Methods Corner

European Journal of Cardiovascular Nursing

ESC European Society of Cardiology

Meta-analysis: How to quantify and explain heterogeneity?

European Journal of Cardiovascular Nursing
2020, Vol. 19(7) 646–652

DOI: 10.1177/1474515120944014
journals.sagepub.com/home/cnu
SAGE

Todd Ruppar

Abstract

The number of systematic reviews and meta-analyses submitted to nursing and allied health journals continues to grow. Well-conducted and reported syntheses of research are valuable to advancing science. One of the common critiques identified in these manuscripts involves how the authors addressed heterogeneity among the studies in their meta-analyses. Methodologically inappropriate approaches regarding heterogeneity introduce error and bias into analyses and may lead to incorrect findings and conclusions. This article will discuss some of the approaches to take as well as avoid when addressing heterogeneity in meta-analyses, including suggestions for how to choose a fixed-effect or random-effects meta-analysis model and steps to follow to address heterogeneity in meta-analysis results.

Keywords

Meta-analysis as topic, research methods, review literature as topic, publication bias, heterogeneity

Date received: 6 April 2020; revised: 26 June 2020; accepted: 29 June 2020

Learning objectives

- Identifying the appropriate meta-analysis model to use for a given research question;
- Understanding how to assess and explain observed heterogeneity properly in meta-analysis results.

The problem: heterogeneity is often poorly addressed in meta-analyses

The number of systematic reviews and meta-analyses submitted to nursing and allied health journals continues to grow. For instance, over the past 2 years six meta-analyses have been published in the *European Journal of Cardiovascular Nursing*.[1–6] Meta-analysis is a valuable tool for quantitatively synthesizing the results of a body of research. Meta-analysis methods permit calculation of an overall mean effect or relationship across a number of studies. Further, moderator analyses (e.g. subgroup analyses, meta-regression) can permit comparative effectiveness analyses to see, for example, if particular intervention types may be more effective than others, or whether interventions may be more effective for particular patient populations.[7]

Unfortunately, many meta-analysis papers submitted to journals contain common and avoidable methodological flaws. In this article, we will discuss the most common of these errors, which center around misunderstandings regarding how to handle variation across observed study effects (statistical heterogeneity) in meta-analytic research.

Stated very simply, in a meta-analysis we calculate an effect size from each study for an outcome of interest, assign weights to those individual study effect sizes using an accepted meta-analysis model, and then calculate a mean overall effect size across all of the included studies. Statistical heterogeneity is the variation of individual study effect sizes.[8] This can be due to differences in study participants, interventions, or outcomes (clinical heterogeneity) as well as variation in study designs or risks of bias (methodological heterogeneity).[9] Heterogeneity is expected in any meta-analysis.[10] There will always be some degree of clinical or methodological heterogeneity. Too much heterogeneity can be concerning, in that it could indicate that the studies are not similar enough to be

Department of Adult Health and Gerontological Nursing, Rush University, USA

Corresponding author:
Todd Ruppar, Department of Adult Health and Gerontological Nursing, College of Nursing, Rush University, 600 S Paulina St, Suite 1080, Chicago, IL 60612, USA.
Email: todd_ruppar@rush.edu
Twitter: @ToddRuppar

quantitatively synthesized. This could be the case, for instance, in a meta-analysis of trials examining the efficacy of a new medication. When synthesizing the results of trials of a single drug, heterogeneity may be considered to be a negative – we could assume that there is a single true mean effect of the medication, because the typical intervention effect would be the same and the included studies would all have nearly identical designs. Thus, clinical and methodological heterogeneity would be assumed to be minimal, and differences between the included studies will likely be small and due to minor sample differences or measurement variation. Significant heterogeneity may indicate excessive sources of error around the effect size estimate (e.g. differences in measurement methods, samples, outcome time points).[11]

