrm{PM}}$ is the estimated between-study variance using the Paule and Mandel estimator (see Sect. 12.2.4.1) [5, 43]. A simulation study suggested that the confidence interval for I^2_{PM} is wider compared to those of I^2_{DL} and that I^2_{PM} maintains coverage close to the nominal level in contrast to I^2_{DL} method [37].

12.2.3.3 R^2 Index

An alternative to H^2 and I^2 measures is the R^2 statistic that describes the quadratic inflation in the confidence interval for the summary intervention effect under the random-effects model compared to that from the fixed-effect model

$$R^2 = \frac{\mathrm{Var}(\mu_{\mathrm{RE}})}{\mathrm{Var}(\mu_{\mathrm{RE}})}$$

where μ_{RE} is the overall intervention effect under the random-effects model with weights $a_i = 1/(v_i + \hat{\tau}^2)$ and μ_{FE} the overall intervention effect under the fixed-effect model with weights $a_i = 1/v_i$. The statistic takes values within the range $(1, \infty)$, and 1 suggests identical inferences under the two meta-analysis models and homogeneity across the study-specific effects. It should be noted that R^2 and H^2 are equal when all study-specific estimates have equal precision. Since R^2 is a function

of $\hat{\tau}^2$ alone (the weights are assumed to be known), one approach to estimate the confidence interval for R^2 is via the calculation of the confidence interval for τ^2. However, note that approaches based on the Cochran's Q-statistic may not be applicable for constructing confidence intervals for R^2.

12.2.3.4 D^2 Index

Wetterslev et al. proposed the D^2 statistic to quantify the relative variance when we change from the random-effects model to the fixed-effect model [33]. The statistic is interpreted as the proportion of the between-study heterogeneity in meta-analysis relative to the total model variance of the included studies and is given by

$$D^2 = \frac{\mathrm{Var}\left(\mu_{\mathrm{RE}}\right) - \mathrm{Var}\left(\mu_{\mathrm{FE}}\right)}{\mathrm{Var}\left(\mu_{\mathrm{RE}}\right)} = 1 - \frac{1}{R^2}$$

or equivalently

$$D^2 = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \sigma_D^2},$$

where

$$\sigma_D^2 = \frac{\hat{\tau}^2 \left(\mathrm{Var}\left(\mu_{\mathrm{FE}}\right)\right)}{\mathrm{Var}\left(\mu_{\mathrm{RE}}\right) - \mathrm{Var}\left(\mu_{\mathrm{FE}}\right)}$$

is the sampling error. Although D^2, similar to I^2, is interpreted as a percentage (taking values between 0 and 1), a simulation study suggested that D^2 is equal to or greater than I^2, irrespective of the chosen effect measure and number of studies in the meta-analysis [33].

12.2.3.5 G^2 Index

Rücker et al. proposed an alternative statistic, called G^2, to measure between-study heterogeneity while adjusting for small-study effects (see also Sect. 12.4) [31]. The statistic can be obtained by

$$G^2 = 1 - \frac{\left[\sum a_i y_i^{\mathrm{Adj}} - \frac{1}{k}\left(\sum \sqrt{a_i}\right)\left(\sum \sqrt{a_i}\, y_i^{\mathrm{Adj}}\right)\right]^2}{\left[\sum a_i - \frac{1}{k}\left(\sum \sqrt{a_i}\right)^2\right]\left[\sum a_i \left(y_i^{\mathrm{Adj}}\right)^2 - \frac{1}{k}\left(\sum \sqrt{a_i}\, y_i^{\mathrm{Adj}}\right)^2\right]},$$

where y_i^{Adj} are the study-specific intervention effect estimates adjusted for small-study effects, $y_i^{\mathrm{Adj}} = \mu_{\mathrm{RE}}^{\mathrm{Adj}} + \sqrt{\hat{\tau}^2 / \left(v_i + \hat{\tau}^2\right)}\left(y_i - \mu_{\mathrm{RE}}^{\mathrm{Adj}}\right)$, with $\mu_{\mathrm{RE}}^{\mathrm{Adj}}$ the summary intervention effect under the random-effects model and adjusted for small-study effects, and $a_i = 1/v_i$.

The G^2 statistic is closely related to the Q-statistic adjusted for small-study effects (see Sect. 12.2.2.3), and it is suggested to quantify heterogeneity in the presence of small-study effects [31]. Similarly to I^2 and D^2, G^2 is interpreted as a percentage (taking values between 0 and 1) and reflects the proportion of the variability in the intervention effect that is not explained under the fixed-effect model that allows for the presence of small-study effects.

12.2.4 Estimating the Between-Study Variance

An important aspect in meta-analysis is to quantify the extent of between-study heterogeneity. The DerSimonian and Laird (DL) between-study variance estimator is the most commonly implemented approach and is the default approach in many statistical software (e.g., RevMan) [39, 44]. However, its use has often been criticized because the method may underestimate the true between-study variance, thereby producing narrow confidence intervals (CIs) for the overall intervention effect, especially for a small number of studies (e.g., $k < 10$) [45]. Hence, several alternative methods have been proposed that vary in popularity and complexity. The estimators for τ^2 are categorized as closed form and iterative methods, and their families presented in the literature to date are:

1. The method of moments estimators (e.g., DL and Paule and Mantel (PM)) [39, 43]
2. The maximum likelihood estimators (e.g., maximum likelihood (ML) [20, 46] and restricted maximum likelihood (REML) [46])
3. The model error variance estimators (e.g., Sidik and Jonkman method) [47]
4. The Bayes estimators (e.g., Rukhin Bayes, full Bayes) [48, 49]
5. The bootstrap estimators [50]

It has been shown that estimating the between-study variance in meta-analyses including only a few studies is particularly inaccurate [50–52]. Therefore, it is recommended to quantify the uncertainty around the point estimates to avoid misleading results. Again, several options exist to quantify the uncertainty in the estimated amount of the between-study variance [20, 24, 53].

In this chapter, we briefly describe the most popular estimators for the between-study variance, as well those recommended for the most frequently encountered meta-analysis. For a comprehensive overview of methods used for estimating the between-study variance and its uncertainty, see Veroniki et al. [5].

