HEIDI ISRAEL, FNP, PhD[1] • RANDY R. RICHTER, PT, PhD[2]

A Guide to Understanding Meta-Analysis

Meta-analysis is a popular and frequently used statistical technique used to combine data from several studies and reexamine the effectiveness of treatment interventions. As the number of articles using meta-analysis increases, understanding of the benefits and drawbacks of the technique is essential.

Well-conducted systematic reviews of randomized controlled trials are regarded as representing a high level of evidence.[28] Practicing in an evidence-based manner is a recognized goal for the profession.[3] Systematic reviews are used to answer questions[7,10,40] about the evidence supporting or refuting the effectiveness or efficacy of an intervention. When certain conditions are met, a systematic review may be extended to include a meta-analysis, a statistical procedure used to numerically summarize the included studies' treatment effect.[56] A meta-analysis provides a single, overall measure of the treatment effect, enhancing the clinical interpretation of findings across several studies. Because of its increasing use, high level of evidence, and enhanced clinical interpretation of treatment effects, interpreting a meta-analysis is an important skill for physical therapists. The purpose of this commentary is to expand on existing articles describing meta-analysis interpretation,[6,13,14,42,61] discuss differences in the results of a meta-analysis based on the treatment questions, explore special cases in the use of meta-analysis, and provide physical therapists guidance in interpreting a meta-analysis.

SYNOPSIS: With the focus on evidence-based practice in healthcare, a well-conducted systematic review that includes a meta-analysis where indicated represents a high level of evidence for treatment effectiveness. The purpose of this commentary is to assist clinicians in understanding meta-analysis as a statistical tool using both published articles and explanations of components of the technique. We describe what meta-analysis is, what heterogeneity is, and how it affects meta-analysis, effect size, the modeling techniques of meta-analysis, and strengths and weaknesses of meta-analysis. Common components like forest plot interpretation, software that may be used, special cases for meta-analysis, such as subgroup analysis, individual patient data, and meta-regression, and a discussion of criticisms, are included. *J Orthop Sports Phys Ther 2011;41(7):496-504. doi:10.2519/jospt.2011.3333*

KEY WORDS: *forest plot, literature review, statistical analysis, systematic review*

WHY META-ANALYSIS

A number of reasons exist for considering the use of meta-analysis techniques. Meta-analyses enable one to combine data and summarize the findings of several clinical trials that evaluate the effectiveness or efficacy of a similar treatment approach on similar group of patients. This technique can prove especially useful when there are several similar clinical trials with or without consistent outcomes, or when there are smaller to medium-sized trials with inconclusive results.

By combining the results from 2 or more studies, a meta-analysis can increase statistical power[18] and provide a single numerical value of the overall treatment effect. The meta-analysis result may show either a benefit or lack of benefit of a treatment approach that will be indicated by the effect size, which is the term used to describe the treatment effect of an intervention. Treatment effect is the gain (or loss) seen in the experimental group relative to the control group. The overall positive or negative change may be hard to discern from individual studies. For example, Clare et al[15] used a meta-analysis to examine the treatment effect of McKenzie therapy for spinal pain. Three studies supported the use of McKenzie therapy for short-term pain. Two of the 3 studies reported a small but similar reduction in pain, which was statistically significant for only 1 of the 2 studies. The third study reported a reduction of pain that was twice the magnitude of the other studies. The results of the meta-analysis indicated an overall treatment effect that was statistically significant and closer

[1]Assistant Professor, Saint Louis University, Department of Orthopaedic Surgery, St Louis, MO. [2]Associate Professor, Saint Louis University, Program in Physical Therapy, St Louis, MO. Address correspondence to Dr Heidi Israel, Saint Louis University, Department of Orthopaedic Surgery, 3635 Vista FDT7N, St Louis, MO 63104. E-mail: Israelha@slu.edu

in magnitude to the 2 studies reporting a small reduction in pain. This example illustrates the potential value of meta-analysis when direction of the treatment effect is the same across all studies but the magnitude and statistical significance of the treatment effect varies. A meta-analysis can also be used to show changes in the treatment effect that occur over time. For example, Zhang et al,[60] in an update of the management of hip and knee osteoarthritis (2006 to 2009), determined that the treatment effect sizes for exercise and acupuncture did not change at multiple time points, while the treatment effect sizes for weight reduction eventually reached statistical significance at later time points, and the treatment effect size for electromagnetic therapy was no longer significant at later time points. These type of results obtained from meta-analysis can be used to make better informed treatment decisions.

To interpret a meta-analysis, the reader needs to understand several concepts, including effect size, heterogeneity, the model used to conduct the meta-analysis, and the forest plot, a graphical representation of the meta-analysis. These concepts are discussed below and summarized in **TABLE 1**.