In nursing research, we are often studying phenomena or interventions in which the effects are more likely to vary across studies. For example, in studies of interventions to increase health behaviors (e.g. physical activity, medication adherence, heart failure self-care) we tend to expect heterogeneity. Not all interventions to modify a given health behavior are the same, and individual patient responses to the same behavioral intervention will vary greater than they would to a medication due to clinical and methodological heterogeneity. As an example, if a meta-analysis was looking to synthesize the findings of interventions to improve physical activity among patients with heart failure, studies may vary due to many factors, such as intervention content, intervention delivery method, intervention intensity or frequency, patients' heart failure severity, comorbidities, physical limitations, or study outcome measurement method, to name a few. All of these are potential sources of heterogeneity between the included studies. When conducting a meta-analysis, it is necessary to assess and explain heterogeneity appropriately.

Assessing heterogeneity

In a meta-analysis, we use several statistics to assess heterogeneity. Heterogeneity is impacted by several factors: (a) the number of studies in the meta-analysis; (b) how much the study effect sizes vary from each other (between-studies variance); and (c) how much variance exists in the observed effect size for each study (within-study variance).

The Q-statistic is the weighted sum of the squared values of each study effect size's deviation from the mean effect size of all studies in the meta-analysis. The Q-statistic is a standardized measure, and is not affected by the effect size metric used. Q follows a chi-squared distribution, and the value of Q is sometimes reported in summary statistics as chi-squared or χ^2. The Q-statistic can be used to test the null hypothesis that all studies share a common effect size. When testing for heterogeneity using Q, we use the number of studies in the analysis (k) to calculate the degrees of freedom (df = k–1). If Q – df is less than zero, then there is not excess heterogeneity beyond what we would expect if all studies shared the same true effect size. If Q – df >0, then there is excess variation in effect sizes among studies that may be attributable to sources of clinical or methodological heterogeneity. The values of Q and the P value of the Q-statistic are dependent on the number of studies in the analysis.

Meta-analysis software will also report a value for T and T^2. T, the estimate of the population variable tau (τ), is the standard deviation of the overall effect size and T^2 represents the variance of the overall effect size. While the Q-statistic is a standardized variable, T^2 depends on the scale of the effect size metric. T and T^2 are useful for helping to understand how much individual study effect sizes are dispersed about the mean effect size, and interpreting the potential impact of observed heterogeneity.

Finally, many authors use I^2 to describe heterogeneity. The I^2 index is a measure of the proportion of unexplained heterogeneity, calculated as $[(Q - \text{df}) \div Q] \times 100\%$. As a ratio, the value of I^2 is not dependent on the number of studies or the effect size metric, but it is subject to over-interpretation.

Figure 1 is a forest plot using data from a meta-analysis previously published in the *European Journal of Cardiovascular Nursing*.[12] In Figure 1, the plot shows some dispersion of study effect sizes around the mean effect size. This is confirmed with the heterogeneity statistics. The Q-statistic is greater than the degrees of freedom, and the P value for the Q-statistic is less than 0.05. I^2 is 81%, indicating that a large proportion of the existing variance is due to unexplained heterogeneity, but I^2 does not tell us about the actual measured value of that variance (measured by T and T^2) or whether the observed between-study variance is statistically significant (Q and Q – df).

A solution: four steps to address heterogeneity

Heterogeneity in meta-analyses does not necessarily need to be a reason to avoid conducting a meta-analysis, but the methodological quality of the meta-analysis will be improved if heterogeneity is appropriately addressed. This paper proposes four steps researchers can use to deal with heterogeneity in meta-analyses.

