REVIEW

A Primer on Systematic Reviews and Meta-Analyses: Part I

Stuart Fisher[1], Melissa J Pearson, PhD, AEP[1], Neil A. Smart, PhD, AEP[1]

ABSTRACT

The conduct of systematic reviews and meta-analyses are a cornerstone source of information required for evidence-based practice in all medical and allied health professions. Meta-analyses are important in the exercise sciences because, for instance, sometimes many small underpowered studies may suggest the optimal treatment deviates from the generic guidelines that suggest 30 minutes to 60 minutes of moderate intensity aerobic activity 3 to 5 times weekly, supplemented by 1 or more sessions of resistance exercise. A systematic review and meta-analysis can help by combining studies to increase power and provide an answer. The signature method of presenting results of meta-analyses is the forest plot, and an ability to interpret these data and the associated funnel plots are essential to the practice of evidence-based exercise programming. This work describes the processes of systematic review and meta-analysis and informs the reader on how these works may be presented, interpreted, and applied. Some examples from the field of kinesiology and exercise physiology are presented to illustrate how the results of a meta-analysis may influence evidence-based practice. *Journal of Clinical Exercise Physiology*. 2021;10(4):160–164.

Keywords: forest plots, evidence-based practice, heterogeneity, data pooling

INTRODUCTION

This is part I of and 2-part series on understanding the general processes of, and how to interpret, a systematic review and meta-analysis paper. The purpose of this paper (Part I) is to provide an introductory explanation of how to conduct a systematic review and meta-analysis and interpret how the results may translate into clinical exercise science practice.

GETTING STARTED

Often results from clinical trials are contradictory, and readers are left wondering which results to believe and which to disregard. One possible clarification approach is to gather the data from all of the relevant studies and add or *pool* the data into 1 analysis. This approach is called meta-analysis. Often readers of journal articles will see the term "systematic review" used in conjunction with "meta-analysis". Both systematic reviews and meta-analyses begin with the development of a research question, and both require a systematic search and review of the literature to find studies that meet the predetermined inclusion criteria used to answer that particular question. Figure 1 illustrates how the data for meta-analysis might come from a search of relevant randomized controlled trials (RCTs). When an RCT is identified, it is then reviewed for predetermined inclusion and exclusion criteria in a systematic process. If the RCT is included, its data will be used in the meta-analysis. This process is presented below.

The inclusion criteria refer to the defined intervention or treatment population (P) (e.g., type 2 diabetes), the intervention or treatment (I) (e.g., aerobic exercise), the comparison group (C) against which the population of interest (P) is compared (e.g., placebo, control, or usual care groups), and the outcome measures (O) (e.g., fasting blood glucose levels). Together these variables are commonly referred to as the *PICO* (Population, Intervention, Comparator, and Outcomes), and in addition to the type of study design, which is often limited to RCTs, the PICO elements collectively form the inclusion criteria.

[1]Department of Clinical Exercise Physiology, University of New England, Armidale, NSW 2350 Australia

Address for correspondence: Neil A. Smart, School of Science and Technology, University of New England, Armidale, NSW 2350, Australia; e-mail: nsmart2@une.edu.au.

Conflicts of Interest and Source of Funding: None

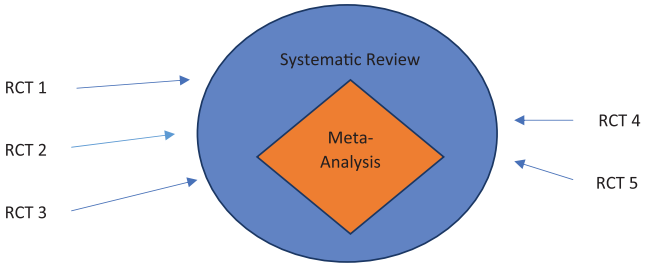


FIGURE 1. Summary of interrelationship between randomized controlled trials (RCTs), systematic review, and meta-analysis.