12.2.4.1 Approaches for the Between-Study Variance Point Estimate

Method of Moments Estimators

The generalized method of moments (GMM) estimator [23] can be derived by equating Q_a (see Sect. 12.2.2) and its expected value:

$$E(Q_a) = \left(\sum a_i v_i - \frac{\sum a_i^2 v_i}{\sum a_i} \right) + \tau^2 \left(\sum a_i - \frac{\sum a_i^2}{\sum a_i} \right)$$

Then, solving for τ^2, we obtain

$$\hat{\tau}^2_{\text{GMM}} = \max \left\{ 0, \frac{Q_a - \left(\sum a_i v_i - \frac{\sum a_i^2 v_i}{\sum a_i} \right)}{\sum a_i - \frac{\sum a_i^2}{\sum a_i}} \right\}$$

The method of moments estimators presented in the following subsections is a special case of the GMM estimator with varying weights a_i.

DerSimonian and Laird (DL)

This method is the most frequently used approach for the estimation of the between-study variance, and many software programs have DL as the default method. The DL estimator is a non-iterative method and is a special case of the GMM estimators with study weights $a_i = 1/v_i$.

Simulation studies have suggested that the DL method performs well when the true between-study variance is small or close to zero and the number of studies in the meta-analysis is large, whereas when τ^2 is large, DL produces estimates with significant negative bias [37, 47, 52, 54–56]. The negative bias that has been reported with respect to the DL estimator seems to be something related to using effect size measures based on 2×2 table data (e.g., odds ratios, risk ratios), where problems arise when using very large τ^2 values in simulation studies. In particular, very large τ^2 can lead to extreme values of the effect size measure, at which point many tables will include zero cells and the accuracy and applicability of the inverse-variance method becomes questionable. Jackson et al. evaluated the efficiency of the DL estimator asymptotically and showed that DL is inefficient when the studies included in the meta-analysis are of different sizes and particularly when τ^2 is large [57]. However, they suggested that the DL estimator performs well and can be efficient for inference on the summary effect when the number of studies included in the meta-analysis is large. The confidence interval for the between-study variance when using the DL method can be ideally estimated using the Jackson's method [24], as they are based on the same statistical principle and are naturally paired.

Paule and Mandel (PM)

Paule and Mandel [43] proposed to profile the generalized Q-statistic (see Sect. 12.2.2.2) until Q_{gen} equals its expected value (i.e., $E(Q_{\text{gen}}) = k - 1$). The PM estimator is an iterative method and a special case of the GMM estimator with $a_i = 1/(\tau^2 + v_i)$.

Rukhin et al. showed that when assumptions underlying the method do not hold, the method is more robust than the DL estimator, which depends on large sample

sizes [58]. It has been shown that the PM method has upward bias for a small number of studies and heterogeneity and downward bias for large number of studies and heterogeneity [52], but generally the method is less biased than its alternatives. One simulation study suggested that PM outperforms the DL and REML (see below) estimators in terms of bias [59]. Panityakul et al. [59] showed that the PM estimator is approximately unbiased for large sample sizes, and Bowden et al. [37] in their empirical study showed that as heterogeneity increases, $\hat{\tau}^2_{\text{PM}}$ becomes greater than $\hat{\tau}^2_{\text{DL}}$. The uncertainty around the between-study variance using the PM method can be ideally estimated using the *Q*-profile method [53], as they are based on the same statistical principle and are naturally paired.

Maximum Likelihood Estimators

The maximum likelihood estimators are iterative methods and are derived after maximizing the (restricted) log-likelihood function [20, 60]. A limitation of the methods is that their success to converge to a solution depends on the selection of the maximization technique (e.g., Newton-Raphson, expectation-maximization algorithm).

Maximum Likelihood (ML)

The method is asymptotically efficient and can be obtained by iterating

$$\hat{\tau}^2_{\text{ML}} = \max\left\{0, \frac{\sum w^2_{i,\text{RE}}\left(\left(y_i - \mu_{\text{RE}}\left(\hat{\tau}^2_{\text{ML}}\right)\right)^2 - v_i\right)}{\sum w^2_{i,\text{RE}}}\right\}$$

and

$$\mu_{\text{RE}}\left(\hat{\tau}^2_{\text{ML}}\right) = \frac{\sum w_{i,\text{RE}} y_i}{\sum w_{i,\text{RE}}}$$

until they converge and do not change from one iteration to the next. The study weights are derived under the random-effects model, $w_{i,\text{RE}} = 1/\left(v_i + \hat{\tau}^2_{\text{ML}}\right)$. An initial estimate of $\hat{\tau}^2_{\text{ML}}$ can be decided a priori as a plausible value of the heterogeneity variance, or it can be estimated with any other non-iterative estimation method. Each iteration step requires nonnegativity.

Simulation studies have suggested that although the ML estimator is efficient, it exhibits large negative bias for large τ^2 when the number and size of studies are small (e.g., for fewer than 10 studies and fewer than 80 participants in each study) [50–52, 56, 59]. It has been shown that the ML method is more efficient than PM, and REML methods, but exhibits the largest amount of bias [51, 52, 60, 61]. However, because of the large amount of bias, it is recommended avoiding the ML estimator [56, 59]. The confidence interval for the between-study variance when using the ML method can be ideally computed using the profile likelihood method [1], as they are based on the same statistical principle and are naturally paired.

Restricted Maximum Likelihood (REML)

The REML method is often used to correct for the negative bias produced by the ML method and can be obtained by

$$\hat{\tau}^2_{\text{REML}} = \max\left\{0, \frac{\sum w^2_{i,\text{RE}}\left(\left(y_i - \mu_{\text{RE}}(\hat{\tau}^2_{\text{REML}})\right)^2 - v_i\right)}{\sum w^2_{i,\text{RE}}} + \frac{1}{\sum w_{i,\text{RE}}}\right\},$$

with study weights derived under the random-effects model, $w_{i,\text{RE}} = 1/(v_i + \hat{\tau}^2_{\text{REML}})$ [39, 52]. The estimator is calculated by an iterative process with a nonnegative initial estimate. Again, each iteration step requires nonnegativity.

Simulation studies suggested that the REML method underestimates the true between-study variance, especially when the data are sparse [47, 52, 54, 56, 62]. For dichotomous outcome data, it was shown that the REML estimator is less biased, but less efficient than the DL estimator [51, 52]. For continuous data, it has been suggested that the REML estimator is less efficient than the ML estimator and comparable to DL estimator [56]. An empirical study [63] with dichotomous outcome data showed that the REML estimator can be smaller or larger in magnitude than the DL method. REML is recommended when large studies are included in the meta-analysis [56]. The uncertainty around the between-study variance when using the REML estimator can be ideally estimated using the profile likelihood method [20].

