Demystifying fixed and random effects meta-analysis

doi:10.1136/eb-2014-101795

ABSTRACT

Objective Systematic reviewers often need to choose between two statistical methods when synthesising evidence in a meta-analysis: the fixed effect and the random effects models. The two approaches entail different assumptions about the treatment effect in the included studies. The aim of this paper was to explain the assumptions underlying each model and their implications in the interpretation of summary results.

Methods We discussed the key assumptions underlying the two methods and the subsequent implications on interpreting results. We used two illustrative examples from a published meta-analysis and highlighted differences in results.

Results The two meta-analytic approaches may yield similar or contradicting results. Even if results between the two models are similar, summary estimates should be interpreted in a different way.

Conclusions Selection between fixed or random effects should be based on the clinical relevance of the assumptions that characterise each approach. Researchers should consider the implications of the analysis model in the interpretation of the findings and use prediction intervals in the random effects meta-analysis.

INTRODUCTION

Meta-analysis is an established tool for evidence-based health decision-making and has a prominent role with Health Technology Assessment agencies. It synthesises information about a particular effect of interest (such as the impact of a risk factor or the effect of a treatment) from relevant studies and is typically a two-stage process. At the first stage, a statistic of interest that measures the effect, often called the effect size (ES), is computed for each study. Typical examples of ESs are the risk ratio (RR) or odds ratio (OR) for dichotomous outcomes and mean difference (MD) or standardised mean difference (SMD) for continuous outcomes. An ES in a given study reflects the magnitude of the effect and assuming that the design and conduct of the study is unflawed, the ES can be interpreted as the most reliable estimate of the true effect in that study. Uncertainty around each study's ES is expressed by the CI and depends mainly on the sample size of each study (and the number of events for dichotomous outcomes). At the second stage, the study-specific ESs are synthesised to produce a summary effect. There are two approaches for synthesising studies; the fixed-effect (FE) and the random effects (RE) models.[1–6]

The selection of the appropriate model to synthesise study findings, adequate justification of the choice and correct interpretation of the meta-analysis result are of great importance but empirical evidence suggests that published meta-analyses do not consider these aspects very often.[7–9] In the present article we aim to elucidate the conceptual and technical differences between the two models, explain why choice of model matters and how the results from each model should be interpreted. We will consider the case of treatment comparison, so that ES refers to the effect of an intervention compared to a reference treatment.

METHODS

The basic assumption of the FE model is that the treatment effect is the same (fixed) in all studies included in the meta-analysis, whereas the RE model allows the treatment effect to vary across studies. Mathematically, the summary effect is a weighted average of the observed ESs in FE and RE models. However, FE and RE differ in the way studies are weighted and in the interpretation of the summary effect.

FE model (FE meta-analysis)

The FE model has dominated the field for many years since the first meta-analyses were published in the late 1970s–early 1980s. Especially in psychology, FE was the rule rather than the exception[10] and until 2006 more than three quarters of meta-analyses were conducted using an FE approach.[3] The prevalence of FE has been also noted in early Cochrane reviews.[11] An FE model assumes that all studies share a common true treatment effect; the relative effect of the treatment compared with the reference is the same in all study settings. Differences between observed ESs are solely attributed to random/sampling error; had all studies infinitely large sample sizes, the sampling variation within a study would have been diminished and all observed study estimates ES would have been equal to the common 'true' relative treatment effect.

The summary treatment effect in FE model is a weighted summary of the study-specific ESs. The weights assigned to each study depend on the study's precision; more specifically each study's weight is equal to the inverse of its variance. The larger the sample size in a trial, the smaller the variance of the ES and the larger the corresponding weight assigned in the meta-analysis. Thus, in an FE model bigger studies contribute more in the estimation of the summary effect and are assigned larger weights, whereas smaller studies convey less information and are assigned smaller weights. An intuitive explanation is that since the effect is the same in all study settings it is sensible to trust more the largest ones. If all studies were of equal precision (eg, equal sample size) then the summary effect obtained from the FE meta-analysis would equal the arithmetic mean of the observed ESs.