The exclusion criteria also must be defined. These, like the inclusion criteria, are at the scientist's discretion but must be clearly stated and defined. For example, a study involving type 2 diabetes would exclude type 1 and gestational diabetes and would also exclude animal studies if we only want to include humans. And the age range might exclude studies of people under 18 years if the focus is upon adults)

To provide more perspective to the information presented in Box 1, the background to the question is an ongoing debate as to whether exercise training can increase survival time in people with heart failure. It is known that those with heart failure have a high annual mortality rate (1). Until the 1980s the advice often provided to people with heart failure was for "bed rest" as physical stress was considered a risk (2). Early clinical trials in the 1980s and 1990s (3,4) quickly illustrated that exercise was actually beneficial in terms of improving cardiorespiratory fitness levels (peak Vo_2). Later, the ExTraMATCH (5) study performed a meta-analysis combining many of those early clinical trials that showed people with heart failure and higher peak Vo_2values had longer survival times. Today peak $Vo_2$is often used as a surrogate measure of survival time in people with heart failure, particularly those being evaluated for advanced treatments (i.e., heart transplant or left ventricular assist device) (6).

The following is an illustration of the principles of meta-analysis using selected publications of RCTs of exercise training in cardiac rehabilitation in those with heart failure. When data from all the included studies was pooled, the mean difference in peak $Vo_2$was different in the high-intensity versus low-intensity trained groups. The change in peak $Vo_2$is presented in Figure 2, where the higher exercise intensity is associated with greater changes in peak Vo_2than

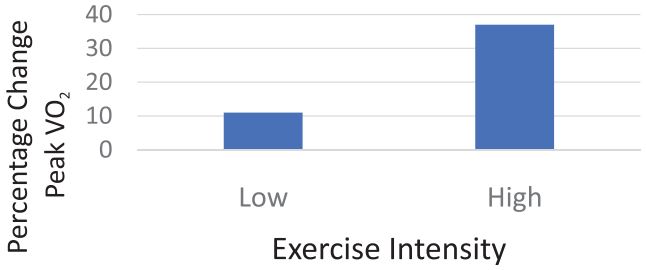


FIGURE 2. Histogram for change in peak Vo_2 high versus low intensity exercise training in people with heart failure.

BOX 1. DEVELOPING A RESEARCH QUESTION USING PICO

The formulation of the research question usually uses the PICO format, for example;

Question:
Does exercise intensity influence the change in cardiorespiratory fitness (Peak Vo_2) in people with heart failure?

PICO:
Population: Heart failure
Intervention: Exercise (categorized in this case by different classifications of intensity)
Comparator: No exercise group receiving only usual care
Outcome(s): Post-training change in peak Vo_2. Although we may wish to measure other outcomes (e.g., mortality and hospitalization), for this question the primary outcome is pre-post training change in peak Vo_2

with lower intensity exercise (7). The meta-analysis, therefore, suggests that those people with heart failure undertaking a cardiac rehabilitation program with an exercise component that uses high-intensity activities are more likely to improve their fitness by a greater amount than with low-intensity exercise training.