TABLE 1 Selected Concepts in Meta-Analysis

Concept	Definition
Meta-analysis	Statistical analysis that integrates results from 2 or more studies,[27] providing a single numerical value of the overall treatment effect for that group of studies.
Effect size	A dimensionless estimate (ie, a measure with no units) that indicates both direction and magnitude of the treatment effect.[16,36]
Odds ratio	The odds the ratio of the probability of an event occurring compared to the event not occurring in a particular group. The odds ratio is the ratio of the odds between 2 groups.[9]
Relative risk	Relative risk is equal to the risk among exposed subjects divided by the risk among unexposed subjects.[49]
Fixed-effects model	A model that assumes that each study included in the meta-analysis is estimating the same population treatment effect, which, in theory, represents the true population treatment effect.[21]
Random-effects model	A model that assumes that the treatment effects of the included studies are part of a distribution of treatment effects that fall along a range of values.[21]
Forest plot	A graph that visually shows the results from the individual studies (treatment effect and confidence interval), as well as the estimate of overall treatment effect and associated confidence interval.[34]
Confidence interval	Confidence intervals (CIs) provide upper and lower limits that capture the range of values around the true but unknown population value. The 95% CI is most commonly used and corresponds with the typical 5% significance level used in hypothesis tests. CIs of continuous measures that include 0 represent nonsignificant results. CIs of odds ratios and relative risk that include 1.0 represent nonsignificant results.[23]

EFFECT SIZE

Studies included in a meta-analysis must have common outcome statistics that allow their results to be combined.[31] Effect sizes, which reflect the magnitude and direction of the treatment effect for each study, serve this purpose. When all the studies to be included in a meta-analysis have the same outcome measure, an effect size in the original units may be calculated. For example, if all studies in the meta-analysis measure a continuous outcome, such as range of motion, the mean difference can be used as the effect size. Standardization of the effect size is needed when treatments are not measured in the same units. Standardization makes data unitless. When the effect sizes are in the original units, the interpretation is clearer. When the effect sizes are in standardized units, the interpretation is more difficult and published guidelines for interpreting effect sizes may be used.[17] Whether standardized or not, the overall effect size derived from the meta-analysis is calculated by combining the effect sizes of the included studies. There are several types of effect sizes. For dichotomous data, such as improved or not improved, odds ratios or relative risks are used for effect sizes. Other types of effect sizes are often reported in meta-analysis, and these are described in **TABLE 2**.

Because several factors, such as sample size, variance, and reliability of the outcome measures, can influence the magnitude and direction of the effect size, the estimates of the effect sizes will vary among studies. In addition, the effect size of the individual studies may be somewhat imprecise and, therefore, lead to an unstable finding when multiple small studies are utilized. Weighting of the standard error based on sample size allows for the best precision of the effect size estimates.[36] Finally, variables such as gender, age differences, or differences in the intervention provided, such as dose, can influence the magnitude and direction of the effect size.

Use of the confidence interval can lend insight into the precision of the treatment estimates of the included studies. A wider confidence interval may be a function of a small sample size, as well as imprecision in the measurement. Larger sample sizes provide more precise estimates of the effect size,[36] whereas smaller studies are less precise, unless these smaller studies have little variance. Confidence intervals, which are reported as a probability (eg, 95% confidence interval), provide a range (upper and lower bounds) that indicate the precision of the estimate of the effect size.[23,48] If the confidence interval of the effect size falls within an area considered as clinically meaningful, then applications of the results in clinical care may be justified.[39,51] Conversely, wide confidence intervals indicate less precise estimates and, coupled with a small sample size, can lead to questions about the stability

TABLE 2 Typical Effect Size Calculations With Various Data Types Used in Meta-Analysis

Type of Data	Type of Effect Size Reported	Examples of Outcome Measures
Continuous data	Standardized mean difference (eg, Cohen *d*)	Various measures of strength[16]
	Unstandardized mean difference	100-mm visual analog pain scale[8,41]
	Pretest-posttest difference	Depression preintervention/postintervention[19]
Binary (dichotomous) data	Odds ratio	Risk factors for persistent problems following whiplash[58]
	Relative risk	Rates of rerupture in Achilles tendon repair[32]

of the effect size estimates. By combining the results of small studies, a meta-analysis may provide a more precise estimate of the treatment effect.

For example, Khan et al[32] examined randomized trials of operative versus nonoperative treatment of Achilles tendon rupture and calculated effect sizes for rerupture, infection rates, and other complications. One aspect of this meta-analysis examined complications other than rerupture for postoperative management that included splinting with casting alone versus casting followed by functional bracing. The confidence intervals of estimates of the relative risks in the individual studies were quite wide, especially in trials with a smaller sample size; but the meta-analysis effect size favored functional bracing with a smaller confidence interval than that of the individual studies. Mollon et al[38] make the point that even when many studies are excluded from a meta-analysis, based on the stringent inclusion criteria of the systematic review protocol, the confidence intervals surrounding the overall estimated effect size is larger when the small studies are included.

FOREST PLOTS

One of the most useful tools used in meta-analysis is the forest plot, which provides a visual summary of the analysis and findings. A forest plot graphically represents estimates of the effect size and corresponding confidence intervals for each study, along with an estimate of overall effect size of all included studies and the corresponding overall confidence interval.[34] Even when a systematic review does not include a meta-analysis, a forest plot can be used to compare the effect size of the included studies.[44]

In addition to illustrating the effect sizes and related confidence intervals of individual studies, a forest plot can illustrate the extent to which the results from individual studies vary. Variability in results among studies on the same topic is called heterogeneity. When the magnitude and direction of the effect sizes among the studies are similar, heterogeneity is less likely and meta-analysis may be appropriate. Conversely, when study results vary, heterogeneity is possible and a meta-analysis may not be appropriate.

A forest plot of a meta-analysis typically includes the numerical value of the treatment effect and variability for each individual study, the modeling technique assumed (random or fixed), the "line of no effect," a test and corresponding value for heterogeneity, and the numerical estimate of overall treatment effect (**FIGURES 1-2**). The forest plot, therefore, provides a quick visual assessment of the individual studies included in the meta-analysis, a visual assessment of heterogeneity, and the overall treatment effect of the individual studies included.