Bayes Estimators

Full Bayes (FB)

The FB approach takes into account the uncertainty of all parameters (including τ^2) in the results. Several investigators claim that in practice the differences between frequentist and Bayesian approaches appear to be small [60, 64]. The FB method uses non-informative priors to approximate a likelihood-based analysis. When the number of studies is large, the choice of the prior does not have a major influence on the results since they are data driven. The choice of prior is particularly important though when the number of studies is small, as it may impact on the estimated between-study variance and hence on the overall intervention effect [65, 66].

A simulation study compared 13 different prior distributions for the heterogeneity variance and suggested that the results might vary substantially when the number of studies is small [65]. The study showed that, in terms of bias, none of the distributions considered performed best for all meta-analysis scenarios. More specifically, inverse-gamma, uniform, and Wishart distributions for the between-study variance all perform poorly when the number of studies is small (<10) and produce estimates with substantial bias. An inverse-gamma prior with small hyper-parameters is often considered to be an approximately non-informative prior, but it was shown that inferences can be sensitive to the choice of hyper-parameters [67, 68]. Informative priors were recently proposed for the between-study variance using the log-odds ratio and standardized mean difference effect measures, and these might

considerably improve estimation when few studies are included in the meta-analysis [69–71]. The uncertainty around the between-study variance when using the FB estimator can be ideally estimated using Bayesian credible intervals.

12.3 Possible Causes and Approaches to Deal with Heterogeneity

Despite the best efforts of investigators to construct a dataset of carefully selected studies where the homogeneity assumption would hold, an imbalance in the distribution of effect modifiers might arise resulting in between-study heterogeneity. The identification of the causes of heterogeneity may help to account for such variation in the results thereby aiding in the interpretation of existing data, as well the planning of future studies. Between-study heterogeneity may be due to clinical and/or methodological heterogeneity, biases, and chance [3, 72]. Clinical heterogeneity suggests that a possible variability in intervention or patient-level characteristics, or in outcomes studied, can influence the intervention effect. Methodological heterogeneity refers to the variability across studies due to study design or quality (e.g., inadequate randomization or allocation concealment, high dropout rates, intention-to-treat versus per-protocol analyses). In addition to biases captured by methodological heterogeneity, there are other biases that might cause between-study heterogeneity, including selection or funding biases. It is also possible that outlier studies show extreme results due to chance (e.g., studies with small sizes and/or event rates).

Quantifying the amount of between-study heterogeneity and exploring its sources are among the most important aspects of meta-analysis. When heterogeneity is identified, the first step researchers should follow is to check the data included in the meta-analysis for potential data abstraction errors. If no errors are found and between-study variability beyond chance is still evident, a different choice in effect measure may improve homogeneity. Empirical studies have shown that relative measures (e.g., odds ratio, risk ratio) are associated with less heterogeneity than absolute measures (e.g., risk difference) [73–75]. Heterogeneity might also be due to intervention effect modifiers. This exploration might include applying subgroup or meta-regression analyses adjusting the estimated intervention effects accordingly. It should be noted that the use of individual patient data in meta-analysis allows for a thorough investigation of potential sources of heterogeneity and a better evaluation of both within- and between-study heterogeneity, avoiding the assumption that a relationship between groups holds between individuals as well [76, 77]. For small to moderate amount of heterogeneity (for a general guideline, see Sect. 12.2.3.2), one can apply the random-effects model assuming that the true study-specific effects are not identical but come from the same distribution. Under the random-effects model, the between-study variation is taken into account in the meta-analysis results, but this is not a remedy for heterogeneity as it still exists.

To facilitate the interpretation of the meta-analysis' result capturing both between-study variance and variance of summary intervention effect, a prediction interval of the possible intervention effect in an individual setting can be calculated [78–80].

A prediction interval indicates the range of values for the true intervention effect when a future study is conducted and can be obtained by

$$\mu_{\mathrm{RE}} \pm t_{1-\frac{a}{2},k-2}\sqrt{\hat{\tau}^2 + \mathrm{var}\left(\mu_{\mathrm{RE}}\right)}$$

where $t_{1-a/2,\,k-2}$ is the $100(1-a/2)\,\%$ quantile of the t_{k-2} distribution. A prediction interval can be calculated when at least three studies are included in the meta-analysis.

12.4 Methods to Appraise Small-Study Effects

The association between size and effect of the studies included in a meta-analysis should be explored, as the presence of selection bias and small-study effects may lead to meaningless conclusions. Funnel plots and statistical tests based on funnel plot asymmetry are popular in meta-analysis for assessing small-study effects. Several methods have been suggested to adjust for small-study effects, including the trim-and-fill method, the Copas selection model, and various regression-based approaches (for a review, see Mavridis and Salanti) [6].

12.4.1 Graphical Representation of Small-Study Effects

Funnel plots facilitate the visual examination for detecting bias or heterogeneity, and often it is not possible to distinguish between the two. A funnel plot (see Fig. 12.5) is a scatter plot of the study-specific intervention effect estimates against a measure of precision or study size. In agreement with forest plots (see Sect. 12.2.1.1) and in contrast to conventional scatter plots, the intervention effect estimates are usually plotted on the *x*-axis, whereas the study size or precision is plotted on the *y*-axis [82–84]. It is recommended to plot the SE (or 1/SE) of the intervention effect

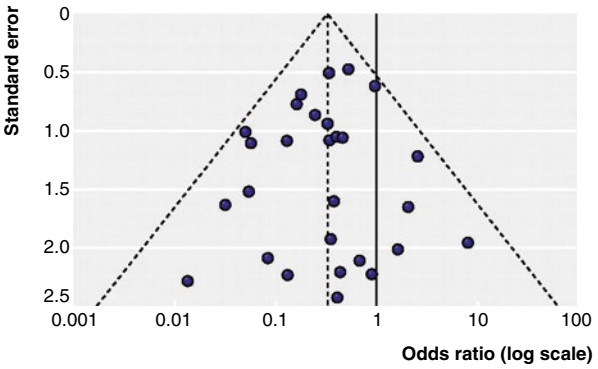

Fig. 12.5 Funnel plot. Example of symmetrical funnel plot (Reproduced with permission [81])

on the vertical axis, rather than study size, as study power is based on several other factors apart from sample size alone (e.g., number of events, standard deviation) [84], and these are summarized by SE. The plot usually includes a triangular 95 % confidence region and a vertical line corresponding to summary intervention effect under the fixed-effect model. In the absence of bias and heterogeneity, 95 % of the studies are expected to lie within the triangular region and be scattered symmetrically around the summary intervention effect. In such a case, the plot resembles a symmetrical and inverted funnel. Small studies are expected to lie at the bottom of the graph and widely spread around the summary intervention effect compared to larger studies. It is advisable to draw funnel plots when ten or more studies are available in the meta-analysis [7].

