Journal of Clinical Epidemiology 107 (2019) 101−106

ORIGINAL ARTICLE

Getting more out of meta-analyses: a new approach to meta-analysis in light of unexplained heterogeneity

Amit Saad[a,b,*,1], Daniel Yekutieli[c,1], Shaul Lev-Ran[b,d], Raz Gross[b,e,f], Gordon Guyatt[g]

[a]*Day Treatment Unit, Shalvata Mental Health Centre, Hod-Hsharon, Israel*
[b]*Department of Psychiatry, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel*
[c]*Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv, Israel*
[d]*Addiction Medicine and Dual Disorders Clinic, Lev-Hasharon Medical Centre, Tsur Moshe, Israel*
[e]*Division of Psychiatry, The Chaim Sheba Medical Centre, Tel-Hashomer, Israel*
[f]*Department of Epidemiology and Preventive Medicine, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel*
[g]*Department of Health Research Methods, Evidence and Impact, McMaster University, Hamilton, Ontario, Canada*

Accepted 30 November 2018; Published online 5 December 2018

Abstract

Background and Objectives: Meta-analyses sometimes summarize results in the presence of substantial unexplained between-study heterogeneity. As GRADE criteria highlight, unexplained heterogeneity reduces certainty in the evidence, resulting in limited confidence in average effect estimates. The aim of this paper is to provide a new clinically useful approach to estimating an intervention effect in light of unexplained heterogeneity.

Methods: We used a random-effects model to estimate the distribution of an intervention-effect across various groups of patients given data derived from meta-analysis. The model provides a distribution of the probabilities of various possible effects in a new group of patients. We examined how our method influenced the conclusions of two meta-analyses.

Results: In one example, our method illustrated that evidence from a meta-analysis did not support authors' highly publicized conclusion that hypericum is as effective as other antidepressants. In the second example, our method provided insight into a subgroup analysis of the effect of ribavirin in hepatitis C, demonstrating clear important benefit in one subgroup but not in others.

Conclusion: Analysing the distribution of an intervention-effect in random-effects models may enable clinicians to improve their understanding of the probability of particular-intervention effects in a new population.

Keywords: Meta-analyses; Heterogeneity; GRADE; Random-effects models; Systematic reviews; I^2 statistic, between study variance

1. Introduction

Meta-analyses of randomized controlled trials (RCTs) play an important role in clinical decision-making [1,2]. Frequently, however, results vary across studies, and attempts to explain variation by subgroup analysis often fail [2−10]. As GRADE criteria highlight, unexplained heterogeneity reduces certainty in evidence [11]. Hence, clinicians may not know the extent to which an estimated mean effect applies to a specific group of patients. To address this problem, statisticians have suggested taking into account the estimated effect in 95% of the population [12]. However, this approach is likely to yield very wide confidence intervals (CIs) that clinicians will find uninformative.

To provide a more clinically useful method to deal with unexplained heterogeneity, we suggest presenting the entire distribution of the heterogeneous effect. Doing so may

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' contribution: A.S. contributed to data collection, conceptual design of the model, data analysis, interpretation of the results, and wrote the first draft of the article. D.Y. (equal contribution) developed the statistical model, wrote the statistical code, and wrote the Supplement. S.L-R. and R.G. revised the article, contributed to the interpretation of the results, and finalized the article. G.G. contributed to interpretation of the results, clarifying the application of the model, and contributed substantially for finalizing the article. All authors read and approved the final version of the article.

The study was approved by Shalvata Mental Health Centre Research Ethics Board.

[1] These authors contributed equally.

* Corresponding author. Health Research Methods, Evidence and Impact, McMaster University, 1280 Main Street West, HSC 2C12, L8S 4K1, Hamilton, Ontario, Canada. Tel.: 1 905 525 9140.

E-mail address: guyatt@mcmaster.ca (A. Saad).

https://doi.org/10.1016/j.jclinepi.2018.11.023
0895-4356/© 2018 Elsevier Inc. All rights reserved.

What is New?