The assumption of a common effect implies that the studies are sufficiently similar in the aspects that might modify the treatment effect. These include population characteristics (eg, age of the participants and baseline risk of the population), study design characteristics (such as duration of follow-up), intervention characteristics (eg, dose and modality) and others. Moreover, the summary treatment effect applies to the common (fixed) setting that is studied in the included trials. Hence, systematic reviewers can employ an FE model when two conditions are met; there is strong evidence that all trials are functionally identical and inference is limited to the population included in the analysis.[1] This assumption should be stated explicitly in the protocol of a systematic review.

RE model (RE meta-analysis)

In several clinical settings it may not be realistic to assume that treatment effects are invariant to study settings. For example, the effectiveness of an antipsychotic might be established compared to placebo, but the magnitude of the benefit might vary in trials depending on whether they include only patients at their first psychotic episode or not. Variability in the 'true' treatment effect across studies (termed as heterogeneity) will result in variability in the observed ESs *additional* to the sampling variability. This is called *statistical heterogeneity* and possible explanations should be sought by examining the impact of patient and study characteristic as described elsewhere.[5 12] However, obstacles such as poor reporting of characteristics in individual studies might make explanation of heterogeneity

impossible. In such cases, an RE meta-analysis could account appropriately for the extra variability in the summary estimate.

In RE meta-analysis we assume that the observed ESs differ because the 'true' treatment effect is different in the settings studied in the trials and because of random error. Under this scenario, even if all studies were of infinite sample size the results in studies would still be variable because of real differences in study settings. Heterogeneity in studies may be caused by clinical factors that differ across settings such as variability in participants, outcomes, interventions and/or design and conduct study characteristics.[4]

An RE model estimates a summary effect accounting for both within study variability (expressed by the CI in each study's ES) but also for between study variability (heterogeneity). As before, the summary effect is estimated as a weighted average and the weights assigned to each observed ES equal the inverse of their variance plus an additional variance component that reflects heterogeneity and is denoted by τ^2.

The summary estimate under the RE model still obtains more information from the larger and more precise studies. However, the distribution of the weights is not as much contrasted as under an FE model, which is important to take into consideration when choosing between these two models (see section Selection of the appropriate model). The less contrasted weights across studies can be conceptually explained by the fact that each study represents a different setting and a different treatment effect. Small studies may be imprecise due to their small sample size but they still give information about effectiveness; in fact a study, small or large, is a unique source of information for the setting it considers. Although an RE model may be appropriate for meta-analysis in many clinical questions, a sizeable number of studies is needed to estimate adequately the heterogeneity and the RE weights.

Comparison of FE and RE meta-analyses

Figure 1 attempts to delineate the conceptual and computational differences between the two models using an imaginary example. Both panels contain the same 12 studies. Each blue square is a study-specific observed ES. Figure 1 left panel shows results from an FE analysis. Observed ES refers to the common fixed treatment effect (vertical solid line/centre of the diamond). Deviations of observed estimates from the summary effect are believed to relate only to random error. The observed ESs are synthesised to obtain the summary effect (the diamond at the bottom), accounting only for sampling variability.

Figure 1 right panel shows an RE meta-analysis. Here it is assumed that the observed ESs in study settings estimate with random error (heuristically expressed in solid horizontal lines) different treatment effects (dotted vertical lines) which are unobserved. Differences between observed ESs are partly attributed to random error and partly to real differences in treatment effects across settings. The setting-specific 'true' treatment effects are different yet related as they all come from a common (normal) distribution shown at the bottom of the right panel. This is often called *the random effects distribution* and it is of clinical interest as it shows the range of the true treatment effect in the various settings. A normal distribution is characterised by its mean and variance; the mean of the REs distribution is the summary treatment effect (the purple diamond) whereas the variance the heterogeneity τ^2.

Both sample size and heterogeneity are taken into account when assigning weights to studies in an RE analysis resulting in more balanced relative weights compared to those of an FE analysis. Very large studies are weighted heavily in the FE model whereas their influence is restricted under the RE model. The amount the relative weights differ between the two models depends on the extent of heterogeneity. The variance