An asymmetric funnel plot suggests there is a relationship between the study-specific effect measure and precision, which might be due to selection bias (including publication bias, language bias, citation bias, and reporting bias), small-study effects, heterogeneity, sampling variation, or chance [10]. An inappropriate choice of effect measure might also result in an asymmetrical funnel plot. It should be noted that some effect measures (e.g. log-odds ratios and standardized mean differences) are correlated with their SEs, and this may produce artificial funnel plot asymmetry. In the presence of small-study effects, the funnel plot will be asymmetrical with small studies missing at the bottom right corner (for an efficacy outcome, and at the left corner for a safety outcome) suggesting an unfavorable effect. Some argue that the visual interpretation of a funnel plot is a subjective issue, and sometimes it is difficult to distinguish between symmetry and asymmetry [85, 86].

Peters et al. proposed a modified version of the conventional funnel plot, in which extra contours representing the statistical significance of each study are added (see Fig. 12.6) [87]. This may aid visual interpretation by suggesting that if the missing studies come from a "nonsignificance area," then asymmetry may be due to selection bias. However, if the missing studies come from a "significance-area" or

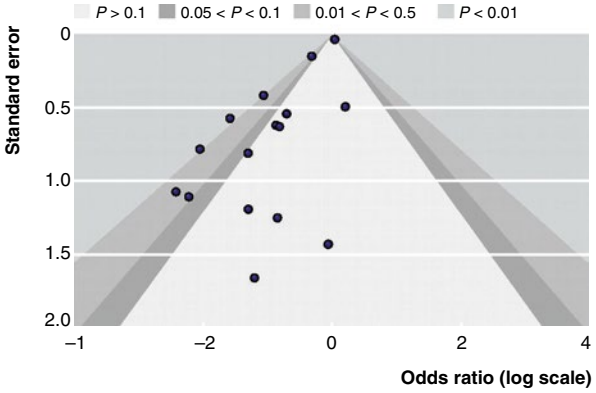

Fig. 12.6 Contour-enhanced funnel plot for trials of the effect of intravenous magnesium on mortality after myocardial infarction. Example of asymmetrical funnel plot (Reproduced with permission [81])

there is a certain direction of the intervention effect, then asymmetry is probably due to factors other than selection bias [81].

12.4.2 Tests for Small-Study Effects and Selection Bias

12.4.2.1 Funnel Plot-Based Tests

Apart from assessing for small-study effects using a visual inspection of funnel plots, several tests have been suggested to statistically assess funnel plot asymmetry. The tests are categorized as (1) rank-correlation tests or (2) linear regression tests. Begg and Mazumdar used a nonparametric rank-correlation method for the examination of the association between the standardized intervention effect estimates and their SEs [88]. When small studies (with large SEs) tend to have larger intervention effect estimates compared to the larger studies, the test identifies a correlation between the two factors. However, the test is associated with low power, and Begg suggests using a very liberal significance level (such as 0.10) [89]. Gjerdevik and Heuch suggested modification of Begg test based on Spearman rho and Kendall tau, to improve type I and II error rates; they suggested that the test based on Spearman rho is preferred for small datasets [90]. Egger et al. proposed a more powerful test compared to Begg test to assess the funnel plot asymmetry based on a regression analysis of Galbraith plot (see also Sect. 12.2.1.2) [83]. The test is based on the weighted linear regression of the standardized intervention effect (z-score) against study precision, with weights equal to the inverse of the variance. The intercept of the regression is used to measure asymmetry; specifically if it is estimated to be statistically significantly different from 0, then there is evidence of selection bias, and a negative intercept would suggest small-study effects are present. Tang and Liu suggested an alternative test using a linear regression of intervention effect estimate on $1/\sqrt{n}$, with weights n the study size [91].

Several modifications of the tests have been presented in the literature, which apply to dichotomous outcome data only. More specifically, for group correlation, the test by Schwarzer et al. could be used [92]. For linear regression, several modifications have been proposed including those by Macaskill et al. [93], Harbord et al. [94], Peters et al. [95], and the "arcsine" test by Rücker et al. [96]. For all aforementioned tests, the cutoff P-value 0.10 is considered to infer asymmetry in the funnel plot.

More specifically, the test proposed by Macaskill et al. is a linear regression of the intervention effect estimate on n, with weights $m_E m_{NE}/n$, where m_E and m_{NE} represent the total number of events and nonevents, respectively [93]. Harbord et al. [94] presented a modified version of the test proposed by Egger et al. [83], based on the efficient score ($Z = a - m_E n_E / n$) and its variance ($V = n_E n_C m_E m_{NE} / n^2 (n-1)$) of the log-odds ratio, where n_E and n_C are the sample sizes of the experimental and control groups, respectively. Peters et al. [95] suggested a slightly modified test compared to Macaskill et al. [93] test using the log-odds ratio effect measure and a linear regression of intervention effect estimate on $1/n$, with weights $m_E m_{NE}/n$, for a better control of type I error. Schwarzer et al. [92] suggested a rank-correlation test

for sparse data, using mean and variance of the noncentral hypergeometric distribution and avoiding correlation between log-odds ratio and its SE. However, for large between-study heterogeneity, the test has low power compared to the other tests [92]. Although the tests by Harbord et al. [94], Peters et al. [95], and Schwarzer et al. [92] have been presented using the odds ratio effect measure, they can be applied for other effect measures too. However, for a dichotomous outcome and the log-odds ratio or log-risk ratio, the intervention effect is statistically dependent on its variance, and hence tests based on these two factors might erroneously suggest the small-study effects' presence. Rücker et al. [96] suggested a test based on arcsine transformation of observed risks avoiding false-positive results when a large intervention effect or substantial between-study heterogeneity is present. In contrast to the other tests, the one suggested by Rücker et al. [96] can model studies with zero events in both arms.