Key findings

- Assessing the distribution of an intervention effect across the RCTs in the meta-analysis can provide estimates of the probable distribution of effects in new groups of patients who are candidates for the interventions of interest.

What this adds to what was known?

- The concept is original, and the article provides a detailed analysis for computing and presenting the confidence-intervals of the distribution of the effect.

What is the implication and what should change now?

- Systematic review authors should consider use of this method to help clinicians to better estimate the expected effect in a new group of patients.

enrich clinical judgment, for clinicians would not need to assume that the expected effect in a new group of patients is the mean effect, with some level of uncertainty. Rather, they would be able to estimate the probability of a range of possible effects in a new group of patients. In the following, we present the essentials of the suggested approach and provide clinical examples showing how our method may enhance clinical decision-making.

2. Methods

The Supplement presents a detailed description of our method. Our goal was to describe the distribution of relative effects estimated by meta-analyses of RCTs comparing two interventions on dichotomous outcomes. We used a random-effects model with the assumption that the distribution of the relative risk (RR) across various groups of patients is normal with unknown mean μ and variance τ^2. We used maximum likelihood estimation (MLE) to derive a bivariate statistic (a function of the data, μ and τ). Using this MLE, we constructed the 95% CIs for μ, τ, and then calculated the cumulative distribution function (CDF) of the relative risk. The CDF provides a distribution of the probabilities of various possible RRs, given results of the meta-analysis. Data were analyzed using R.

3. Results

3.1. Clinical example—A

A Cochrane review examines the effectiveness of hypericum (St. John's wart) extracts for treating major depression [13]. The authors' primary conclusion was that hypericum is as effective as other antidepressants. Observers considered this finding sufficiently important that the review was included in the Top Ten List of Cochrane Reviews [14]. The primary meta-analysis of the review (18 RCTs, 3064 participants) suggests that, compared to placebo, hypericum increases by 17.4% the response rates (defined by reduction in depression in standard measurement instruments) in patients with major depression (RR = 1.48, $CI_{95\%}$ 1.23 to 1.77, DerSimonian-Laird random-effects model). Nevertheless, the analysis revealed substantial heterogeneity (I^2 = 75%), even after subgroup exploration.

Consider this review in relation to another Cochrane review that examines the effect of amitriptyline—a "benchmark" antidepressant—for major depression [15]. The main meta-analysis (31 RCTs, 3228 participants) reveals that amitriptyline increases the likelihood of response comparing to placebo by 22.2% (RR = 1.66, $CI_{95\%}$ 1.51 to 1.81, DerSimonian-Laird random-effects model). In contrast to the hypericum evidence, the effect of amitriptyline was largely consistent across studies (I^2 = 9%).

These results may give the impression that amitriptyline and hypericum have similar effect—indeed, this was the conclusion of the authors of the hypericum study. What, however, is the difference in the inconsistency of results in the two data sets? To clarify the clinical significance of the difference, we present the distribution of the RRs, in superiority/inferiority plots (Fig. 1). These plots are derivatives of the CDF plots of the RR. The x axis represents values of the RR, and the y-axis the associated probabilities. The plots provide the probability that the RR would be smaller than 1 (red line, with 95% CI in brighter red) or above 1 (green line, 95% CI in brighter green). By examining the values on the left side of the plot (left of the line RR = 1), clinicians can deduce the probabilities that the drug is less effective than placebo (i.e., the RR is truly <1). By examining the values on the right side, they can assess the probability that the drug is more effective than the comparator.

As shown by the red curve in Fig. 1A, the probability that hypericum is not superior to placebo is 13.7% [$CI_{95\%}$ of 1.3% to 40.2%]. In contrast, the probability that amitriptyline is superior to placebo (Fig. 1B) is 100% [$CI_{95\%}$ of 97.5% to 100%]. Clinicians can therefore be quite confident that amitriptyline will be effective, whatever the patient population to whom it is offered. The same is not, however, true of hypericum, in which the probability that it is inferior to placebo may be as much as 40% (the upper boundary of the CI around that probability).