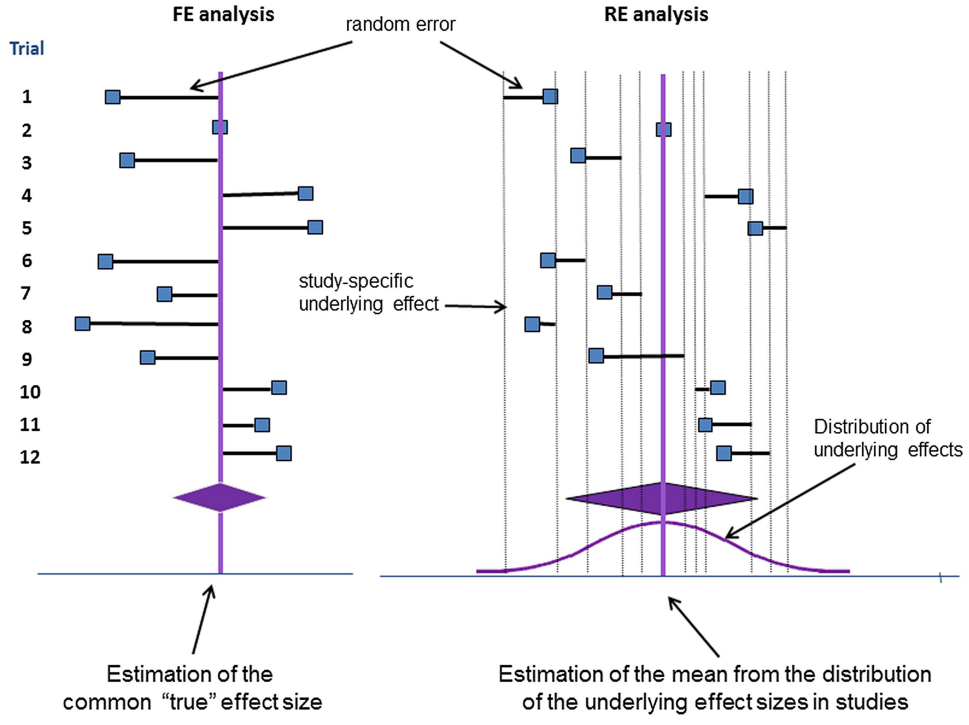

Figure 1 Forest plots from a fixed effect (left panel) and a random effects (right panel) meta-analysis. Observed effect sizes (squares) are identical both figures. Solid horizontal lines relate to the random error in each trial. Dotted lines in the right panel represent the underlying study-specific 'true' relative treatment effects.

under the RE model is calculated as the sum of the within study variances plus heterogeneity. The relative ratio of these two sources of uncertainty has an impact on the weighting scheme under the RE model. If heterogeneity is small, the weights assigned under an RE model will be similar to those under an FE model. If heterogeneity is large then weights under an RE model will be mainly driven by heterogeneity and will be substantially different from those obtained using an FE model.

If it is believed that effectiveness varies with studies, more uncertainty is associated with the summary estimate so that it captures all variability in the studies. The 95% CI around the summary estimate from an RE analysis is often larger than the corresponding interval from an FE analysis; hence RE analysis often yields more conservative results. This may not be true in the presence of small study effects; the phenomenon where small studies give different results from large studies. When smaller trials are associated with exaggerated ESs then the FE model yields a more conservative (closer to null) summary effect than the RE model. This is because the influence of the small studies is greater under the RE model and the summary estimate will be pooled towards the results from smaller studies.[13] Actually, the comparison of results from RE and FE meta-analyses is a method to detect the presence of small study effects and suggests that the assumed distribution for the underlying study effects does not hold.[1] Small study effects can be also identified by funnel plot asymmetry and Sterne *et al*[14] recommend its use in choosing between the two models. Table 1 summarises several scenarios where the two meta-analysis models may yield similar or contradicting results.

Interpretation of FE and RE meta-analyses

The conceptual differences between FE and RE suggest that the interpretation of the summary estimate depends heavily on the method used to synthesise the study findings. In the FE model the summary effect is the best estimate of the common treatment effect and together with its uncertainty they are the only information of relevance. The RE summary effect is often misinterpreted as if it were the estimate of a common overall effect[15] whereas in reality it is an estimation of the average of a collection of possible treatment effects in various settings. As the 'average effect' might not actually occur in any setting, the entire distribution of possible effects (as shown in figure 1 right panel) might be of greater interest than the just the mean. As an alternative to present the entire distribution of possible effects, it has been suggested to include predictive intervals in an RE meta-analysis result. Predictive intervals refer to the 95% of the possible treatment effects in individual settings and can be interpreted as the predicted range for the 'true' treatment effect in a new study.[7 15] CIs and predictive intervals are not interchangeable; CIs reflect uncertainty about the estimation of the mean whereas predictive intervals express the dispersion of the true ESs. The latter is of importance in an RE model that allows inference for studies that are not included in the meta-analysis.