Step 1: choose the appropriate meta-analysis model

The choice of meta-analysis model is best made a priori, based on the review's research question, inclusion criteria, and an understanding of clinical and methodological reasons for potential statistical heterogeneity.[8,13] While there is not universal agreement on the best approach for exploring heterogeneity in a meta-analysis, the major meta-analysis

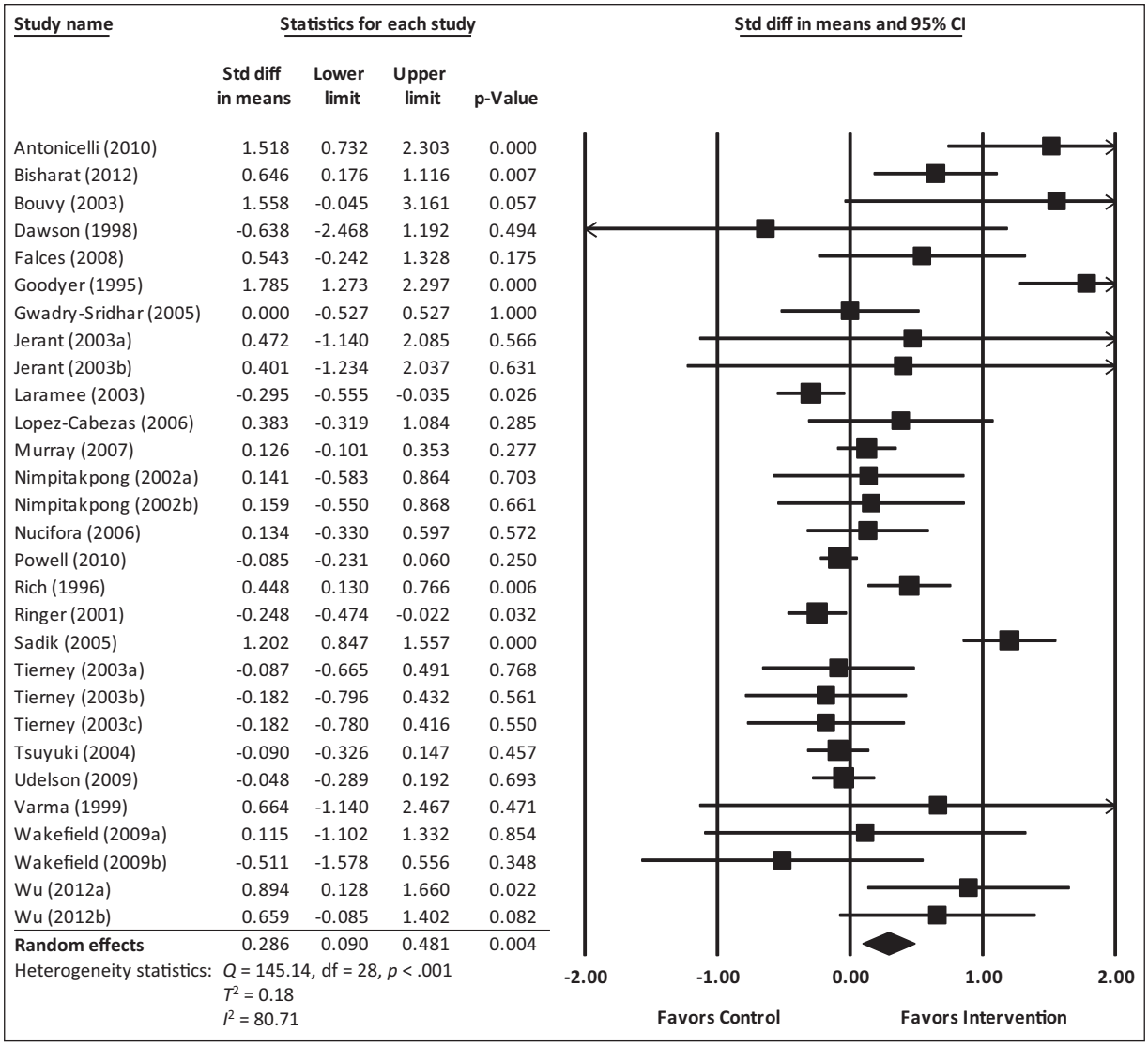

Figure 1. Forest plot of heart failure medication adherence interventions.