The clinical context or clinical significance of the findings must be considered when interpreting effect sizes. Some researchers use the term "minimal clinically important difference" (MCID) to indicate clinical versus statistical significance. Because statistical significance does not always translate into clinical significance, the confidence intervals of an effect size can be used to interpret the results of a meta-analysis in a clinically meaningful manner.[39,51] For example, if a 10° change in range of motion is considered clinically meaningful and the lower bound of the 95% confidence interval is 12° and the upper bound is 18°, the statistically significant difference is also clinically meaningful, as the confidence interval exceeds the clinically meaningful value of 10°. This type of finding based on randomized controlled trials should prompt the adoption of the intervention that lead to those gains. However, if the MCID falls within the lower and upper bounds of the confidence interval, the clinician will have to determine if adoption of the intervention is warranted or if additional evidence is needed.[51]

HETEROGENEITY

Heterogeneity is a term used to describe variability among studies,[21] and both statistical and clinical heterogeneity need to be considered.[54] Statistical heterogeneity occurs when the treatment effect estimates of a set of studies vary among one another.[21] Because some variation in treatment effect among studies would be expected by chance, statistical heterogeneity refers to the amount of variation in treatment effect present beyond chance.[21] By convention, statistical heterogeneity is referred to as just heterogeneity.[21] Studies with methodological flaws and small studies may overestimate treatment effects[45,59] and can contribute to statistical heterogeneity. Statistical heterogeneity can be examined and quantified using statistical tests.

Review: Laser knee osteoarthritis

Comparison: 05 Pain outcome RCTs, knee osteoarthritis

Outcome: 06 Pain relief on 100-mm VAS, end of therapy with optimal dose

Study/Subcategory	n	Treatment (Mean ± SD)	n	Control (Mean ± SD)	WMD (Fixed) 95% CI*	Weight, %†	WMD (Fixed) 95% CI
01 Electroacupuncture							
Yurtkuran	25	25.00 ± 12.00	25	5.00 ± 8.90		22.83	20.00 (14.14, 25.86)
Sangdee	48	48.00 ± 24.30	47	23.00 ± 24.20		8.23	25.00 (15.25, 34.75)
Vas	48	53.50 ± 21.40	49	28.50 ± 35.20		5.85	25.00 (13.43, 36.57)
Subtotal (95% CI)	121		121			36.92	21.91 (17.30, 26.51)
Test for heterogeneity: $\chi^2 = 1.07$, $df = 2$ ($P = .59$), $I^2 = 0\%$§							
Test for overall effect: $z = 9.32$ ($P<.00001$)							

‡

Line of no effect

FIGURE 1. Forest plot suggesting little heterogeneity. Abbreviations: CI, confidence interval; RCT, randomized controlled trial; VAS, visual analog scale; WMD, weighted mean difference. *A fixed-effects model, which assumes a common, fixed treatment effect. Between-study differences are assumed due to chance and not incorporated into the model. †Larger studies have a greater influence (weight) than smaller studies. ‡A qualitative visual analysis of the studies' results suggests little between-study variability. The individual study point estimates of the treatment effect (blue squares) are on the same side of the line of no effect and closely line up on a vertical axis, indicating a similar treatment effect magnitude. The confidence intervals for each study's treatment effect (horizontal line) overlap one another, and none cross the line of no effect, indicating a similar estimation of the population treatment effect between studies. These qualitative results suggest that there is little heterogeneity. §The chi-square test for heterogeneity was nonsignificant. The I^2 value was zero. These quantitative results suggest that there was little between-study variability (ie, heterogeneity). Adapted from Bjordal JM, Johnson MI, Lopes-Martins RA, Bogen B, Chow R, Ljunggren AE. Short-term efficacy of physical interventions in osteoarthritic knee pain. A systematic review and meta-analysis of randomised placebo-controlled trials. *BMC Musculoskelet Disord*. 2007;8:51. © 2007 Bjordal et al; licensee BioMed Central Ltd.

However, despite having statistical tests for statistical heterogeneity, there are no accepted guidelines for when a meta-analysis should not be completed due to statistical heterogeneity, and it is left to the author's discretion to determine if a meta-analysis is appropriate.

Clinical heterogeneity refers to differences in study methods that affect the ability to compare and/or combine data from different studies. Examples of differences in study methods that may lead to clinical heterogeneity include differences in participant demographics, such as risk or severity of disease, the settings in which the research was conducted, the frequency and intensity of the intervention, and how outcomes were measured across studies.[54] While there are statistical tests to estimate the extent of statistical heterogeneity, there are no tests to determine the extent of clinical heterogeneity. Researchers and clinicians must decide if the studies contributing to a meta-analysis are similar enough clinically to make meta-analysis sensible.

Whether the amount of clinical heterogeneity is too great to warrant meta-analysis is a matter of judgment. In some instances, authors may decide not to conduct a meta-analysis because the clinical heterogeneity is too great[2]; in others, they may decide to do so, using a subgroup analysis to explore the origin of the clinical heterogeneity[8] (for example, to insure that certain included populations are not affected adversely by a generally overall favorable intervention for the general population in the meta-analysis). Other authors may attempt to minimize clinical heterogeneity within a meta-analysis by limiting study eligibility. Carey et al[12] decided which studies should be included in an allograft meta-analysis before starting the study, based on acceptable clinical heterogeneity and the quality of assessment for inclusion. This approach, while reducing heterogeneity, typically results in the total number of articles included on a topic to be reduced.