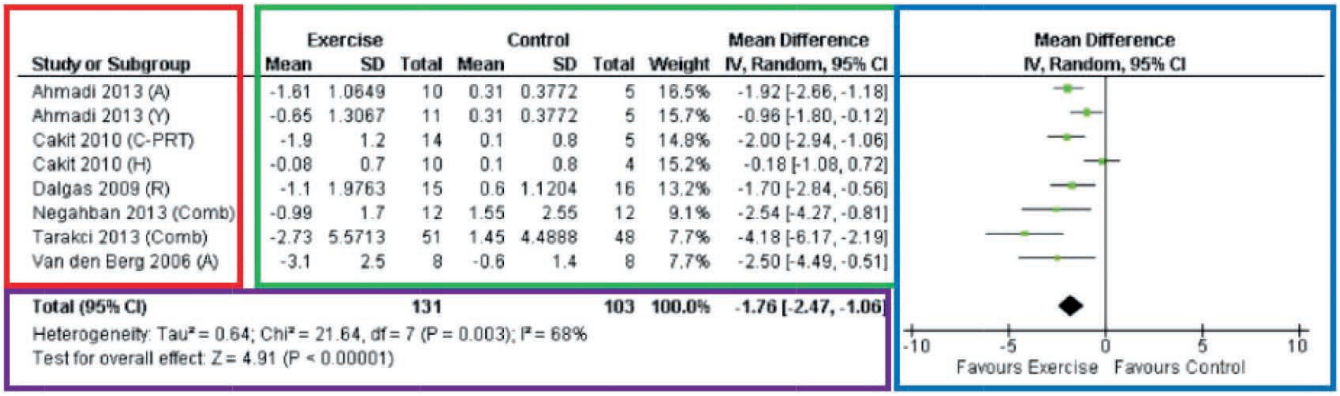
GROUP-LEVEL VERSUS INDIVIDUAL PATIENT-LEVEL META-ANALYSES

Note that meta-analyses are almost always based upon group-level data and not from individual subject-level data. Thus, the mean difference values for data are taken for the intervention group (e.g., high-intensity exercise group) and these are compared to mean values for the control group (e.g., low-intensity exercise group). This type of meta-analysis is most commonly encountered in the published literature, and statisticians often refer to it as a group-level meta-analysis as opposed to an individual patient data meta-analysis that requires included study authors to provide their original datasets with each patient's individual data.

Typically results from a meta-analysis are not often presented in bar graph format. More often results of a group-level meta-analysis are presented in a graph called a forest plot (Figure 3). Figure 4 explains the constituent components of the graphical component of the forest plot.

Figure 5 contains hypothetical data for the postexercise training program change in a blood marker of heart failure severity called brain natriuretic peptide (BNP), which is released when the myocardium is stretched. BNP is usually higher (which is unfavorable) in people who have heart failure versus those who do not. We can see below that BNP is changed by a mean value (difference) of −79.20 $pg \cdot mL^{-1}$ (which is a favorable outcome) after exercise training in people with heart failure because the *P* value is less than 0.05 (5% level of significance) that is traditionally used. The

REVIEW



1. Included Studies
2. Effect Estimate Information
3. Overall Statistics
4. Forest Plot – graphical representation of the meta-analysis

FIGURE 3. A traditional forest plot with constituent components color-coded. Change in 10m WT (s) after exercise in pwMS. A, aerobic; Comb, combined aerobic and resistance training; C-PRT, cycling and progressive resistance training; H, home exercise; IV, inverse variance; R, resistance training; Y, yoga.

following paragraphs examine in detail the aspects of forest plot.

It is important to note that historically forest plots were designed to illustrate whether outcomes such as mortality and hospitalization were lower (therefore more favorable) in treated versus untreated patients. Most statistical software programs today use a default plot where lower values are considered better, therefore it is important to remember in many fields including kinesiology or exercise science that authors prefer that many outcomes have a higher value (e.g., peak Vo_2 is better if higher). Note in this example though BNP is better if lower. The way authors most commonly adjust for having higher values as better by simply swapping the axes at the foot of the forest plot (favors exercise – favors control) when an outcome is better if higher. In Figure 3 the default (i.e., lower is better) occurs, and thus in this case the lower BNP values are better and there is no need to swap the "favors exercise favors control" axes, as would be done for a peak Vo_2outcome measure because higher peak $Vo_2$is better.

When examining the forest plot in Figure 5 the far right column denotes the statistical weighting (%) assigned to each study (e.g., Barnes is 13.82% in Figure 5). The study by Parkes has the highest weighting and thus influences the outcome of the meta-analysis the most of the 5 studies listed. The number of participants is 1, but not the only factor that determines weighting. Weighting is also partially related to study variability. So the larger the SD reported in the study, the lower the weighting that will be assigned. Although the effect of variance is not obvious in our figure, we can see the effect of the reported SDs has some bearing on weighting because the hypothetical study by Barnes 2012 has 41 (23 exercise + 18 control) participants, yet it is weighted slightly lower (13.82% vs. 14.35%) than the study of Jones 2006 that only has 40 (21 exercise + 19 control) participants. The