expert groups do agree that authors should not choose whether to use a fixed-effect or random-effects meta-analysis model based on a test of heterogeneity.[8,9,13,14] When reviewing meta-analysis papers, I often see authors stating that they used a fixed-effect approach, found significant heterogeneity using the *Q*-statistic, and switched to a random-effects model. Sometimes, rather than the *Q*-statistic, authors will use an arbitrary cut-off of the I^2 index to determine that there is too much heterogeneity for a fixed-effect model. Both of these approaches are considered methodologically inappropriate.[13]

Random-effects model meta-analyses include an estimate of between-study variance when weighting studies in the overall effect size estimate and when calculating the variance of the overall effect size. The inclusion of between-study variance is because random-effects models assume that there is not one single true effect size for the phenomenon being studied. Rather, the studies in a random effects analysis are a sample of effect size estimates from among a distribution of possible true effects. The fixed-effect model does not include this between-studies variance component, and assumes that there is only one single true effect. Any observed variance in a fixed-effect model is assumed to be due to sampling error in selecting the sample of studies in the analysis. Tightly controlled interventions, such as drug trials, typically use a fixed-effect approach, as there is not a conceptual difference in the expected average effect between studies. In many types of nursing research, however, we expect differences. For instance, in a physical activity intervention meta-analysis, we would expect variation in the true effect based on differences in intervention content. None of the studies is likely to be testing the exact same intervention. As a result, between-study heterogeneity is expected, and