MODELING DATA: FIXED- AND RANDOM-EFFECTS MODELS

The 2 most frequently used models to conduct a meta-analysis are the fixed- and random-effects models,[53] each of which handles statistical heterogeneity differently. Although the assumptions of each model differ, they frequently lead to similar results when heterogeneity is not extreme. These findings may lead one to conclude that the choice of a fixed- or random-effects model is not critical, which is an incorrect conclusion. Understanding the assumptions of each model sheds light on when one model will be more appropriate than the other.

The fixed- and random-effects models differ in assumptions related to the observed differences among study results. The 2 models are actually answering slightly different questions. In the fixed-effects model, the question is "What is the best estimate of the population effect size?"[18] An assumption of the fixed-effects model is that among a fixed set of studies there is a common treatment effect[18,26] and between-study differences in results occur by chance.[22,53] In other words, the true treatment effect is assumed to be fixed[22] and variability of between-study results is not incorporated into the model. Because of this assumption of fixed treatment effect, larger studies are given greater weight than the smaller studies.[11] Different calculation methods are available under the fixed-effects model.[22] Three common fixed-effect methods are the inverse variance method, the Mantel-Haenszel method, and the Peto method.[22]

The random-effects model, which

Review: Laser knee osteoarthritis
Comparison: 05 Pain outcome RCTs, knee osteoarthritis
Outcome: 06 Pain relief on 100-mm VAS, end of therapy with optimal dose

Study/Subcategory	n	Treatment (Mean ± SD)	n	Control (Mean ± SD)	WMD (Fixed) 95% CI	Weight, %	WMD (Fixed) 95% CI
02 Transcutaneous electrical stimulation							
Fargas Babjak	19	56.00 ± 37.50	18	10.70 ± 64.00		0.68	45.30 (11.26, 79.34)
Yurtkuran	25	25.00 ± 12.00	25	5.00 ± 8.90		22.83	20.00 (14.14, 25.86)
Cheing	16	58.00 ± 27.00	16	4.96 ± 42.40		1.29	53.04 (28.41, 77.67)
Adeboyin	15	68.70 ± 24.00	15	43.00 ± 24.00		2.65	25.70 (8.52, 42.88)
Cheing	30	35.20 ± 41.50	10	30.00 ± 24.10		1.75	5.20 (–15.94, 26.34)
Defrin	33	28.40 ± 11.10	21	5.70 ± 13.40		16.59	22.70 (15.83, 29.57
Law	25	46.70 ± 24.40	9	17.00 ± 39.70		1.02	29.70 (2.06, 57.34)
Subtotal (95% CI)	163		114			46.82	22.21 (18.12, 26.30)
Test for heterogeneity: χ^2 = 11.28, *df* = 6 (*P* = .08), I^2 = 46.8%†							
Test for overall effect: *z* = 10.65 (*P*<.00001)							

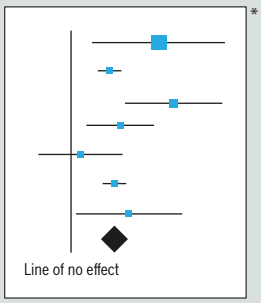

FIGURE 2. Forest plot suggesting heterogeneity. Abbreviations: CI, confidence interval; RCT, randomized controlled trial; VAS, visual analog scale; WMD, weighted mean difference. *Qualitative analysis of heterogeneity. A qualitative visual analysis of the studies' results suggests between-study variability. The individual study point estimates of the treatment effect (blue squares) are on the same side of the line of no effect but do not line up on a vertical axis, indicating a difference in treatment effect magnitude among studies. The confidence intervals for each study's treatment effect (horizontal lines) overlap one another, but the upper and lower limits of the CI do not consistently line up on a vertical axis, indicating differences in estimation of the population treatment effect among studies. These qualitative results suggest that there is heterogeneity. †Quantitative tests of heterogeneity. The chi-square test for heterogeneity was significant at a level of less than .10. The I^2 value was 47%. These quantitative results suggest there was study variability (ie, heterogeneity). Adapted from Bjordal JM, Johnson MI, Lopes-Martins RA, Bogen B, Chow R, Ljunggren AE. Short-term efficacy of physical interventions in osteoarthritic knee pain. A systematic review and meta-analysis of randomised placebo-controlled trials. *BMC Musculoskelet Disord*. 2007;8:51. © 2007 Bjordal et al; licensee BioMed Central Ltd.

assumes a distribution of treatment effects, answers the question "What is the average treatment effect?" The random-effects model assumes a distribution of the treatment effect for some populations, meaning that the treatment effect falls along a range of values, not a single value, as in the fixed-effects model. Because of this distribution, the effect size may be positive for some populations but may be negative or harmful for others.[18,29]

Studies included in the meta-analysis using a random-effects model are assumed to represent a random sample of a population of studies. The results of each study included in the meta-analysis represent a study-specific effect size that varies around a mean population effect size.[18,26] In other words, the results of each study in the meta-analysis are assumed to represent a unique effect. Because of this assumption, larger studies are given proportionally less weight, while smaller studies are given proportionally more weight.[11] In the random-effects model, the unique effect of each study is accounted for in the calculation.[22,26] A calculation method under the random-effects model is the DerSimonian and Laird method.[22]

When heterogeneity is present, the random-effects model will weight the studies comprising the meta-analysis more equally, resulting in smaller studies having greater relative influence on the combined overall effect than in the fixed-effects model.[50] To the extent that smaller studies overestimate treatment effects, a random-effects model may overestimate treatment effects when heterogeneity is present. In this case, one recommendation is to compare the fixed- and random-effects models.[50]

EXAMING STUDIES FOR HETEROGENEITY

High heterogeneity may indicate that it is inappropriate to combine studies in a meta-analysis. Heterogeneity can be visualized using forest plots (**FIGURES 1** and **2**); however, like any graph, the interpretation is not absolute.