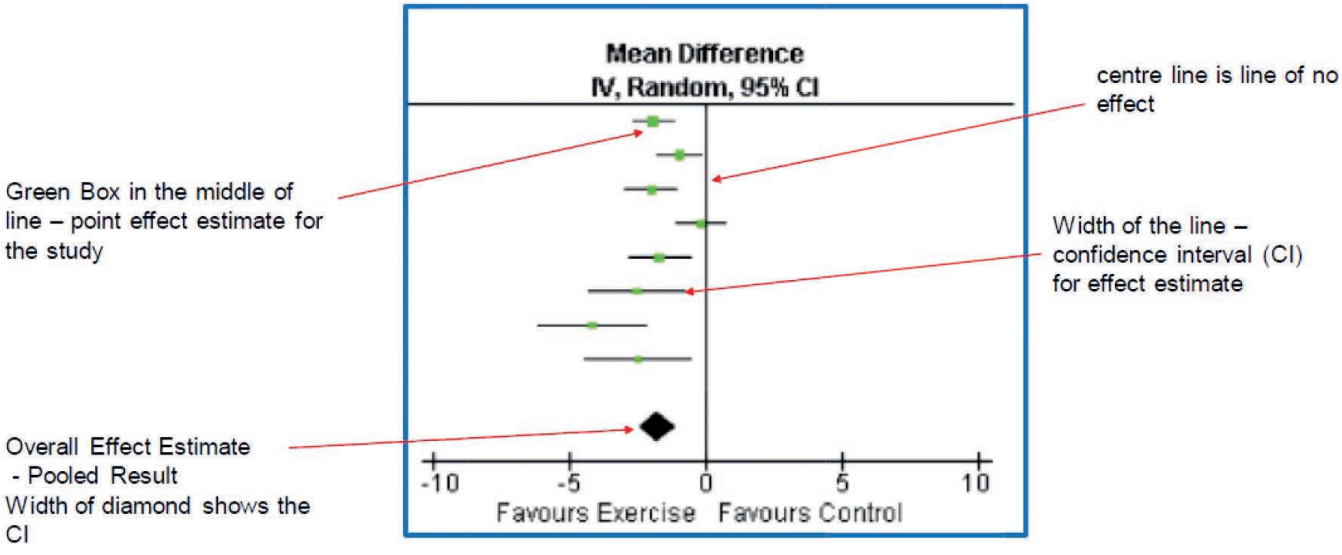


FIGURE 4. Constituent components of the graphical component of the forest plot.

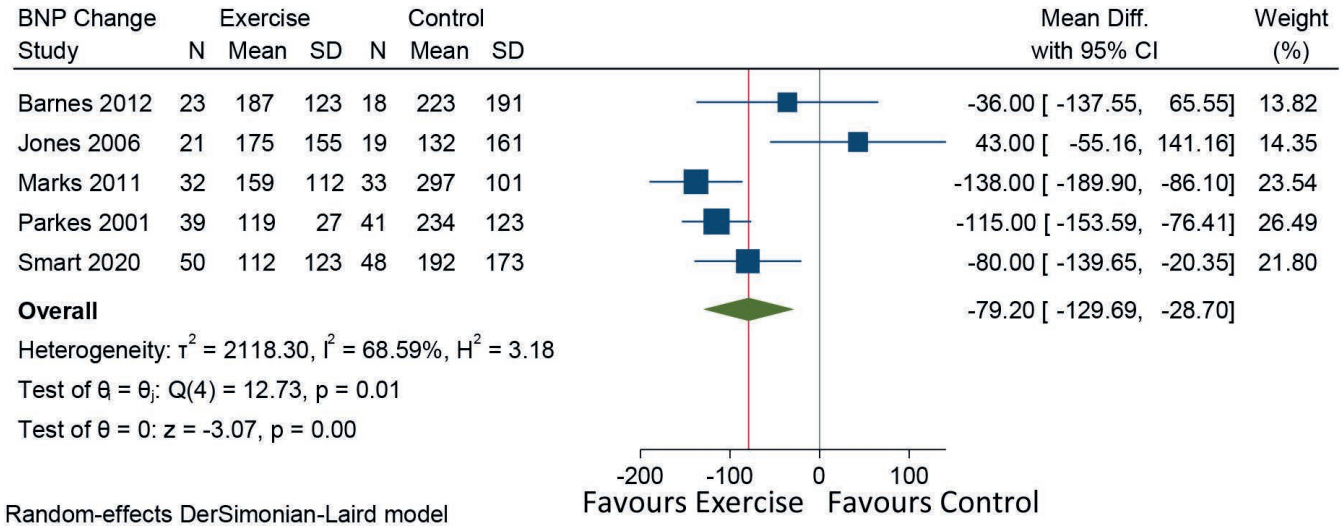


FIGURE 5. Random effects forest plot of postexercise training change in brain natriuretic peptide in people with heart failure.

lower weighting of Barnes is partly because of the higher SD of 191 for the control group.