Sterne et al. [81] advise using regression tests to address selection bias and small-study effects as they have larger power compared to rank tests as well as avoiding tests for funnel plot asymmetry if all studies are of similar sizes and hence of similar SEs. The Egger test has greater power for continuous outcomes than for dichotomous outcomes and is suggested for testing for funnel plot asymmetry. For dichotomous outcomes, the Harbord, Peters, and Rücker tests are suggested, as they have greater power compared to the other tests and avoid the mathematical association between log-odds ratio and its SE (this is also known as "regression to the mean"). It should be noted though that the performance of the tests deteriorates as the between-study heterogeneity increases. A general recommendation is to select one of the Harbord, Peters, and Rücker tests for small heterogeneity ($\tau^2 < 0.1$) and to use Rücker test for large heterogeneity ($\tau^2 > 0.1$) [3, 81].

12.4.3 Adjusting Intervention Effect Estimates for Small-Study Effects

12.4.3.1 Trim-and-Fill Method

The trim-and-fill method is a nonparametric method and aims to correct for funnel plot asymmetry due to small-study effects. The method is a four-step process:

1. The smaller studies are "trimmed" (i.e., removed) so that a symmetrical funnel plot is produced.
2. The summary intervention effect from the "trimmed" funnel plot is estimated.
3. The omitted studies are returned to the funnel plot and their "missing counterparts" are imputed or "filled" as their mirror images.
4. An adjusted overall intervention effect with its corresponding confidence interval is estimated using the complete set of studies [97, 98].

This is a nonparametric method and provides an estimate of both the number of missing studies and of the summary intervention effect adjusted for selection bias.

Although no assumptions are required about the mechanism leading to selection bias, the trim-and-fill method assumes that the small-study effect is solely caused by selection bias and that in truth there should be a symmetric funnel plot. However, the adjusted intervention effect should be interpreted with caution as it is not necessarily the intervention effect that would have been observed in the absence of selection bias.

Simulation studies have shown that the method performs well in the presence of selection bias, but it underestimates the intervention effect when there is large between-study heterogeneity and no selection bias [99, 100].

12.4.3.2 Selection Models

To evaluate the potential impact of missing studies on the results of a meta-analysis, selection models have been suggested that account for the mechanism by which studies are published. Selection models assume that missing studies are not missing at random, and the observed studies are due to certain characteristics (e.g., sample size, quality of design) that increase their propensity for publication. These models associate each observed study with an a priori probability to be published, and then estimate the summary intervention effect from the distribution of the observed sample.

A popular selection model in meta-analysis is the one developed by Copas [101], in which the probability that a study is observed depends on its SE. Although selection models correct effect estimates for selection bias, they have not been widely used probably because of their complexity, the large number of studies needed and the strong modeling assumptions about the severity of selection bias (i.e., that the factor causing small-study effects is selection bias). Copas [101] suggested applying a sensitivity analysis so that the researcher has the full picture of the estimated values of the intervention effect (and its uncertainty) under a range of assumptions about the severity of selection bias. It has been alternatively suggested to use expert opinion to inform the probabilities of publication [102]. A Copas selection model accounts for the correlation between the observed intervention effect and the probability that a study is published, which is:

1. Zero in the absence of selection bias.
2. Positive for a large intervention effect and large propensity for publication (e.g., for safety outcomes).
3. Negative for a large intervention effect and small propensity for publication (e.g., for efficacy outcomes; harms are less likely to be studied in trials and hence less likely to be published) [101, 103, 104].

Empirical studies using large collections of meta-analyses with dichotomous data suggest that the Copas selection model is preferable than the trim-and-fill method, as the latter produces systematically larger SEs and *P*-values [105, 106].

12.4.3.3 Extrapolation Methods

Extrapolation approaches model the relationship between the observed intervention effects and a measure of their uncertainty (e.g., SE). Stanley [107] and Copas and Malley [108] are early proponents of the regression-based approaches, with Stanley [107] adjusting the estimated intervention effect and Copas and Malley [108] adjusting the *P*-values for small-study effects. The approach suggested by Moreno et al. [109, 110] regresses the study-specific effects against their precision and computes the "unbiased" intervention effect as the extrapolation of the regression line to predict the intervention effect in a study with infinite sample size (or zero SE). The slope of the meta-regression is used to test for funnel plot asymmetry (see also Sect. 12.4.2.1), and the intercept is interpreted as the estimated intervention effect of a study with infinite sample size and hence infinite precision, adjusted for selection bias.

A key concern in these methods, as already stated in Sect. 12.4.2.1, is the mathematical association between some effect measures (e.g., log-odds ratio) and its SE, which might erroneously suggest the presence of small-study effects. Also, the performance of these methods depends on the variability of the meta-analysis' study sizes; if, for example, all studies are small, then the methods will not perform well. The regression-based methods, as any meta-regression model, suffer from lack of power to detect existing associations when few studies are available and in the presence of substantial heterogeneity. Simulation studies suggest that extrapolation within funnel plots outperform the trim-and-fill method, but still the adjusted effect estimates should be interpreted with caution [4, 109].

12.5 Moderators and Confounders

The impact of moderators and confounders is best viewed in light of the prior sections on heterogeneity issues and small-study effects, as any meaningfully important moderator or confounder is likely going to have an impact on homogeneity and symmetry of effects. The typical approaches to moderator and confounders include subgroup analyses and regression methods, which can be undertaken in the context of meta-analysis as well as more comprehensive overviews of reviews. As always, it remains important to recognize the presence of clustering and to minimize, especially in umbrella reviews, the risk of duplicate entry of trials with multiple arms as this may have a biasing effect on the accuracy and precision of the overall estimates.

12.6 Discussion

This chapter illustrates a vast range of approaches to evaluate the presence and estimate the magnitude of between-study heterogeneity as well as a wide variety of methods to test and adjust for small-study effects, which can easily be extrapolated to the analysis of key moderators and confounders. Heterogeneity and selection bias

are two of the greatest threats in meta-analysis and may lead to meaningless and/or overoptimistic intervention effect estimates. Researchers should routinely address and explore reasons for their presence and assess the extent to which these may influence the meta-analysis results.