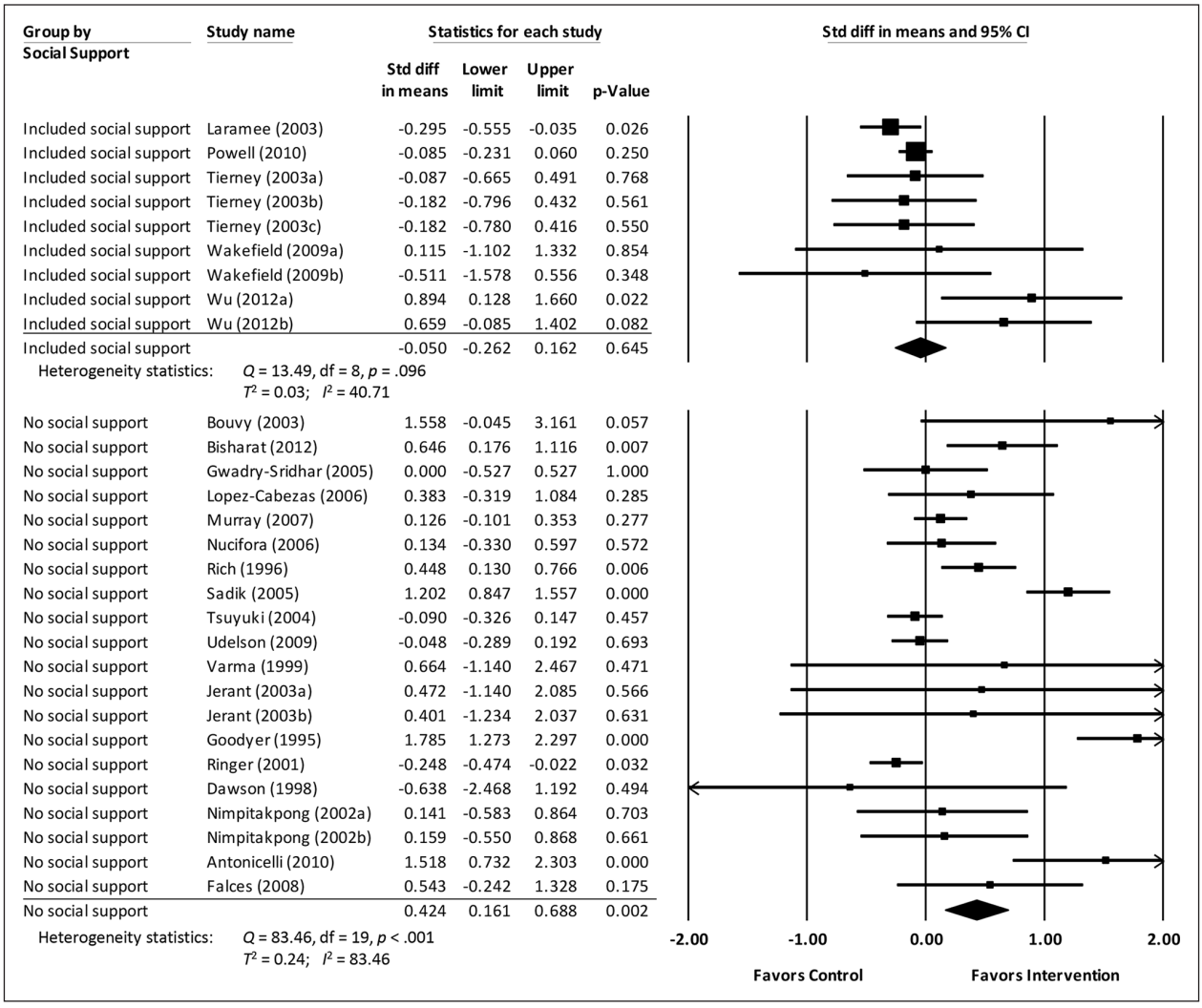

Figure 2. Subgroup analysis.

it would be most appropriate to use a random-effects model in this case. The random-effects approach is more conservative in terms of rejecting the null, because it includes both within-study and between-studies variance when weighting studies and calculating the variance of the overall effect size.

Step 2: assess outliers

Statistical outliers can also be sources of heterogeneity in a meta-analysis. As an initial evaluation, authors may be able to see from their forest plot if one or more studies may be outliers. While some authors will do an analysis to check the effect size and heterogeneity with each study removed one at a time,[15] this may not be sufficient if there are multiple outliers present in a large meta-analysis with many studies. It may be better for authors to explore for potential outliers statistically by checking for significant standardized residuals, and then check those possible outliers both for data extraction or coding errors and also to see if the studies are, in fact, outliers that should be eliminated from the meta-analysis.[15]

Step 3: explore the heterogeneity

If your meta-analysis has a sufficient number of studies, it is possible to conduct moderator analyses using meta-regression and/or subgroup analyses to try to explain some of the potential sources of heterogeneity. For instance, subgroup analyses allow to you compare effect sizes of subgroups of studies in your review and assess whether heterogeneity is lowered when grouping analyses by subgroups. An example of a subgroup analysis is shown in Figure 2, which divided the studies from Figure 1 into whether the intervention included a component designed to enhance social support for medication adherence. Interventions including a social support component had less variation in effect sizes. The *Q*-statistic for the social