The hypericum review provides a further meta-analysis in which reviewers have compared hypericum to traditional antidepressants. The meta-analysis did not reveal a significant difference in the response rates (RR = 0.96, $CI_{95\%}$ 0.88 to 1.05 DerSimonian-Laird random-effects model), which led the authors to conclude that hypericum extracts are similarly effective as other antidepressants. However, the results once again showed substantial inconsistency

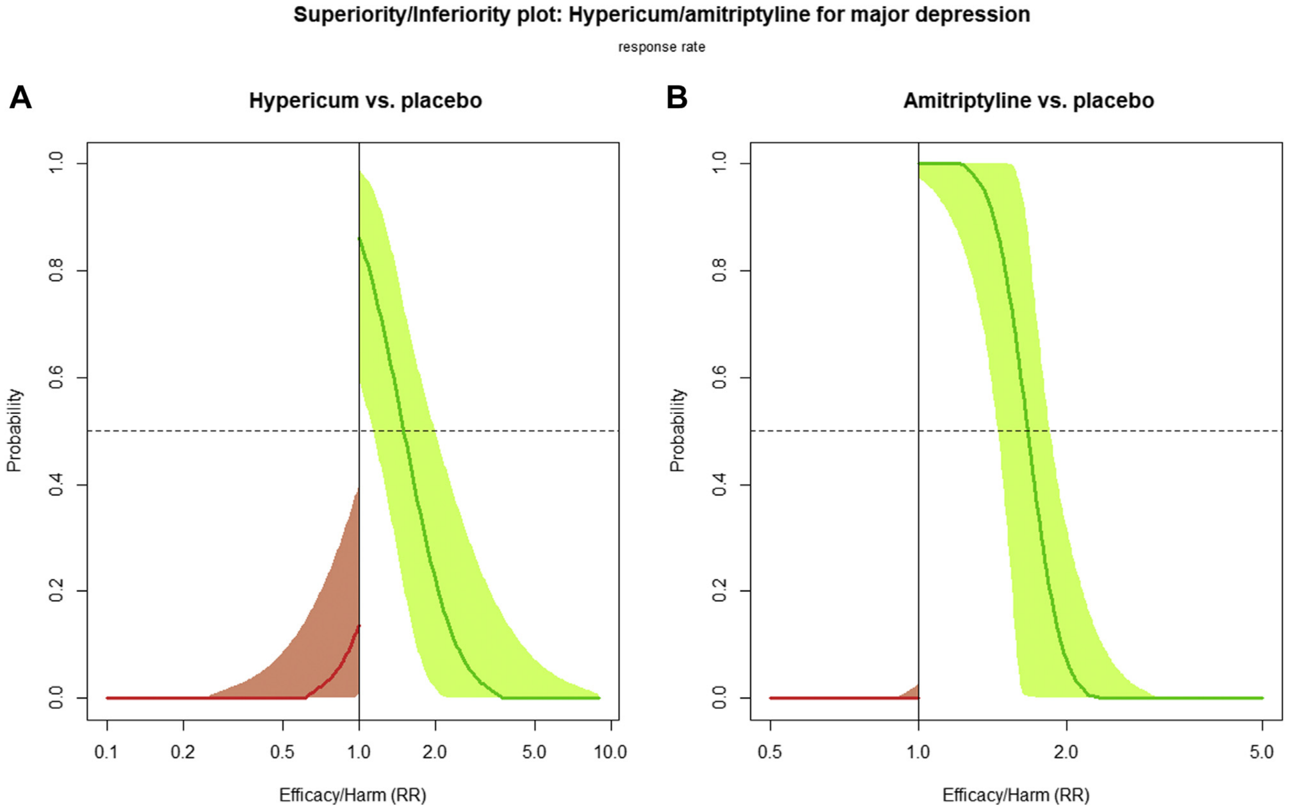

Fig. 1. Superiority/inferiority plot: hypericum/amitriptyline for major depression. Response rate of (A) hypericum vs placebo and (B) amitriptyline vs placebo. RR, relative risk.

(I^2 = 42.5%), raising questions about the confidence one can place in this conclusion.