Recent methodological research supports use of the random-effects model when completing a meta-analysis because it accounts for the between-study heterogeneity [3, 111, 112]. The random-effects model is considered more realistic than the fixed-effect model in most contexts. The new methodologies in meta-analysis help us incorporate heterogeneity and adjust for small-study effects and general funnel plot asymmetries as parts of the modeling that can also be reflected in the results. As presented in this chapter, both heterogeneity and selection bias can be examined using graphical methods, statistical tests, subgroup, and meta-regression analyses.

When selection bias is present, it is advisable that researchers make efforts to reduce (or if possible to eliminate) it, such as identifying unpublished or difficult to locate material from the "gray" literature for potential inclusion in the meta-analysis [113]. Also, exploration of heterogeneity should always take place when conducting a meta-analysis but should be interpreted with caution if individual participant data is not used in the statistical modeling. When few studies are included in a meta-analysis, we suggest conducting a sensitivity analysis using a variety of methods for addressing heterogeneity and small-study effects, before reaching definitive conclusions.

Acknowledgements AAV is funded by the CIHR Banting Postdoctoral Fellowship Program.

We would also like to thank Dr. Sharon E. Straus for her comments on a previous draft of this chapter.

References

1. Sterne JA, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. J Clin Epidemiol. 2000;53(11):1119–29.
2. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Random-effects model. Introduction to meta-analysis [Internet]. John Wiley & Sons, Ltd; 2009 [cited 2014]. p. 69–75. Available from: http://onlinelibrary.wiley.com/doi/10.1002/9780470743386.ch12/summary.
3. Higgins JPT, Green S. Cochrane handbook for systematic reviews of interventions [Internet]. Version 5.1.0. The Cochrane Collaboration; 2011. Available from: www.cochrane-handbook.org.
4. Rücker G, Carpenter JR, Schwarzer G. Detecting and adjusting for small-study effects in meta-analysis. Biom J. 2011;53(2):351–68.
5. Veroniki AA, Jackson D, Viechtbauer W, Bender R, Bowden J, Knapp G, Kuss O, Higgins JPT, Langan D, Salanti G. Methods to estimate the between-study variance and its uncertainty in meta-analysis. Res Synth Method. 2015. doi: 10.1002/jrsm.1164.
6. Mavridis D, Salanti G. Exploring and accounting for publication bias in mental health: a brief overview of methods. Evid Based Ment Health. 2014;17(1):11–5.
7. Anzures-Cabrera J, Higgins JPT. Graphical displays for meta-analysis: an overview with suggestions for practice. Res Synth Methods. 2010;1(1):66–80.

8. Bax L. MIX 2.0 – professional software for meta-analysis in Excel [Internet]. BiostatXL; 2011. Available from: http://www.meta-analysis-made-easy.com.
9. Tricco AC, Alateeq A, Tashkandi M, Mamdani M, Al-Omran M, Straus SE. Histamine H2 receptor antagonists for decreasing gastrointestinal harms in adults using acetylsalicylic acid: systematic review and meta-analysis. Open Med. 2012;6(3):e109–17.
10. Egger M, Smith GD, Phillips AN. Meta-analysis: principles and procedures. BMJ. 1997;315(7121):1533–7.
11. Ried K. Interpreting and understanding meta-analysis graphs – a practical guide. Aust Fam Physician. 2006;35(8):635–8.
12. Moja L, Moschetti I, Liberati A, Gensini GF, Gusinu R. Understanding systematic reviews: the meta-analysis graph (also called "forest plot"). Intern Emerg Med. 2007;2(2):140–2.
13. DuMouchel W. Predictive cross-validation of Bayesian meta-analyses. In: Bernardo JM, Berger JO, Dawid AP, Smith AFM, editors. Bayesian statistics 5. Oxford: Oxford University Press; 1996. p. 107–27.
14. Thompson SG. Controversies in meta-analysis: the case of the trials of serum cholesterol reduction. Stat Methods Med Res. 1993;2(2):173–92.
15. Galbraith RF. Some applications of radial plots. J Am Stat Assoc. 1994;89(428):1232–42.
16. Galbraith RF. A note on graphical presentation of estimated odds ratios from several clinical trials. Stat Med. 1988;7(8):889–94.
17. Song F. Exploring heterogeneity in meta-analysis: is the L'Abbé plot useful? J Clin Epidemiol. 1999;52(8):725–30.
18. L'Abbé KA, Detsky AS, O'Rourke K. Meta-analysis in clinical research. Ann Intern Med. 1987;107(2):224–33.
19. Baujat B, Mahé C, Pignon J-P, Hill C. A graphical method for exploring heterogeneity in meta-analyses: application to a meta-analysis of 65 trials. Stat Med. 2002;21(18):2641–52.
20. Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. Stat Med. 1996;15(6):619–29.
21. Viechtbauer W. Hypothesis tests for population heterogeneity in meta-analysis. Br J Math Stat Psychol. 2007;60(1):29–60.
22. Cochran WG. The combination of estimates from different experiments. Biometrics. 1954;10(1):101–29.
23. DerSimonian R, Kacker R. Random-effects model for meta-analysis of clinical trials: an update. Contemp Clin Trials. 2007;28(2):105–14.
24. Jackson D. Confidence intervals for the between-study variance in random effects meta-analysis using generalised Cochran heterogeneity statistics. Res Synth Methods. 2013;4(3):220–9.
25. Biggerstaff BJ, Tweedie RL. Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. Stat Med. 1997;16(7):753–68.
26. Biggerstaff BJ, Jackson D. The exact distribution of Cochran's heterogeneity statistic in one-way random effects meta-analysis. Stat Med. 2008;27(29):6093–110.
27. Hardy RJ, Thompson SG. Detecting and describing heterogeneity in meta-analysis. Stat Med. 1998;17(8):841–56.
28. Fleiss JL. Analysis of data from multiclinic trials. Control Clin Trials. 1986;7(4):267–75.
29. Petitti DB. Approaches to heterogeneity in meta-analysis. Stat Med. 2001;20(23):3625–33.
30. Barbui C, Hotopf M, Freemantle N, Boynton J, Churchill R, Eccles MP, et al. WITHDRAWN: treatment discontinuation with selective serotonin reuptake inhibitors (SSRIs) versus tricyclic antidepressants (TCAs). Cochrane Database Syst Rev. 2006;3:CD002791.
31. Rücker G, Schwarzer G, Carpenter JR, Binder H, Schumacher M. Treatment-effect estimates adjusted for small-study effects via a limit meta-analysis. Biostatistics. 2011;12(1):122–42.
32. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002;21(11):1539–58.
33. Wetterslev J, Thorlund K, Brok J, Gluud C. Estimating required information size by quantifying diversity in random-effects model meta-analyses. BMC Med Res Methodol. 2009;9(1):86.
34. Huedo-Medina TB, Sánchez-Meca J, Marín-Martínez F, Botella J. Assessing heterogeneity in meta-analysis: Q statistic or I2 index? Psychol Methods. 2006;11(2):193–206.