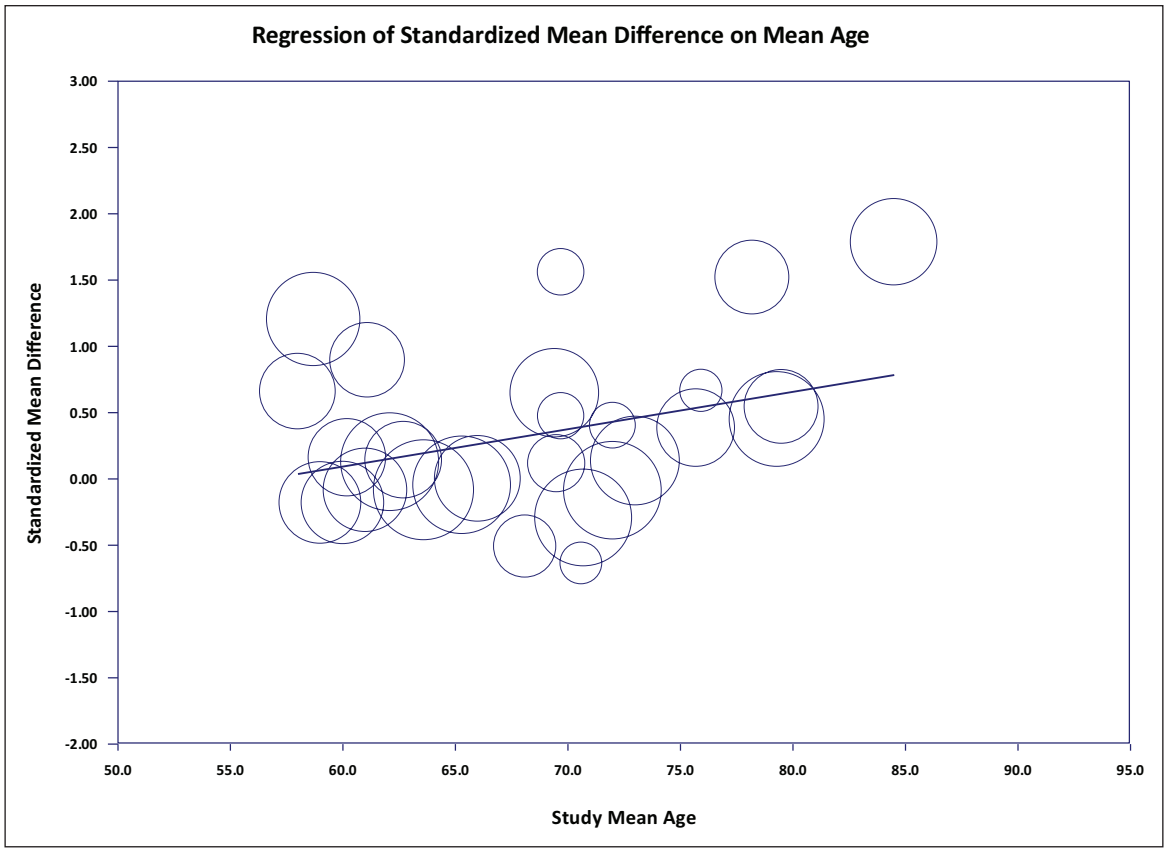

Figure 3. Meta-regression plot.

support studies was 13.49 with df of 8 (P=0.096). Within this subgroup, the I^2 dropped to 40.71%.

Using the subgroup variable as a moderator, it is also possible to analyze whether the groups' effect sizes are significantly different. In the Figure 2 example, interventions without a social support component had a larger standardized mean difference (d=0.42) than did interventions with a social support component (d=–0.05). We would look at the between-studies variance (not shown in the figure) to see that the subgroups' effect sizes are significantly different ($Q_{between}$=7.55, P=0.006).

An example of meta-regression is shown in Figure 3, exploring whether effect sizes vary due to the mean age of the study samples. In this example, the regression coefficient is 0.028, indicating that on average, for every one year increase in mean sample age, the effect size increases by 0.028. As the regression coefficient is statistically significant (P=0.045) it is unlikely that the slope is zero. Looking at heterogeneity statistics, the study dispersion due to the regression model is significant, indicating that the relationship between mean age and intervention effects is stronger than would be expected due to chance (Q_{model}=4.02, df=1, P=0.045).

If too few studies are included in a moderator analysis, however, the analysis will be underpowered and will not produce useful information. It is recommended that at least 10 studies be available for each meta-regression (minimum five studies per group for subgroup analyses).[9] Moderator analyses may require the involvement of a statistician and/or the use of specialized software.

Step 4: acknowledge the limitations

Authors should address in the limitations section of their discussion that it is possible that limitations exist in the searching and screening approach used, leading to eligible studies being inadvertently left out of the review. It is also possible that there is publication bias inherent in the body of literature. While heterogeneity statistics are not designed to detect publication bias, publication bias may contribute to observed heterogeneity. Publication bias is the tendency of studies to be more likely to be published when they report positive and statistically significant results, or have larger effect sizes.[9] Publication bias also encompasses the tendency of systematic reviews and meta-analyses to include only published studies or only studies published in peer-reviewed journals.[8] Publication bias can be assessed visually, by assessing the symmetry of a funnel plot of the included studies (see example in Figure 4). We can also check for publication bias statistically using tests such as Egger's test;[16] however, the utility and accuracy of such tests are dependent on several factors, such as the number