Two statistical methods to analyze statistical heterogeneity that are frequently reported are the Cochran Q test (also known as chi-square test for heterogeneity or the chi-square test for homogeneity) and the I^2 (also known as Higgins I^2).

The Cochran Q tests whether the individual studies' treatment effects are farther away from the common effect, beyond what is expected by chance.[22,30] When the chi-square test is significant, statistical heterogeneity is present. This test has low power when few studies make up the meta-analysis,[30] and, as a result, a nonsignificant test may lead to the wrong conclusion regarding heterogeneity. A compensation for the low power of the Cochran Q is to test for heterogeneity at an alpha level of .10, rather than at .05, thereby increasing the chance of finding heterogeneity.

The test can have excessive power when there are many large studies,[30] which is similar to the problems encountered in other statistics with large sample

sizes. Finally, the Cochran Q reduces the question of heterogeneity to a dichotomy based on the *P* value, so there is no quantification of the amount of heterogeneity, just whether or not there is statistically significant heterogeneity.

A different question to ask is "How much heterogeneity is present?" The I^2 statistic was developed to answer this question.[30] The range of I^2 is from 0% to 100%. This percentage represents the percentage of total variation across studies due to heterogeneity. The test is not influenced by the number of studies in the meta-analysis, and, rather than a dichotomy, the results indicate how much heterogeneity is present. Another advantage of I^2 is that this test can be interpreted similarly, regardless of the type of outcome data and choice of effect measure. An important disadvantage of I^2 is that there are no empirically developed cut-points to determine when there is too much heterogeneity to do a meta-analysis. Higgins et al[30] have suggested rule-of-thumb interpretations, such as 25% equals low heterogeneity, 50% equals medium heterogeneity, and 75% equals high heterogeneity. Other authors may choose a cut-off for heterogeneity when choosing whether a fixed or random model is appropriate.

The decision to move forward with the meta-analysis or stop at the systematic review should be made based on the results of the test of heterogeneity and clinical judgment. High heterogeneity implies dissimilarity in the studies, and a meta-analysis should be conducted with caution. The question that the informed clinician should evaluate is whether this amount of heterogeneity is so large that the results of the meta-analysis are problematic.

OTHER CONSIDERATIONS

Event Rarity

Event rarity usually leads to overestimate of effect size.[36] For example, the result that functional bracing instead of operative intervention prevented Achilles tendon rupture[32] was acknowledged by the authors to be a possible result of both low event occurrence and numerous small randomized clinical trials included in the meta-analysis. When the event in question happens rarely or infrequently, or is captured in small numbers within small trials, caution should be exercised in interpreting the meta-analysis. While a random-effects model is advocated by many authors, a fixed-effects model should be considered, because the random effects model is influenced more by smaller studies.

Subgroup Analysis

The meta-analysis produces an average effect for all the trials included in the analysis. A question that may be asked is, does this average treatment effect hold true for all types of groups included in the meta-analysis? A subgroup analysis can be used to compare the overall effect of the meta-analysis to a particular subgroup of patients within the meta-analysis. A subgroup effect can occur by chance.[20] Because of the possibility of a chance finding, care should be taken when interpreting the results of a subgroup analysis. Some authors suggest determining prior to the analysis which factors to include in a subgroup analysis.[21] Risk stratification encourages the same phenomenon, in which those at higher risk are more likely to benefit from a treatment. In fact, detrimental effects of a treatment may outweigh the benefits in the low-risk group. If the results in the subgroups are very different, then using meta-analysis to produce an average effect may not be appropriate. Lastly, including randomized trials in the meta-analysis does not mean that comparisons between the trials are random,[20] making interpretation of subgroup analysis difficult.

Sensitivity Analysis

While a subgroup analysis attempts to estimate a treatment effect for a particular subgroup, a sensitivity analysis is used to determine if the meta-analysis findings change when different decisions related to the systematic review/meta-analysis process are made.[21] For example, a sensitivity analysis could be conducted to determine if a fixed- versus random-effects analysis reach different conclusions.[21]

Meta-Regression

Often, the studies included in a meta-analysis vary in their study characteristics (eg, variations in participant characteristics). Rather than not accounting for these differences, a meta-regression tries to relate the size of the effect to characteristics of the studies involved.[55] Conceptually, meta-regression is similar to regression. The predictor variables are the characteristics of the studies (ie, sample size, randomization) that influence the effect size, which is the outcome variable.[21] The associations derived from meta-regressions are observational and have a weaker interpretation than the causal relationships derived from randomized studies. This applies particularly when averages of patient characteristics in each trial are used as covariates in the regression. Data dredging is the main pitfall in reaching reliable conclusions from meta-regression. It can only be avoided by prespecification of variables that are believed to be potential sources of heterogeneity.[55] While some sources of heterogeneity may be expected due to differences in study design (use and nonuse of randomization), others sources of heterogeneity related to patient characteristics require expert knowledge of the clinical area.