The next step in the analysis is to assess the mean difference values. These are the differences between mean values for the exercise and control groups. In the forest plot example for Barnes, the mean difference is calculated as 187 − 223 = −36. This is done for all included studies. Note in the example in Figure 5 that only the Jones study had a mean difference in the opposite (higher BNP in the exercise group) resulting in a difference in mean values of 43.

Heterogeneity is the variability in outcomes beyond what is expected due to measurement error. In this case the heterogeneity, expressed as I^2% in the bottom left of Figure 5, is relatively high at 68.6%. This means that overall, the studies are inherently different from one another. Although subjective, some authors will not pool data (the process of conducting meta-analyses and developing forest plots) if heterogeneity is too high, with 75% often considered the threshold. In this example the *P* value for the test of heterogeneity (not the effect size for the outcome measure) is $P = 0.01$, which means there is statistically significant heterogeneity between studies. This heterogeneity probably stems from the large difference in SD values (range 27 to 191) across the 5 studies.

Recall that the significance test result for the outcome measure is $P < 0.01$. Without knowing the *P* value, the reader can look at the forest plot and see that the effect size or point estimate for the mean difference is statistically significant because the horizontal component of the green diamond (corresponding to the 95% confidence interval of the meta-analysis) does not touch or cross the line of no effect (i.e., vertical line going through value zero). If the green diamond crossed the line of no effect, then this would indicate the effect is not statistically significant and the *P* value would be greater than 0.05 in this case.

At the bottom of the forest plot there is a notation that a random effects model has been used, in this case specifically the Der Simonian-Laird random effects model. There are

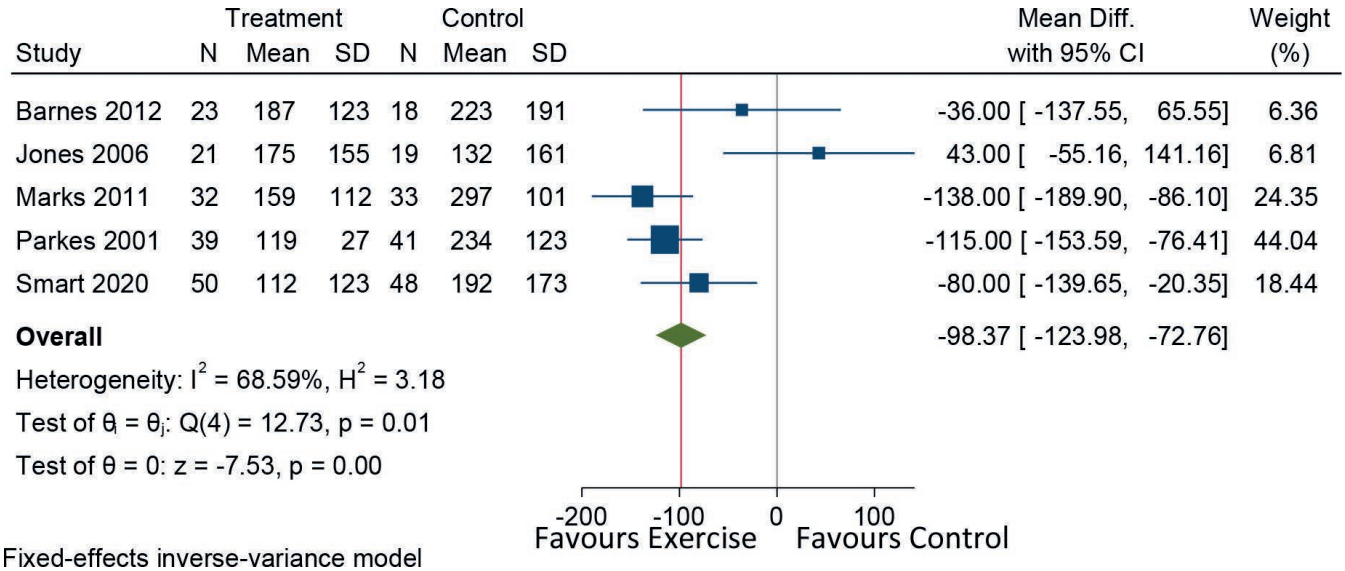


FIGURE 6. Fixed effects model forest plot of postexercise training change in brain natriuretic peptide in people with heart failure.