Table 1 Factors that impact on the meta-analysis summary estimate under the FE and the RE models

Scenario	Impact on results
Heterogeneity variance is estimated to be 0	The results are identical under the two meta-analyses models
Heterogeneity is greater than 0 and the studies' size is unrelated to its effect size	The two models yield similar mean summary estimates. The CI of the RE meta-analysis is wider than that of the FE meta-analysis
Small studies are associated with more extreme (eg, more beneficial for the active intervention) effect sizes (small study effect)	Under the FE meta-analysis, the summary estimate tends to be pooled towards the summary effect from the larger studies (usually towards the no effect line). The influence of the big studies is less pronounced under the RE meta-analysis, resulting in a less conservative result
A mega-trial is included in the meta-analysis	Under the FE meta-analysis, the summary estimate tends to be pooled towards the effect of the mega-trial. The influence of the mega-trial is less pronounced under the RE meta-analysis

FE, fixed-effect; RE, random effect.

Quantifying heterogeneity

Clinical and/or methodological heterogeneity may result into statistical heterogeneity, that is, differences between observed ESs. There are several methods to quantify statistical heterogeneity and their performance varies depending on the magnitude and the number of studies available. DerSimonian and Laird[16] estimator together with a maximum likelihood estimator are available in most statistical software. Interpretation of τ^2 to infer about its magnitude depends on the studied outcome, the treatments compared and the summary treatment effect. Recent empirical evidence pinpointed what constitutes a large, average or small heterogeneity value for dichotomous outcomes.[17] Another measure that is often used to quantify heterogeneity and is also a standard output in most software used for meta-analysis is an index, denoted by I^2, reflecting the proportion of variability in summary estimates that is attributed to heterogeneity rather than sampling error.[18]

Many meta-analyses do not include enough studies to estimate adequately the extent of heterogeneity. As a rule of thumb, at least three studies are needed to estimate heterogeneity. Lack of data to accurately estimate heterogeneity does not entail its absence. With few studies, the mean I^2 or τ^2 value might be 0 but their 95% CIs may include values of extreme heterogeneity.

Selection of the appropriate model

A common malpractice in the past has been to choose the meta-analysis model based on the significance of a test for homogeneity (the χ^2 test). It has been found that the test has low power when studies have small sample sizes or are few in number, and a lack of statistical significance does not guarantee the absence of heterogeneity.[19] This strategy is flawed and is strongly discouraged.[1] Higgins *et al*[8] argue that if there are clinical or methodological differences in the included studies then statistical heterogeneity is inevitable. Using an FE model in the presence of heterogeneity may result in underestimation of the treatment variability and in conclusions that do not apply to the settings of interest.

Researchers should consider the identified studies and examine whether they have been conducted under similar conditions and in similar populations, then the assumption underlying the FE model is likely to hold. Clinical understanding of the research question and an a priori hypothesis about whether the effect of interest is likely to be similar in various settings are crucial in deciding on the meta-analysis model.

Researchers should also consider the fact that small studies are assigned larger weights in an RE model compared to an FE model. Finally, it is important to note that RE analysis is not a remedy for extreme heterogeneity. Differences in effects, when

present should be explored and if possible explained via subgroup analyses or meta-regression.[5 12 20]

RESULTS

Leucht *et al*[21] examined the comparative efficacy and acceptability of 15 antipsychotic drugs in schizophrenia. They synthesised data from 212 trials and they examined seven outcomes using an RE meta-analysis.

Efficacy of risperidone versus haloperidol for reducing symptoms in schizophrenia

The primary outcome in the analysis was efficacy and it was measured by the mean overall change in symptoms on the PANSS scale. The relative ESs were calculated with the SMD. For the purposes of the present paper, we focused only on the head-to-head comparison between risperidone with haloperidol (13 studies). Figure 2 displays the forest plot from the FE and RE meta-analyses.