Meta-Analysis Using Individual Patient Data (IPD)

Some critics of meta-analysis have stated that group-level analysis can lead to overestimation of low-occurrence events and nondetection of important subset effects. In theory, using IPD allows for a larger sample size and potentially better standardization of the data, because the data are reanalyzed by the researchers doing the meta-analysis, rather than relying

on the estimates provided in group-level data. For these reasons, some researchers recommend use of IPD, rather than group-level data alone, when possible.[5] However, IPD analysis has a number of drawbacks. For example, IPD analysis techniques are not standardized, IPD data require sharing of the data among investigators, the use of IDP requires all data to be available,[52] and the research questions, recruitment criteria, and measures of treatment effect need to be similar. Differences in how data are collected can lead to limitations in the use of IDP. When any of these elements differ, uncertainty or error is introduced into the overall analysis. However, IPD is certainly an alternative that can be appealing when answering questions related to group or subset differences.

When Meta-Analysis Results are Counterintuitive

When results of the meta-analysis are counterintuitive, clinical judgment will be needed to decipher the unexpected results. This judgment, based on experience, education, and current practices, will help the clinician to determine whether to accept the findings or question the statistical technique. One option is to go back to the original articles, reassess the inclusion of the articles into the analysis, and judge whether the inclusion made clinical sense. In meta-analysis there can be a loss of the original individual study assumptions. Similar to assumptions regarding the research question and data collection in a given study, assumptions of the original research question can be lost when the studies are combined together.[25]

Software Packages

There are a variety of free and proprietary software packages for meta-analysis.[4] One well-known software package offered by the Cochrane Collaboration is RevMan (http://www.cc-ims.net/revman). Another free meta-analysis software package was recently described by Wallace et al.[57]

CRITICISMS OF META-ANALYSIS

Meta-analysis is not without its critics. Rosenthal[43] cautions that, while there are many positive aspects to meta-analysis, researchers must be cognizant that studies may vary considerably in their operational definitions of the independent and dependent variables, methods of measuring variables, data-analytic approaches, and results.

Eysenck[24] has published numerous criticisms of the assumptions and the techniques of meta-analysis. Eysenck[24] believes that meta-analysis encourages a narrow focus on the effect size, without consideration of other aspects of the included studies, such as methods or individual study outcomes that are in opposite directions. This may lead to an erroneous conclusion. As with many other statistical techniques, a focus on singular aspects of the data (eg, the overall effect) can lead to conflicting interpretations, thus the entire body of evidence should be part of the interpretation of any statistical analysis. Appraising a systematic review using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement[35,37] or a standardized appraisal tool, such as the AMSTAR (a measurement tool to assess the methodological quality of systematic reviews) tool[46,47] will help clinicians focus on all aspects of the systematic review. A meta-analysis should, at a minimum, include forest plots with effect size estimates and confidence intervals for each included study, a measure of heterogeneity, and the meta-analysis overall treatment effect and related confidence interval.

Eysenck[24] also argues that only meta-analysis of a simple question is valid and, when several studies are positive but not significant because of insufficient statistical power, using meta-analysis to examine effect size can lead to spurious conclusions. Clinicians should examine the results of the systematic review and the protocol for article inclusion to judge whether a research question can be explored with meta-analysis.[2]

Other critics[25,33] have claimed that meta-analysis techniques represent the destruction of the scientific methods formed to provide statistical accuracy and reproducibility in research. The aggregation of the data is a concern and, as discussed earlier in this paper, without a good systematic review and a protocol for inclusion of studies into the meta-analysis, these concerns may be valid.

Most of these criticisms can be addressed by conducting a quality systematic review and then deciding whether meta-analysis is appropriate. Having a study team with expertise in the area of the research topic, in searching databases, and in the technique of meta-analysis can help guard against some of these potential problems.

CONCLUSION

Meta-analysis is a valuable tool for researchers, because this technique allows a reexamination of treatment effect of several studies using a larger sample than is possible for most researchers to recruit on their own. In many ways, meta-analysis allows, without additional clinical resources, exploration of potential treatment benefits or drawbacks and utilizes information made available in smaller clinical trials.

Clinicians reading the results of a meta-analysis should have a clear understanding of the strengths and limitations of the technique. In clinical medicine, many small studies are performed due to lack of access to patients, resources for conducting studies, or other forces that drive clinical practice. Meta-analysis provides a way to reevaluate the results of a particular clinical question. Meta-analysis can be misleading if the studies included are dissimilar in their research question or collect different types of outcome data. Meta-analysis, like any other statistical method, is unable to identify whether the data being utilized are appropriate. It is the responsibility of clinicians and researchers in the field to be

well informed about the evaluation and interpretation of the research information before them, so as to make good clinical decisions for their patients.