many different statistical tests that can be selected and typically available as a default setting of common software packages. Each are valid and have various specific attributes. Readers who wish to learn more can obtain information from the following reference (8). The Der Simonian-Laird model is most used by Cochrane systematic review authors

FIXED VERSUS RANDOM EFFECTS

Most statistical software offers a choice of using either a random or fixed effects model when generating forest plots. There are many opinions on the choice of model, but generally it is agreed that a random effects model (Figure 5) is most conservative. So, a fixed effects model (Figure 6) is often avoided in meta-analysis as it is considered less conservative than a random effects model and therefore has a greater chance of achieving statistical significance and thus an increased risk of a type 1 error (i.e., incorrect rejection of null hypothesis). This is most simply explained because the 95% confidence interval is narrow and therefore less likely to cross the black vertical line of no effect.

SUMMARY

Meta-analysis is a key tool medical and health practitioners use to clarify whether a treatment or an approach to delivering a treatment is effective or not in the presence of conflicting data from different publications. In Part II the primer will address the issue of publication bias using funnel plots to identify if nonsignificant studies exist that have not been published, because they produced "negative findings". In addition, sub-analyses and meta-regression will be examined as two of the most used techniques used to identify if particular study characteristics (e.g., intervention type or volume) lead to more favorable or worse changes in the clinical outcomes of interest. Finally, in Part II we will introduce readers to the more sophisticated approach to meta-analysis that uses individual patient data and not group-level "average" data, as Part I was limited to the latter.

REFERENCES

1. Virani SS, Alonso A, Benjamin EJ, Bittencourt MS, Callaway CW, Carson AP, Chamberlain AM, Chang AR, Cheng S, Delling FN, Djousse L, Elkind MSV, Ferguson JF, Fornage M, Khan SS, Kissela BM, Knutson KL, Kwan TW, Lackland DT, Lewis TT, Lichtman JH, Longenecker CT, Loop MS, Lutsey PL, Martin SS, Matsushita K, Moran AE, Mussolino ME, Perak AM, Rosamond WD, Roth GA, Sampson UKA, Satou GM, Schroeder EB, Shah SH, Shay CM, Spartano NL, Stokes A, Tirschwell DL, VanWagner LB, Tsao CW; American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. Heart disease and stroke statistics-2020 update: a report from the American Heart Association. Circulation. 2020 Mar 3;141(9):e139–e596
2. Thaulow E, Ihlen H, Kjekshus J, Forfang K, Storstein O. Effects of red rest and prazosin in congestive heart failure. Acta Med Scand. 1982;212(3):131–6
3. Sullivan MJ, Higginbotham MB, Cobb FR. Exercise training in patients with chronic heart failure delays ventilatory anaerobic threshold and improves submaximal exercise performance. Circulation. 1989;79(2):324–9
4. Coats AJ, Adamopoulos S, Meyer TE, Conway J, Sleight P. Effects of physical training in chronic heart failure. Lancet. 1990;335(8681):63–6
5. Piepoli MF, Davos C, Francis DP, Coats AJ. Exercise training meta-analysis of trials in patients with chronic heart failure (ExTraMATCH). BrMed J. 2004;328(7433):189. doi:10.1136/bmj.37938.645220.EE.
6. Sarullo FM, Fazio G, Brusca I, Fasullo S, Paterna S, Licata P, Novo G, Novo S, Di Pasquale P. Cardiopulmonary exercise testing in patients with chronic heart failure: prognostic comparison from peak VO2 and VE/VCO2 slope. Open Cardiovasc Med J. 2010 May 26;4:127–34
7. Ismail H, McFarlane JR, Nojoumian AH, Dieberg G, Smart NA. Clinical outcomes and cardiovascular responses to different exercise training intensities in patients with heart failure: a systematic review and meta-analysis. JACC Heart Fail. 2013;1(6):514–22
8. Fernandez-Castilla B, Jamshidi L, Declercq L, Beretvas SN, Onghena P, Van den Noortgate W. The application of meta-analytic (multi-level) models with multiple random effects: a systematic review. Behav Res Methods. 2020;52(5):2031–52