The heterogeneity variance τ^2 was estimated as 0.04 with $I^2 = 52\%$ suggesting a moderate amount of heterogeneity.[8] This was supported also by the χ^2 statistic which was equal to 24.82 with a corresponding p value equal to 0.02. The extent of heterogeneity resulted in the weights being quite different between the two approaches. Under the FE meta-analysis, Peuskens 1995 contributed more than one-third of the total information (36.17%) while its influence was more than halved in the RE meta-analysis (relative weight: 15.55%). In contrast, small studies like Min 1993 and Claus 1992 were assigned relatively larger weights under the RE compared to those of FE.

The summary SMD of the FE meta-analysis was −0.09 (95% CI −0.19 to 0) suggesting a relative benefit of risperidone over haloperidol, though the upper limit of the CI showed that there might be not difference between the two interventions. However, the result under the FE meta-analysis could only be supported if the studies are believed to be functionally identical. In this case the summary results along with its CI represent the best estimate of the common true treatment effect within the population included in the analysis and cannot be generalised in other settings.

The RE summary result was −0.10 (95% CI −0.27 to 0.06). As the CI included 0 there is no evidence that on an average risperidone is more effective than haloperidol. It may be that across studies real differences in treatment effectiveness exist and this can be represented by the prediction interval (−0.57 to 0.37). The prediction interval crossed 0 showing that neither intervention was more effective than the other in all considered settings. As it included values that could be seen as clinically important favouring either risperidone or haloperidol, we are not able to say whether the relative effectiveness in a new study would suggest a higher benefit with risperidone, haloperidol or no difference at all.

Acceptability of chlorpromazine versus placebo for all cause discontinuation in schizophrenia

In the same systematic review,[21] 11 studies compared the number of dropouts due to any reason (all cause discontinuation) for chlorpromazine versus placebo. The outcome was measured OR. The variation in the estimates attributed to heterogeneity was moderate to low ($I^2 = 34.2\%$) with $\tau^2 = 0.17$. The χ^2 test for heterogeneity was equal to 15.20 and yielded a p value=0.13 conventionally interpreted as supportive for an FE model. However, the mean ORs varied from 0.15 to 3.55 while some CIs did not overlap (see figure 3).

The FE summary estimate was 0.68 (95% CI 0.49 to 0.93) while under the RE estimate was 0.75 (95% CI 0.47 to 1.19). The FE meta-analysis suggested a significant reduction of dropout with chlorpromazine. In contrast, the RE meta-analysis suggested that chlorpromazine might be associated with more, same or larger dropouts compared to placebo. In this case the conclusions drawn from the two approaches are contradicting. However, the summary estimates from the two approaches should not be interpreted as if the two approaches were interchangeable. Under the assumption that all studies aim to estimate the same true relative effectiveness of chlorpromazine versus placebo, the FE estimate provides the best estimate of this effectiveness. Leucht *et al* acknowledged that several known (such as treatment dose, sponsorship and year of publication) or unknown factors may cause heterogeneity and therefore the RE setting might be more appropriate. The 95% predictive interval was 0.25 to 2.22 and reflects the expectation that in 95% of cases the true effect in a future study will lie within this range.