REFERENCES

1. Ada L, Dorsch S, Canning CG. Strengthening interventions increase strength and improve activity after stroke: a systematic review. *Aust J Physiother*. 2006;52:241-248.
2. Alexander LD, Gilman DR, Brown DR, Brown JL, Houghton PE. Exposure to low amounts of ultrasound energy does not improve soft tissue shoulder pathology: a systematic review. *Phys Ther*. 2010;90:14-25. http://dx.doi.org/10.2522/ptj.20080272
3. American Physical Therapy Association. APTA Vision Statement for Physical Therapy 2020. Available at: http://www.apta.org/AM/Template.cfm?Section=Vision_20201&Template=/TaggedPage/TaggedPageDisplay.cfm&TPLID=285&ContentID=32061. Accessed July 22, 2000.
4. Bax L, Yu LM, Ikeda N, Moons KG. A systematic comparison of software dedicated to meta-analysis of causal studies. *BMC Med Res Methodol*. 2007;7:40. http://dx.doi.org/10.1186/1471-2288-7-40
5. Berlin JA, Santanna J, Schmid CH, Szczech LA, Feldman HI. Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. *Stat Med*. 2002;21:371-387.
6. Bhandari M, Devereaux PJ, Montori V, Cina C, Tandan V, Guyatt GH. Users' guide to the surgical literature: how to use a systematic literature review and meta-analysis. *Can J Surg*. 2004;47:60-67.
7. Bhandari M, Montori VM, Devereaux PJ, Wilczynski NL, Morgan D, Haynes RB. Doubling the impact: publication of systematic review articles in orthopaedic journals. *J Bone Joint Surg Am*. 2004;86-A:1012-1016.
8. Bjordal JM, Johnson MI, Lopes-Martins RA, Bogen B, Chow R, Ljunggren AE. Short-term efficacy of physical interventions in osteoarthritic knee pain. A systematic review and meta-analysis of randomised placebo-controlled trials. *BMC Musculoskelet Disord*. 2007;8:51. http://dx.doi.org/10.1186/1471-2474-8-51
9. Bloch J. Interpreting odds ratios and logistic regression: a neccessity for evidence-based practice. *Am J Nurse Pract*. 2008;12:39-43.
10. Bolland MJ, Grey A, Reid IR. The randomised controlled trial to meta-analysis ratio: original data versus systematic reviews in the medical literature. *N Z Med J*. 2007;120:U2804.
11. Borenstein M, Hedges L, Higgins J, Rothstein H. *Introduction to Meta-Analysis*. Chichester, West Sussex, UK: John Wiley & Sons; 2009.
12. Carey JL, Dunn WR, Dahm DL, Zeger SL, Spindler KP. A systematic review of anterior cruciate ligament reconstruction with autograft compared with allograft. *J Bone Joint Surg Am*. 2009;91:2242-2250. http://dx.doi.org/10.2106/JBJS.I.00610
13. Chan KS, Morton SC, Shekelle PG. Systematic reviews for evidence-based management: how to find them and what to do with them. *Am J Manag Care*. 2004;10:806-812.
14. Chung KC, Burns PB, Kim HM. A practical guide to meta-analysis. *J Hand Surg Am*. 2006;31:1671-1678. http://dx.doi.org/10.1016/j.jhsa.2006.09.002
15. Clare HA, Adams R, Maher CG. A systematic review of efficacy of McKenzie therapy for spinal pain. *Aust J Physiother*. 2004;50:209-216.
16. Cochrane Collaboration. Glossary of Terms in The Cochrane Collaboration Version 4.2.5. Available at: http://www.cochrane.org/sites/default/files/uploads/glossary.pdf. Accessed March 22, 2005.
17. Cohen J. *Statistical Power Analysis for the Behavioural Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc; 1988.
18. Cohn LD, Becker BJ. How meta-analysis increases statistical power. *Psychol Methods*. 2003;8:243-253. http://dx.doi.org/10.1037/1082-989X.8.3.243
19. Cuijpers P, van Straten A, Warmerdam L. Problem solving therapies for depression: a meta-analysis. *Eur Psychiatry*. 2007;22:9-15. http://dx.doi.org/10.1016/j.eurpsy.2006.11.001
20. Davey Smith G, Egger M, Phillips A. Meta-analysis. Beyond the grand mean? *BMJ*. 1997;315:1610-1614.
21. Deeks J, Higgins JP, Altman D. Analysing data and undertaking meta-analyses. In: Higgins JP, Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions*. Chichester, West Sussex, UK: John Wiley & Sons; 2008:243-296.
22. Deeks JJ, Altman D, Bradburn M. Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis. In: Egger M, Smith GD, Altman D, eds. *Systematic Reviews in Health Care: Meta-Analysis in Context*. London, UK: BMJ Books; 2001:285-312.
23. Doll H, Carney S. Statistical approaches to uncertainty: P values and confidence intervals unpacked. *Equine Vet J*. 2007;39:275-276.
24. Eysenck HJ. Meta-analysis of best-evidence synthesis? *J Eval Clin Pract*. 1995;1:29-36.
25. Feinstein AR. Meta-analysis: statistical alchemy for the 21st century. *J Clin Epidemiol*. 1995;48:71-79.
26. Fleiss JL. The statistical basis of meta-analysis. *Stat Methods Med Res*. 1993;2:121-145.
27. Glass GV. Primary, secondary, and meta-analysis of research. *Educ Res*. 1976;5:3-8.
28. Guyatt G, Rennie D. *Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice*. Chicago, IL: American Medical Association; 2002.
29. Hedges LV, Pigott TD. The power of statistical tests in meta-analysis. *Psychol Methods*. 2001;6:203-217.
30. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ*. 2003;327:557-560. http://dx.doi.org/10.1136/bmj.327.7414.557
31. Kane RL, Saleh KJ, Wilt TJ, Bershadsky B. The functional outcomes of total knee arthroplasty. *J Bone Joint Surg Am*. 2005;87:1719-1724. http://dx.doi.org/10.2106/JBJS.D.02714
32. Khan RJ, Fick D, Keogh A, Crawford J, Brammar T, Parker M. Treatment of acute achilles tendon ruptures. A meta-analysis of randomized, controlled trials. *J Bone Joint Surg Am*. 2005;87:2202-2210. http://dx.doi.org/10.2106/JBJS.D.03049
33. Lau J, Ioannidis JP, Schmid CH. Summing up evidence: one answer is not always enough. *Lancet*. 1998;351:123-127. http://dx.doi.org/10.1016/S0140-6736(97)08468-7
34. Lewis S, Clarke M. Forest plots: trying to see the wood and the trees. *BMJ*. 2001;322:1479-1480.
35. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Ann Intern Med*. 2009;151:W65-94.
36. Lipsey M, Wilson D. *Practical Meta-Analysis*. Thousand Oaks, CA: Sage Publications; 2000.
37. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Ann Intern Med*. 2009;151:264-269, W264.
38. Mollon B, da Silva V, Busse JW, Einhorn TA, Bhandari M. Electrical stimulation for long-bone fracture-healing: a meta-analysis of randomized controlled trials. *J Bone Joint Surg Am*. 2008;90:2322-2330. http://dx.doi.org/10.2106/JBJS.H.00111
39. Montori VM, Kleinbart J, Newman TB, et al. Tips for learners of evidence-based medicine: 2. Measures of precision (confidence intervals). *CMAJ*. 2004;171:611-615. http://dx.doi.org/10.1503/cmaj.1031667
40. Patsopoulos NA, Analatos AA, Ioannidis JP. Relative citation impact of various study designs in the health sciences. *JAMA*. 2005;293:2362-2366. http://dx.doi.org/10.1001/jama.293.19.2362
41. Pittler MH, Brown EM, Ernst E. Static magnets for reducing pain: systematic review and meta-analysis of randomized trials. *CMAJ*. 2007;177:736-742. http://dx.doi.org/10.1503/cmaj.061344
42. Ried K. Interpreting and understanding meta-analysis graphs--a practical guide. *Aust Fam Physician*. 2006;35:635-638.
43. Rosenthal R, DiMatteo MR. Meta-analysis: recent developments in quantitative methods for literature reviews. *Annu Rev Psychol*. 2001;52:59-82. http://dx.doi.org/10.1146/annurev.psych.52.1.59
44. Santamaria LJ, Webster KE. The effect of fatigue on lower-limb biomechanics during single-limb landings: a systematic review. *J Orthop Sports Phys Ther*. 2010;40:464-473. http://dx.doi.org/10.2519/jospt.2010.3295

45. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA*. 1995;273:408-412.
46. Shea BJ, Grimshaw JM, Wells GA, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Med Res Methodol*. 2007;7:10. http://dx.doi.org/10.1186/1471-2288-7-10
47. Shea BJ, Hamel C, Wells GA, et al. AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. *J Clin Epidemiol*. 2009;62:1013-1020.
48. Sim J, Reid N. Statistical inference by confidence intervals: issues of interpretation and utilization. *Phys Ther*. 1999;79:186-195.
49. Sistrom CL, Garvan CW. Proportions, odds, and risk. *Radiology*. 2004;230:12-19. http://dx.doi.org/10.1148/radiol.2301031028
50. Sterne JA, Egger M, Moher D. Addressing reporting biases. In: Higgins JP, Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions*. Chichester, West Sussex, UK: John Wiley & Sons; 2008:297-333.
51. Stratford PW. The added value of confidence intervals. *Phys Ther*. 2010;90:333-335. http://dx.doi.org/10.2522/ptj.2010.90.3.333
52. Sud S, Douketis J. The devil is in the details...or not? A primer on individual patient data meta-analysis. *Evid Based Med*. 2009;14:100-101.
53. Sutton AJ, Higgins JP. Recent developments in meta-analysis. *Stat Med*. 2008;27:625-650. http://dx.doi.org/10.1002/sim.2934
54. Thompson SG. Why and how sources of heterogeneity should be investigated. In: Egger M, Davey Smith G, Altman D, eds. *Systematic Reviews in Health Care: Meta-Analysis in Context*. London, UK: BMJ Publishing Group; 2001:157-175.
55. Thompson SG, Higgins JP. How should meta-regression analyses be undertaken and interpreted? *Stat Med*. 2002;21:1559-1573. http://dx.doi.org/10.1002/sim.1187
56. Thompson SG, Pocock SJ. Can meta-analyses be trusted? *Lancet*. 1991;338:1127-1130.
57. Wallace BC, Schmid CH, Lau J, Trikalinos TA. Meta-Analyst: software for meta-analysis of binary, continuous and diagnostic data. *BMC Med Res Methodol*. 2009;9:80. http://dx.doi.org/10.1186/1471-2288-9-80
58. Walton DM, Pretty J, MacDermid JC, Teasell RW. Risk factors for persistent problems following whiplash injury: results of a systematic review and meta-analysis. *J Orthop Sports Phys Ther*. 2009;39:334-350. http://dx.doi.org/10.2519/jospt.2009.2765
59. Wood L, Egger M, Gluud LL, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *BMJ*. 2008;336:601-605. http://dx.doi.org/10.1136/bmj.39465.451748.AD
60. Zhang W, Nuki G, Moskowitz RW, et al. OARSI recommendations for the management of hip and knee osteoarthritis: part III: Changes in evidence following systematic cumulative update of research published through January 2009. *Osteoarthritis Cartilage*. 2010;18:476-499. http://dx.doi.org/10.1016/j.joca.2010.01.013
61. Zlowodzki M, Poolman RW, Kerkhoffs GM, Tornetta P, 3rd, Bhandari M. How to interpret a meta-analysis and judge its value as a guide for clinical practice. *Acta Orthop*. 2007;78:598-609. http://dx.doi.org/10.1080/17453670710014284

MORE INFORMATION
WWW.JOSPT.ORG