Fig. 2 reveals that the probability that hypericum is less effective than other antidepressants (RR $<$ 1) is 65.2% [$CI_{95\%}$ of 0 to 100%], and that it is maybe substantially less effective. For example, the upper boundary of the $CI_{95\%}$ suggests that the probability of RR $<$ 0.8 may be as high as 31.4%. There is, however, also a substantial probability that hypericum is more effective than other antidepressants. In any case, the conclusion that hypericum is similarly effective as other antidepressants is over simplistic, and perhaps frankly misleading. The more accurate conclusion is that the effect of hypericum relative to other antidepressants in a new population, or a new patient, is uncertain. Given the proven effectiveness of other antidepressants, risk-averse patients are likely to choose the medication in which they can be more confident.

3.2. A second clinical example

A Cochrane review has examined the impact of adding ribavirin to interferon for patients with chronic hepatitis C [16]. The main meta-analysis (60 RCTs, 9146 patients) showed that compared to placebo, ribavirin significantly reduces the probability of failing to achieve sustained virological response (the RR of failure to achieve such a response was 0.75, $CI_{95\%}$ 0.71 to 0.79, DerSimonian-Laird random-effects model, absolute effect of 24% reduction). The results showed, however, high variability across studies (I^2 = 83%). To address the heterogeneity, the authors explored the possible impact of ribavirin on subgroups that differed according to previous response to viral therapy: patients naïve to antiviral treatment (*naïve*), patients who had relapses after responding to prior therapy (*relapsers*), and patients who did not respond to previous antiviral treatment (*nonresponders*). The authors reported that ribavirin was beneficial in all three groups, but the magnitude of the benefit differed: naïve patients—absolute effect of 24.7%, RR 0.72, $CI_{95\%}$ 0.68 to 0.75, I^2 = 34%; relapsers—absolute effect of 36.6% RR 0.62, $CI_{95\%}$ 0.54 to 0.70, I^2 = 57%; nonresponders—absolute effect of 13.7%, RR 0.89, $CI_{95\%}$ 0.84 to 0.93, I^2 = 63%. A test of interaction demonstrated that adding ribavirin was significantly more beneficial in naive (P $<$ 0.0001) and relapsers (P $<$ 0.0001) compared to nonresponders (16).

Although, the subgroup analysis explains some of the variability, considerable heterogeneity remains in each of the three subgroups as reflected in the I^2 statistics. Given the high rates of adverse effects, the uncertainty in these positive findings may still lead clinicians to wonder whether they should recommend ribavirin. To assist clinical decision-making, we present the distribution of the RRs, in superiority/inferiority plots (Fig. 3).

Fig. 3 shows that the probability that ribavirin is more effective than placebo (RR $<$ 1) varies across the three