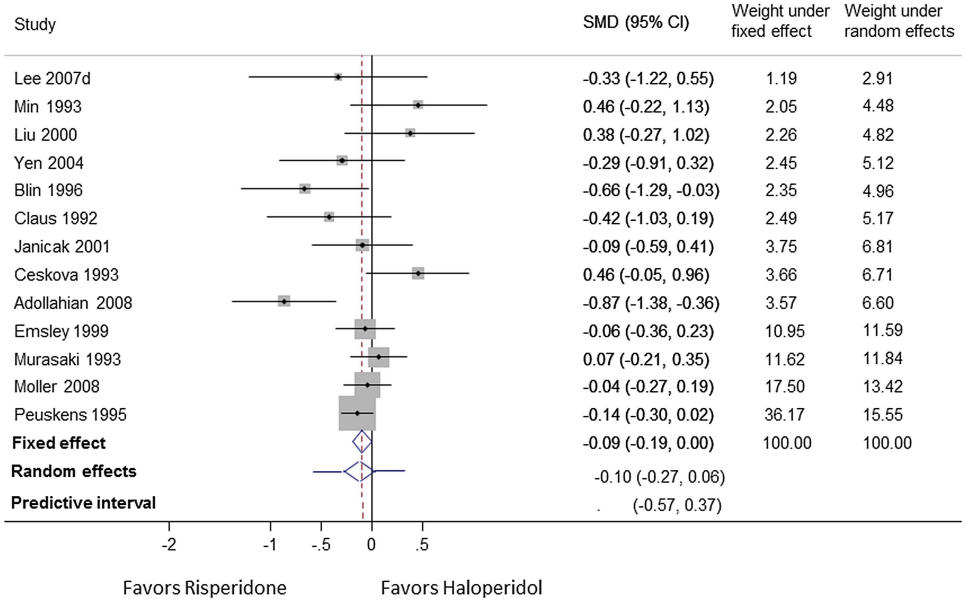

Figure 2 Forest plot displaying results for the efficacy of risperidone versus haloperidol for reducing symptoms in schizophrenia. Negative values of standardised mean difference favor risperidone. The studies are ordered by their sample size so that any presence of small study effects is depictable.

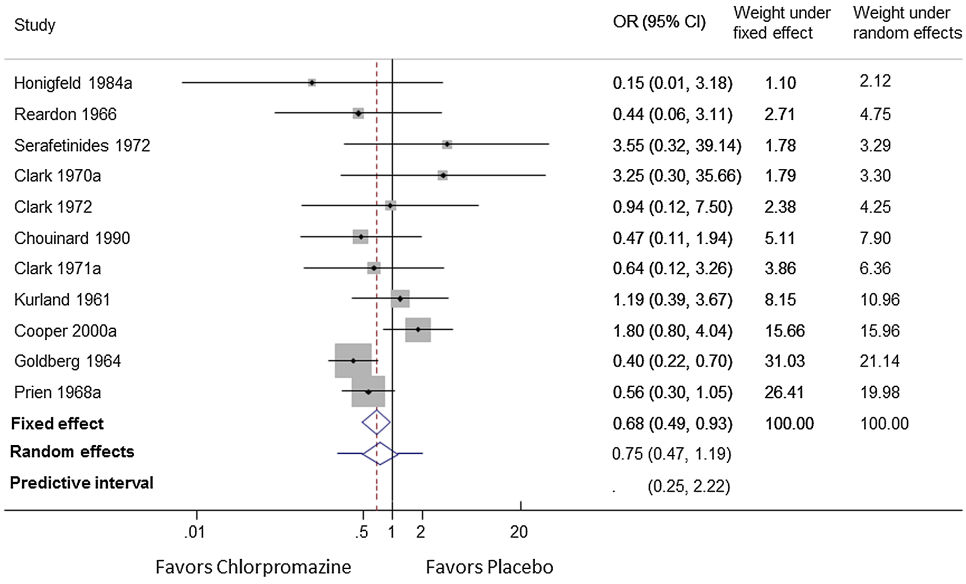

Figure 3 Forest plot displaying results for the acceptability of chlorpromazine versus placebo for all cause discontinuation in schizophrenia. Values of OR smaller than 1 indicate a relative reduction in dropout in favour of chlorpromazine compared with placebo. The studies are ordered by their sample size so that any presence of small study effects is depictable.

DISCUSSION

Dealing with heterogeneity is one of the most challenging tasks in meta-analysis. If possible effect modifiers are known or suspected then subgroups analysis and meta-regression may be employed to explore heterogeneity, identify its sources and estimate subgroup-specific treatment effects.[20] The RE model offers the benefit of allowing the true effects underlying the studies to differ and thus accounts for unexplained heterogeneity. Many investigators consider the RE approach to be a more natural choice than FE in medical research fields.[16 22 23] The choice of the model should not be made based on the test of heterogeneity as heterogeneity may exist even if it goes undetected from this test.

Unfortunately, the appropriate choice of the computational model is not a common practice. Schroll *et al*[11] examined 60 Cochrane reviews with substantial heterogeneity ($I^2 > 50\%$) in order to investigate how authors deal with this source of uncertainty. More than half used an FE model (55%) for their analysis and over a third (33%) had major methodological problems mainly due to the choice of computational model. Their conclusions concerning the problematic handling of heterogeneity and the choice of computational model were consistent with the findings of Higgins *et al*[24] and those of Hahn *et al*.[25]

Adriani Nikolakopoulou,[1] Dimitris Mavridis,[1,2] Georgia Salanti[1]

[1]Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece

[2]Department of Primary Education, University of Ioannina, Ioannina, Greece

Correspondence to: Dr Dimitris Mavridis, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, 45110, Greece; dimi.mavridis@googlemail.com

Competing interests AN, DM and GS received research funding from the European Research Council (IMMA 260559).