35. Zou GY. On the estimation of additive interaction by use of the four-by-two table and beyond. Am J Epidemiol. 2008;168(2):212–24.
36. Donner A, Zou GY. Closed-form confidence intervals for functions of the normal mean and standard deviation. Stat Methods Med Res. 2012;21(4):347–59.
37. Bowden J, Tierney JF, Copas AJ, Burdett S. Quantifying, displaying and accounting for heterogeneity in the meta-analysis of RCTs using standard and generalised Q statistics. BMC Med Res Methodol. 2011;11(1):41.
38. Birge RT. The calculation of errors by the method of least squares. Phys Rev. 1932;40(2):207–27.
39. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986;7(3):177–88.
40. Mittlböck M, Heinzl H. A simulation study comparing properties of heterogeneity measures in meta-analyses. Stat Med. 2006;25(24):4321–33.
41. Rücker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on I(2) in assessing heterogeneity may mislead. BMC Med Res Methodol. 2008;8:79.
42. Thorlund K, Imberger G, Johnston BC, Walsh M, Awad T, Thabane L, et al. Evolution of heterogeneity (I2) estimates and their 95% confidence intervals in large meta-analyses. PLoS One. 2012;7(7):e39471.
43. Paule RC, Mandel J. Consensus values and weighting factors [Internet]. National Institute of Standards and Technology; 1982 [cited 25 Mar 2014]. Available from: http://archive.org/details/jresv87n5p377_A1b.
44. The Nordic Cochrane Centre. Review manager (RevMan) [Computer program]. Version 5.3. Copenhagen: The Cochrane Collaboration; 2014.
45. Cornell JE, Mulrow CD, Localio R, Stack CB, Meibohm AR, Guallar E, et al. Random-effects meta-analysis of inconsistent effects: a time for change. Ann Intern Med. 2014;160(4):267–70.
46. Raudenbush SW. Analyzing effect sizes: Random-effects models. In: Cooper H, Hedges LV, Valentine JC, (Eds.). The handbook of research synthesis and meta-analysis (2nd ed). New York: Russell Sage Foundation; 2009.
47. Sidik K, Jonkman JN. Simple heterogeneity variance estimation for meta-analysis. J R Stat Soc Ser C Appl Stat. 2005;54(2):367–84.
48. Rukhin AL. Estimating heterogeneity variance in meta-analysis. J R Stat Soc Ser B Stat Methodol. 2013;75(3):451–69.
49. Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta-analysis: a comparative study. Stat Med. 1995;14(24):2685–99.
50. Kontopantelis E, Springate DA, Reeves D. A re-analysis of the Cochrane Library data: the dangers of unobserved heterogeneity in meta-analyses. PLoS One. 2013;8(7):e69930.
51. Chung Y, Rabe-Hesketh S, Choi I-H. Avoiding zero between-study variance estimates in random-effects meta-analysis. Stat Med. 2013;32(23):4071–89.
52. Sidik K, Jonkman JN. A comparison of heterogeneity variance estimators in combining results of studies. Stat Med. 2007;26(9):1964–81.
53. Viechtbauer W. Confidence intervals for the amount of heterogeneity in meta-analysis. Stat Med. 2007;26(1):37–52.
54. Novianti PW, Roes KCB, van der Tweel I. Estimation of between-trial variance in sequential meta-analyses: a simulation study. Contemp Clin Trials. 2014;37(1):129–38.
55. Sidik K, Jonkman JN. A note on variance estimation in random effects meta-regression. J Biopharm Stat. 2005;15(5):823–38.
56. Viechtbauer W. Bias and efficiency of meta-analytic variance estimators in the random-effects model. J Edu Behav Stat. 2005;30(3):261–93.
57. Jackson D, Bowden J, Baker R. How does the DerSimonian and Laird procedure for random effects meta-analysis compare with its more efficient but harder to compute counterparts? J Stat Plan Inference. 2010;140(4):961–70.
58. Rukhin AL, Biggerstaff BJ, Vangel MG. Restricted maximum likelihood estimation of a common mean and the Mandel–Paule algorithm. J Stat Plan Inference. 2000;83(2):319–30.

59. Panityakul T, Bumrungsup C, Knapp G. On estimating residual heterogeneity in random-effects meta-regression: a comparative study. J Stat Theory Appl. 2013;12(3):253.
60. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. Stat Med. 1999;18(20):2693–708.
61. William H, Swallow JFM. Monte Carlo comparison of ANOVA, MIVQUE, REML, and ML estimators of variance components. Technometrics. 1984;26(1):47–57.
62. Goldstein H, Rasbash J. Improved approximations for multilevel models with binary responses. J R Stat Soc Ser A Stat Soc. 1996;159(3):505.
63. Thorlund K, Wetterslev J, Awad T, Thabane L, Gluud C. Comparison of statistical inferences from the DerSimonian–Laird and alternative random-effects model meta-analyses – an empirical assessment of 920 Cochrane primary outcome meta-analyses. Res Synth Methods. 2011;2(4):238–53.
64. Morris CN. Parametric empirical Bayes inference: theory and applications. J Am Stat Assoc. 1983;78(381):47.
65. Lambert PC, Sutton AJ, Burton PR, Abrams KR, Jones DR. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. Stat Med. 2005;24(15):2401–28.
66. Senn S. Trying to be precise about vagueness. Stat Med. 2007;26(7):1417–30.
67. Chung Y, Rabe-Hesketh S, Dorie V, Gelman A, Liu J. A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. Psychometrika. 2013;78(4):685–709.
68. Gelman A. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). Bayesian Anal. 2006;1(3):515–34.
69. Rhodes KM, Turner RM, Higgins JPT. Predictive distributions were developed for the extent of heterogeneity in meta-analyses of continuous outcome data. J Clin Epidemiol. 2015;68(1):52–60.
70. Turner RM, Davey J, Clarke MJ, Thompson SG, Higgins JPT. Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane Database of Systematic Reviews. Int J Epidemiol. 2012;41(3):818–27. doi: 10.1093/ije/dys041.
71. Pullenayegum EM. An informed reference prior for between-study heterogeneity in meta-analyses of binary outcomes. Stat Med. 2011;30(26):3082–94.
72. Bailey KR. Inter-study differences: how should they influence the interpretation and analysis of results? Stat Med. 1987;6(3):351–60.
73. Friedrich JO, Adhikari NKJ, Beyene J. Ratio of means for analyzing continuous outcomes in meta-analysis performed as well as mean difference methods. J Clin Epidemiol. 2011;64(5):556–64.
74. Engels EA, Schmid CH, Terrin N, Olkin I, Lau J. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. Stat Med. 2000;19(13):1707–28.
75. Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. Stat Med. 2002;21(11):1575–600.
76. Lambert PC, Sutton AJ, Abrams KR, Jones DR. A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. J Clin Epidemiol. 2002;55(1):86–94.
77. Smith CT, Williamson PR, Marson AG. Investigating heterogeneity in an individual patient data meta-analysis of time to event outcomes. Stat Med. 2005;24(9):1307–19.
78. Riley RD, Higgins JPT, Deeks JJ. Interpretation of random effects meta-analyses. BMJ. 2011;342:d549.
79. Guddat C, Grouven U, Bender R, Skipka G. A note on the graphical presentation of prediction intervals in random-effects meta-analyses. Syst Rev. 2012;1:34.
80. Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. J R Stat Soc Ser A Stat Soc. 2009;172(1):137–59.
81. Sterne JAC, Sutton AJ, Ioannidis JPA, Terrin N, Jones DR, Lau J, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ. 2011;343:d4002.