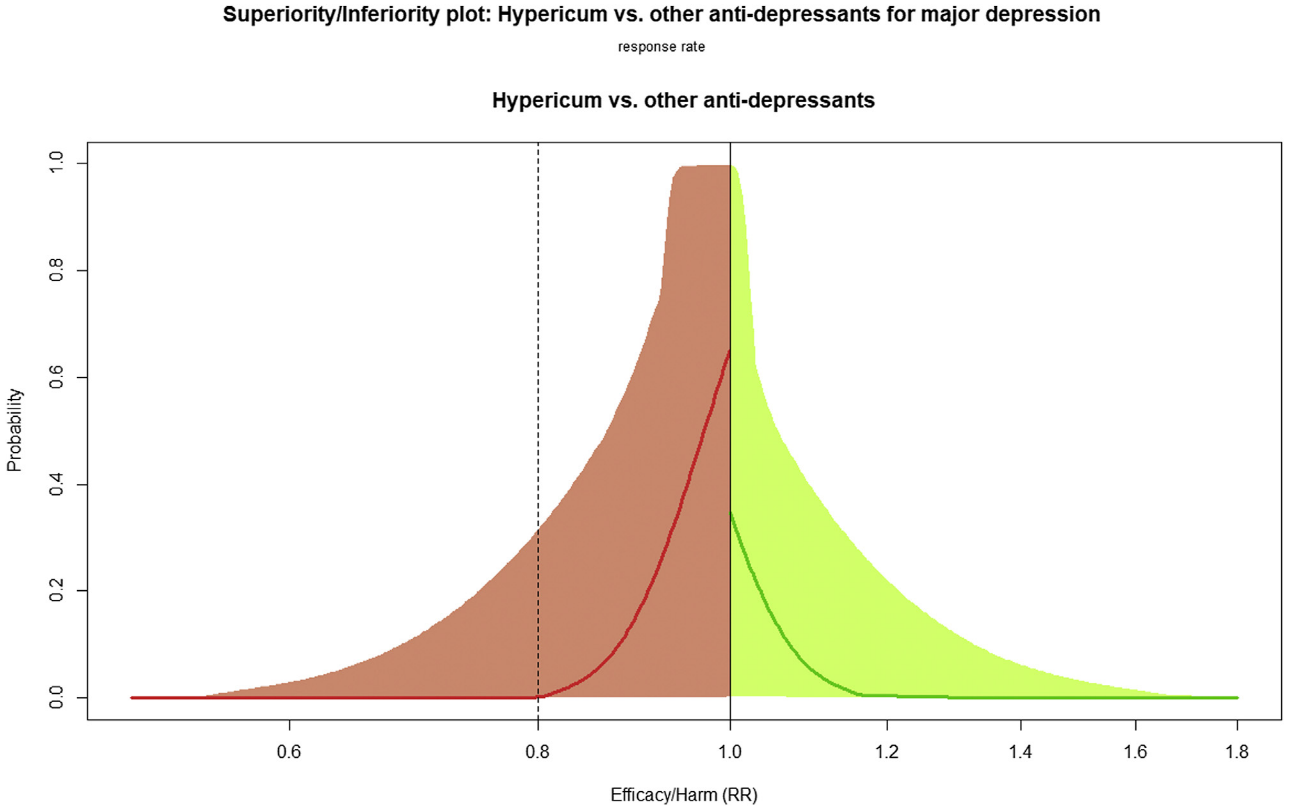

Fig. 2. Superiority/inferiority plot: hypericum vs other antidepressants for major depression. Response rate of hypericum vs other antidepressants. RR, relative risk.

groups. It is virtually certain that ribavirin is effective for naïve patients (probability of 100% with $CI_{95\%}$ of 98.8% to 100%); the effect is less certain for relapsers (probability of 100% with $CI_{95\%}$ of 86.3% to 100%) and even less certain for nonresponders (probability of 94.2% with $CI_{95\%}$ of 73.4% to 100%). Moreover, clinicians can expect a substantial effect in naïve patients, but a substantial effect is less certain for relapsers and unlikely for nonresponders. For example, the probability of $RR < 0.8$ is still 100% for naïve patients ($CI_{95\%}$ of 72.8% to 100%) and 96.6% ($CI_{95\%}$ of 63.3% to 100%) for relapsers, but only 5.3% ($CI_{95\%}$ of 0 to 30.1%) for nonresponders. This assessment helps inform whether, in the face of side effects, fully informed patients will choose or decline ribavirin. One might expect that naïve patients would usually or always choose the drug, but prior nonresponders who place a low value on a small likelihood of benefit are likely to decline.

4. Discussion

Although GRADE specifies that inconsistency reduces the certainty associated with a body of evidence, how clinicians should deal with this uncertainty in the clinical decision-making arena remains open to question. The approach we are suggesting provides additional information beyond traditional presentations of meta-analysis results that can usefully inform clinical decisions in the face of studies with disparate findings.

To optimally estimate intervention effects, researchers need to address both imprecision and inconsistency. The CI around the estimated mean effect addresses imprecision, but in fixed effect models does not address inconsistency at all, and in random-effects models addresses it only partially. Referring to our first example, the comparisons of hypericum and amitriptyline to placebo yield similar CIs using D-L random effects, but the amitriptyline effect is consistent, whereas the hypericum effect is not.