REFERENCES

1. **Borenstein M,** Hedges LV, Higgins JPT, *et al*. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res Synthesis Methods* 2010;**1**:60–86.
2. **Hedges Larry V,** Vevea JL. Fixed and random-effects models in meta-analysis. *Psychol Methods* 1998;**3**:486–504.
3. **Schmidt FL,** Oh IS, 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**:97–128.
4. **Higgins JPT,** Green S. *Cochrane handbook for systematic reviews of interventions*. Chichester: Wiley, 2008.
5. **Borenstein M,** Hedges LV, Higgins JPT, *et al*. *Introduction to meta-analysis*. Wiley, 2009.
6. **Guyatt GH,** Oxman AD, Schunemann HJ, *et al*. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. *J Clin Epidemiol* 2011;**64**:380–2.
7. **Higgins JP,** Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *J R Stat Soc Ser A Stat Soc* 2009;**172**:137–59.
8. **Higgins JP,** Thompson SG, Deeks JJ, *et al*. Measuring inconsistency in meta-analyses. *BMJ* 2003;**327**:557–60.
9. **Riley RD,** Gates S, Neilson J, *et al*. Statistical methods can be improved within Cochrane pregnancy and childbirth reviews. *J Clin Epidemiol* 2011;**64**:608–18.
10. **Schulze R.** *Meta-analysis: a comparison of approaches*. Toronto: Hogrefe & Huber, 2004.
11. **Schroll JB,** Moustgaard R, Gotzsche PC. Dealing with substantial heterogeneity in Cochrane reviews. Cross-sectional study. *BMC Med Res Methodol* 2011;**11**:22.
12. **Thompson SG.** Why sources of heterogeneity in meta-analysis should be investigated. *BMJ* 1994;**309**:1351–5.
13. **Poole C,** Greenland S. Random-effects meta-analyses are not always conservative. *Am J Epidemiol* 1999;**150**:469–75.
14. **Sterne JA,** Sutton AJ, Ioannidis JP, *et al*. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ* 2011;**343**:d4002.
15. **Riley RD,** Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. *BMJ* 2011;**342**:d549.
16. **DerSimonian R,** Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;**7**:177–88.
17. **Turner RM,** Davey J, Clarke MJ, *et al*. Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane Database of Systematic Reviews. *Int J Epidemiol* 2012;**41**:818–27.
18. **Higgins JP,** Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med* 2002;**21**:1539–58.
19. **Hardy RJ,** Thompson SG. Detecting and describing heterogeneity in meta-analysis. *Stat Med* 1998;**17**:841–56.
20. **Thompson SG,** Higgins JP. How should meta-regression analyses be undertaken and interpreted? *Stat Med* 2002;**21**:1559–73.
21. **Leucht S,** Cipriani A, Spineli L, *et al*. Comparative efficacy and tolerability of 15 antipsychotic drugs in schizophrenia: a multiple-treatments meta-analysis. *Lancet* 2013;**382**:951–62.
22. **Ades AE,** Lu G, Higgins JP. The interpretation of random-effects meta-analysis in decision models. *Med Decis Making* 2005;**25**:646–54.
23. **Fleiss JL,** Gross AJ. Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke and lung cancer: a critique. *J Clin Epidemiol* 1991;**44**:127–39.
24. **Higgins J,** Thompson S, Deeks J, *et al*. Statistical heterogeneity in systematic reviews of clinical trials: a critical appraisal of guidelines and practice. *J Health Serv Res Policy* 2002;**7**:51–61.
25. **Hahn S,** Garner P, Williamson P. Are systematic reviews taking heterogeneity into account? An analysis from the Infectious Diseases Module of the Cochrane Library. *J Eval Clin Pract* 2000;**6**:231–3.

Correction

Nikolakopoulou A, Mavridis D, Salanti G. Demystifying fixed and random effects meta-analysis. *Evid Based Mental Health* 2014;17:53–57. The authors apologise for the omission of the following Acknowledgement statement from their article: 'We acknowledge Professor Julian Higgins (University of Bristol) for Figure 1.'

Evid Based Mental Health 2014;**17**:89. doi:10.1136/ebmental-2014-101795corr1