82. Light RJ, Pillemer DB. Summing up: the science of reviewing research. Highlightingth ed. Cambridge: Harvard University Press; 1984. 191 p.
83. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ. 1997;315(7109):629–34.
84. Sterne JAC, Egger M, Smith GD. Investigating and dealing with publication and other biases in meta-analysis. BMJ. 2001;323(7304):101–5.
85. Ioannidis JPA, Trikalinos TA. The appropriateness of asymmetry tests for publication bias in meta-analyses: a large survey. CMAJ. 2007;176(8):1091–6.
86. Terrin N, Schmid CH, Lau J. In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. J Clin Epidemiol. 2005;58(9):894–901.
87. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. J Clin Epidemiol. 2008;61(10):991–6.
88. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. Biometrics. 1994;50(4):1088–101.
89. Cooper H, Hedges LV, editors. The handbook of research synthesis. 1st ed. New York: Russell Sage; 1994. 573 p.
90. Gjerdevik M, Heuch I. Improving the error rates of the Begg and Mazumdar test for publication bias in fixed effects meta-analysis. BMC Med Res Methodol. 2014;14:109.
91. Tang J-L, Liu JL. Misleading funnel plot for detection of bias in meta-analysis. J Clin Epidemiol. 2000;53(5):477–84.
92. Schwarzer G, Antes G, Schumacher M. A test for publication bias in meta-analysis with sparse binary data. Stat Med. 2007;26(4):721–33.
93. Macaskill P, Walter SD, Irwig L. A comparison of methods to detect publication bias in meta-analysis. Stat Med. 2001;20(4):641–54.
94. Harbord RM, Egger M, Sterne JAC. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. Stat Med. 2006;25(20):3443–57.
95. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Comparison of two methods to detect publication bias in meta-analysis. JAMA. 2006;295(6):676–80.
96. Rücker G, Schwarzer G, Carpenter J. Arcsine test for publication bias in meta-analyses with binary outcomes. Stat Med. 2008;27(5):746–63.
97. Duval S, Tweedie R. A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. J Am Stat Assoc. 2000;95(449):89–98.
98. Duval S, Tweedie R. Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics. 2000;56(2):455–63.
99. Terrin N, Schmid CH, Lau J, Olkin I. Adjusting for publication bias in the presence of heterogeneity. Stat Med. 2003;22(13):2113–26.
100. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. Stat Med. 2007;26(25):4544–62.
101. Copas J. What works?: selectivity models and meta-analysis. J R Stat Soc Ser A Stat Soc. 1999;162(1):95–109.
102. Mavridis D, Sutton A, Cipriani A, Salanti G. A fully Bayesian application of the Copas selection model for publication bias extended to network meta-analysis. Stat Med. 2013;32(1):51–66.
103. Copas JB, Shi JQ. A sensitivity analysis for publication bias in systematic reviews. Stat Methods Med Res. 2001;10(4):251–65.
104. Copas J, Shi JQ. Meta-analysis, funnel plots and sensitivity analysis. Biostat Oxf Engl. 2000;1(3):247–62.
105. Schwarzer G, Carpenter J, Rücker G. Empirical evaluation suggests Copas selection model preferable to trim-and-fill method for selection bias in meta-analysis. J Clin Epidemiol. 2010;63(3):282–8.

106. Carpenter JR, Schwarzer G, Rücker G, Künstler R. Empirical evaluation showed that the Copas selection model provided a useful summary in 80% of meta-analyses. J Clin Epidemiol. 2009;62(6):624–31.e4.
107. Stanley TD. Meta-regression methods for detecting and estimating empirical effects in the presence of publication selection. Oxf Bull Econ Stat. 2008;70(1):103–27.
108. Copas JB, Malley PF. A robust P-value for treatment effect in meta-analysis with publication bias. Stat Med. 2008;27(21):4267–78.
109. Moreno SG, Sutton AJ, Ades AE, Stanley TD, Abrams KR, Peters JL, et al. Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. BMC Med Res Methodol. 2009;9(1):2.
110. Moreno SG, Sutton AJ, Ades AE, Cooper NJ, Abrams KR. Adjusting for publication biases across similar interventions performed well when compared with gold standard data. J Clin Epidemiol. 2011;64(11):1230–41.
111. Schmidt FL, Oh I-S, Hayes TL. Fixed- versus random-effects models in meta-analysis: model properties and an empirical comparison of differences in results. Br J Math Stat Psychol. 2009;62(1):97–128.
112. Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. Stat Med. 2001;20(6):825–40.
113. Searching grey literature: grey matters [Internet]. [cited 29 Mar 2015]. Available from: http://www.nccmt.ca/registry/view/eng/130.html.