To help clinicians understand the degree of inconsistency, researchers present the I^2 statistic. I^2 has, however, many limitations: when data are sparse, I^2 may be low despite inconsistent results, and when there is a great deal of data, I^2 may be large in the face of unimportant between-study variability. Moreover, the I^2 depends on the mean effect (see the Supplement). Finally, the clinical significance of the I^2 statistics depends on the intervention effect. The I^2 statistics on the ribavirin example suggest moderate inconsistency for both relapsers and nonresponders. The implications of this inconsistency, however, differ markedly for the two groups: it casts serious doubt on use of the drug for nonresponders, but not for relapsers. Our superiority/inferiority plots that yield the distribution

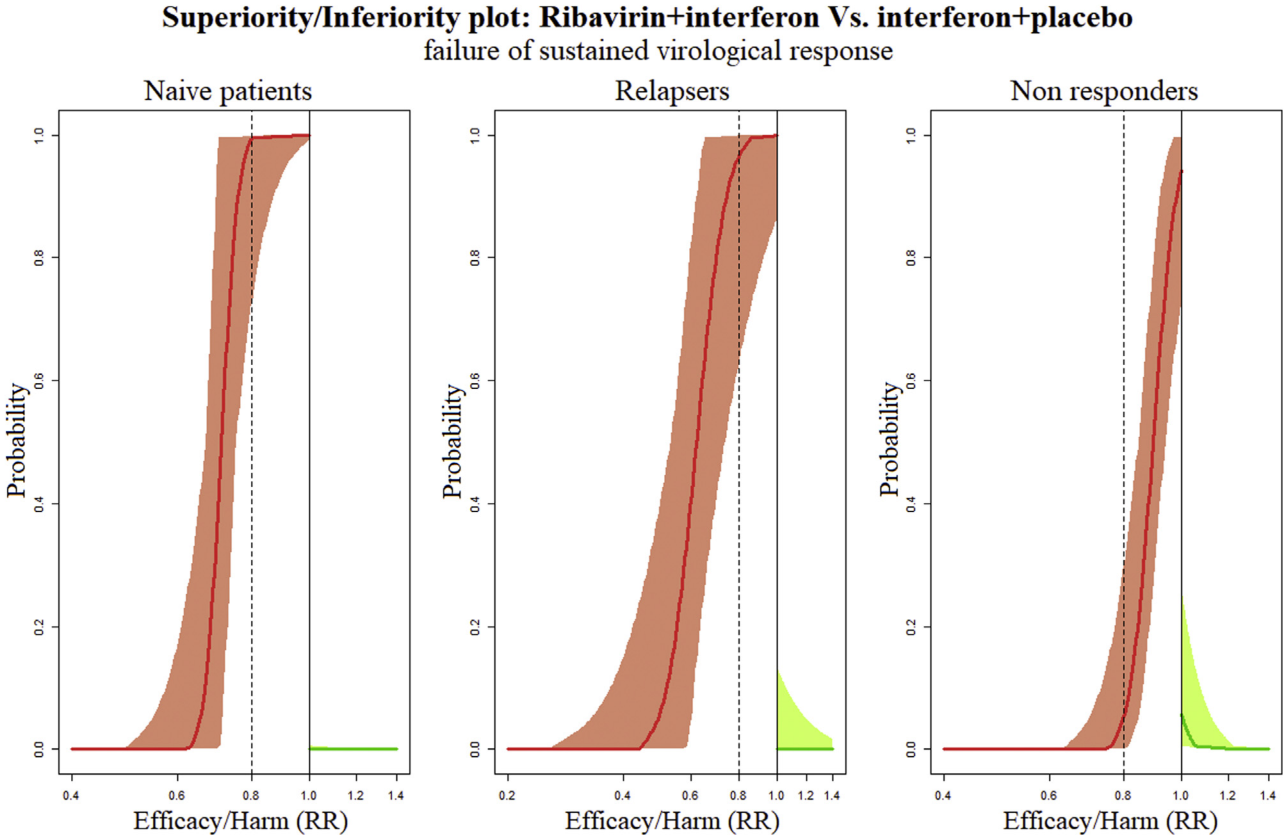

Fig. 3. Superiority/inferiority plot: ribavirin + interferon vs interferon + placebo. Failure of sustained virological response of (A) naive patients, (B) relapsers, and (C) nonresponders. RR, relative risk.

of probable effects in a new population effectively highlight the different implications and provide a quantitative method for assessing the implications of inconsistency for clinical decision-making.