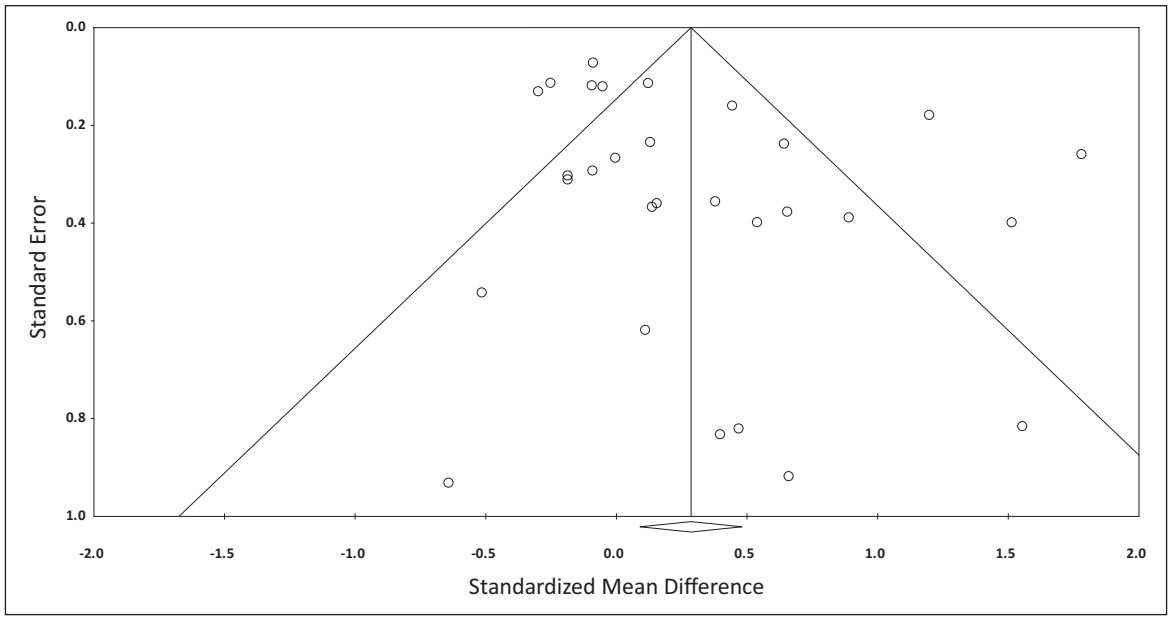

Figure 4. Funnel plot for publication bias.

of included studies, the degree of dispersion across the studies, and the effect size index used in the analysis.[8]

The choice of software used for the meta-analysis may also be a limitation. Some meta-analysis software packages allow for far fewer different types of data to be used for calculating study effect size. This may lead authors to exclude studies on the basis of not having sufficient data for calculating an effect size, when an effect size could be calculated using a different software package. Such exclusions would be methodologically inappropriate.

Software

Most statistical packages that can run a meta-analysis can use either a fixed-effect or random-effects model. Statistically assessing outliers and conducting moderator analyses requires additional coding or macros in SAS, SPSS, STATA, or R. Software specific to meta-analysis, such as RevMan or Comprehensive Meta-Analysis, make these analyses easier. The free or publicly available versions of some software packages have limited functionality and may not be able to conduct all analyses necessary to address heterogeneity. Researchers planning a meta-analysis should review the statistical packages and statistician support available to them prior to conducting their review to see that the necessary resources are available to ensure methodological rigor in their meta-analysis.

Reporting

When writing papers of meta-analysis projects, authors should address each of these four steps for addressing heterogeneity. Explicitly state the meta-analysis model used (fixed-effect or random-effects), with the rationale for the model choice. Describe how outliers were identified and handled. If significant heterogeneity is present, conduct moderator analyses to explore potential sources of heterogeneity, if enough studies are included to permit moderator analyses. Finally, be sure to discuss all potential methodological limitations of the review methods thoroughly and objectively, including potential sources of unexplained heterogeneity.

Conclusion

Maintaining methodological rigor in a meta-analysis is important to ensure valid and correctly interpreted results. Managing heterogeneity is one of the methodological areas typically under-addressed in meta-analyses submitted to nursing and allied health journals, but following the four steps outlined previously will help authors to improve the quality of their meta-analysis research.

Acknowledgements

The author has received funding for consulting services from Becton Dickinson and The Medicines Company (Novartis), unrelated to the work on this manuscript. The data presented in the examples are from research funded by the PhRMA Foundation (PI: Ruppar).

Declaration of conflicting interests

The author(s) declare that there are no conflicts of interest.

Funding

The author(s) received no financial support for the authorship and/or publication of this article.

References

1. Gao J, Chen G, He H, et al. The effect of auricular therapy on blood pressure: a systematic review and meta-analysis. *Eur J Cardiovasc Nurs* 2020; 19: 20–30.
2. Sua YS, Jiang Y, Thompson DR, et al. Effectiveness of mobile phone-based self-management interventions for medication adherence and change in blood pressure in patients with coronary heart disease: a systematic review and meta-analysis. *Eur J Cardiovasc Nurs* 2020; 19: 192–200.
3. Jin K, Khonsari S, Gallagher R, et al. Telehealth interventions for the secondary prevention of coronary heart disease: a systematic review and meta-analysis. *Eur J Cardiovasc Nurs* 2019; 18: 260–271.
4. Parappilly BP, Field TS, Mortenson WB, et al. Effectiveness of interventions involving nurses in secondary stroke prevention: a systematic review and meta-analysis. *Eur J Cardiovasc Nurs* 2018; 17: 728–736.
5. Liu T, Chan AWK, Liu YH, et al. Effects of Tai Chi-based cardiac rehabilitation on aerobic endurance, psychosocial well-being, and cardiovascular risk reduction among patients with coronary heart disease: a systematic review and meta-analysis. *Eur J Cardiovasc Nurs* 2018; 17: 368–383.
6. Driscoll A, Grant MJ, Carroll D, et al. The effect of nurse-to-patient ratios on nurse-sensitive patient outcomes in acute specialist units: a systematic review and meta-analysis. *Eur J Cardiovasc Nurs* 2018; 17: 6–22.
7. Conn VS, Ruppar TM, Phillips LJ, et al. Using meta-analyses for comparative effectiveness research. *Nurs Outlook* 2012; 60: 182–190.
8. Borenstein M, Hedges LV, Higgins JPT, et al. *Introduction to Meta-Analysis*. Chichester: Wiley, 2009.
9. Higgins JPT, Thomas J, Chandler J, et al. (eds). *Cochrane Handbook for Systematic Reviews of Interventions*. 6.0. Cochrane; 2019. www.training.cochrane.org/handbook (accessed 19 July 2020).
10. Higgins JPT. Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. *Int J Epidemiol* 2008; 37: 1158–1160.
11. Cooper HM, Hedges LV and Valentine JC (eds). *The Handbook of Research Synthesis and Meta-Analysis*, 2nd ed. New York: Russell Sage Foundation, 2009.
12. Ruppar TM, Delgado JM and Temple J. Medication adherence interventions for heart failure patients: a meta-analysis. *Eur J Cardiovasc Nurs* 2015; 14: 395–404.
13. Cooper H. *Research Synthesis and Meta-Analysis: A Step-by-Step Approach*, 5th ed. Los Angeles: Sage, 2017.
14. Shadish WR and Haddock CK. Combining estimates of effect size. In: Cooper H, Hedges LV, Valentine JC (eds) *The Handbook of Research Synthesis and Meta-Analysis*, 2nd ed. New York: Russell Sage Foundation, 2009.
15. Hedges LV and Olkin I. *Statistical Methods for Meta-Analysis*. San Diego: Academic Press, 1985.
16. Egger M, Smith GD, Schneider M, et al. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997; 315: 629–634.