Our inferential framework builds on the standard theory for the random-effects model [17], according to which the intervention effect may differ across studies, and one can view the RR in a new treatment group as independently sampled from the random-effects model. Higgins et al. [18] have already suggested this notion, but while these authors consider a full Bayesian model in which μ and τ are fitted a prior distribution, we use the normal likelihood to construct a confidence region for the RR. Our analysis, hence, is frequentist and does not rely on the Bayesian requirement for establishing priors.

Our approach has limitations. First, it is designed solely for cases in which standard methods for explaining heterogeneity (e.g., control-rate metaregression [19]) fail. Second, our model is based on the assumption that the results of individual RCTs included in meta-analyses are independent and randomly distributed, which is not always the case [18].

In summary, inconsistency across studies is a serious concern in systematic reviews of randomized trials, decreasing certainty in the evidence. Estimating the average effect using random-effects models, by widening CIs, captures some but not all of the additional uncertainty. Analyzing both the standard deviation and the mean effect in random-effects models may enable clinicians to improve their understanding of the distribution of probability of particular intervention effects in a new population. The method thus is a practical and rational method for facilitating optimal clinical decision-making in light of unexplained heterogeneity.

References

[1] Egger M, Smith GD. Meta-analysis. Potentials and promise. BMJ 1997;315:1371–4.

[2] Guyatt GH, Haynes RB, Jaeschke RZ, Cook DJ, Green L, Naylor CD, et al. Users' guides to the medical literature: XXV. Evidence-based medicine: principles for applying the users' guides to patient care. Evidence-based medicine working group. JAMA 2000;284:1290–6.

[3] Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to meta-analysis. Hoboken, NJ: John Wiley & Sons, Ltd; 2009.

[4] Smith GD, Egger M, Phillips AN. Meta-analysis: beyond the grand mean. BMJ 1997;315:1610–4.

[5] Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? Stat Med 2002;21:1559–73.

[6] Sun X, Briel M, Busse JW, You JJ, Akl EA, Mejza F, et al. Credibility of claims of subgroup effects in randomised controlled trials: systematic review. BMJ 2012;344.

[7] Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. BMJ 1994;309:1351–5.

[8] Turner RM, Davey J, Clarke MJ, Thompson SG, Higgins JP. 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.

[9] Higgins JP, Green S. Cochrane handbook for systematic reviews of interventions. Chichester, UK: John Wiley & Sons, Ltd.; 2008.

[10] Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ Br Med J 2003;327(7414):557–60.

[11] Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 7. Rating the quality of evidence - Inconsistency. J Clin Epidemiol 2011;64:1294–302.

[12] Riley RD, Higgins JPT, Deeks JJ. Interpretation of random effects meta-analyses. BMJ 2011;342:d549.

[13] Linde K, Mm B, Kriston L. St John's wort for major depression (review). Cochrane Libr 2008;(4):1–110.

[14] Where's the Evidence? A Top Ten List of Cochrane Reviews | Cochrane Canada [Internet]. Available at http://canada.cochrane.org/news/where's-evidence-top-ten-list-cochrane-reviews. Accessed January 20, 2018.

[15] Leucht C, Huhn M, Leucht S. Amitriptyline versus placebo for major depressive disorder (Review). Cochrane Database Syst Rev 2012;1–149.

[16] Brok J, Gluud LL, Gluud C. Ribavirin plus interferon versus interferon for chronic hepatitis C. Cochrane Database Syst Rev 2010; CD005445.

[17] Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. Stat Med 2001;20:825–40.

[18] 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.

[19] Lau J, Terrin N, Fu R. Expanded guidance on selected quantitative synthesis topics. Methods guide for effectiveness and comparative effectiveness reviews. Rockville, MD: Agency for Healthcare Research and Quality (US); 2008.