How to interpret a meta-analysis and judge its value as a guide for clinical practice

Michael Zlowodzki[1], Rudolf W Poolman[2], Gino M Kerkhoffs[3], Paul Tornetta III[4], and Mohit Bhandari[1]

On behalf of the International Evidence-Based Orthopedic Surgery Working Group

[1]Division of Orthopedic Surgery, McMaster University, Hamilton, Ontario, Canada, Departments of Orthopedic Surgery, [2]Onze Lieve Vrouwe Gasthuis, Amsterdam, the Netherlands, [3]Kantonsspital, St. Gallen, Switzerland, [4]Boston University, Boston, MA, USA
Correspondence: ieboswg@gmail.com
Submitted 07-05-10. Accepted 07-08-22

In the era of evidence-based orthopedics, the number of meta-analyses has dramatically increased in the last decade (Bhandari et al. 2001). Meta-analyses statistically combine the results of multiple studies and are considered to be the highest level of evidence when the results of high-quality randomized trials are combined in an appropriate way (Wright et al. 2003). Results from meta-analyses can vary widely from the truth because of large variation in the quality of the studies that have been pooled, and clinical and methodological differences between the pooled studies (Thompson and Pocock 1991, Detsky et al. 1992, Khan et al. 1996, Bhandari et al. 2001). The popularity of meta-analysis in the orthopedic literature provides one compelling reason for understanding the principles of critical appraisal.

The purpose of this article is to educate the reader about how to interpret a meta-analysis. We explain important aspects of conducting a meta-analysis as described in the guidelines for the reporting of meta-analyses of randomized controlled trials (QUOROM) (Moher et al. 1999a). An understanding of the basic terminology and concepts of a meta-analysis can help readers to evaluate the quality of a meta-analysis and to assess the potential relevance of its results for an individual patient.

Narrative reviews vs. systematic reviews

Several types of summary articles exist. Traditional narrative summaries do not, by definition, systematically identify the studies they include and do not combine their results statistically. Inclusion of studies in narrative reviews is largely selective and often based on the authors' point of view. Narrative reviews can be associated with large degrees of bias due to unsystematic study selection. Sometimes narrative reviews are even inconsistent with evidence, or lag behind evidence (Antman et al. 1992). As implied by the name, systematic reviews use a systematic approach for identification and selection of studies. A systematic review can be called a meta-analysis when statistical techniques are used to combine the data of the individual studies. Whereas narrative reviews frequently give a broad overview of a topic, systematic reviews investigate a more focused clinical question that may provide more insight into a specific aspect of patient management.

A classic example of a narrative review would be a book chapter on hip fractures that encompasses epidemiology, anatomy, radiology, and several treatment options (Baumgaertner and Higgins 2001). In such a review several studies might be cited, but they would not usually be identified in a systematic manner—or the results combined statistically. As opposed to a narrative review, a meta-

DOI 10.1080/17453670710014284

Text box 1. Evaluating a meta-analysis (Oxman and Guyatt 1991, Bhandari et al. 2004a)

Was the research question focused and clearly described?
Was the literature research systematic and reproducible?
Was the study selection process systematic?
Were the characteristics of the studies included presented?
Was a quality assessment of the studies included performed?
Were the outcome parameters objective and clinically relevant?
Were the statistical methods used to combine the studies reported?
Were the pooled studies homogenous?
If not, were sensitivity analyses conducted to explore sources of heterogeneity?
Was publication bias assessed?

analysis concentrates on a specific aspect of hip fractures and identifies and combines the available data in a systematic manner: e.g., a meta-analysis comparing the relative merits of internal fixation and arthroplasty in displaced femoral neck fractures (Bhandari et al. 2003).

Because of the systematic nature of study selection and data synthesis, systematic reviews and meta-analyses are less prone to bias than narrative reviews. Probably for this reason, rigorous systematic reviews have been found to be cited twice as often as narrative reviews and to receive more citations than any other form of study design (Bhandari et al. 2004b, Patsopoulos et al. 2005).

Where do meta-analyses fit into the hierarchy of evidence?

Randomized controlled trials (RCTs) give less biased estimates of treatment effects than observational studies because randomization balances known and unknown prognostic factors between the interventions that are being compared. Thus, meta-analyses are often used to summarize the results of RCTs. Although observational studies can also be summarized in a meta-analysis, the inherent biases associated with observational studies and confounding factors make the results of a meta-analysis of observational studies more prone to bias. Because of the balancing of prognostic factors afforded by randomization, RCTs and meta-analyses of RCTs are considered to be the highest level of evidence concerning a research question about a therapy. However, drawing conclusions from single RCTs can be problematic, because (1) they often lack the power to detect clinically important differences, and (2) several RCTs investigating the same question can contradict each other. Meta-analysis can overcome this disadvantage and provide more precise estimates of an effect by combining study results, thereby increasing sample size. Thus, conclusions drawn from meta-analyses often have a greater influence on clinical practice than single RCTs.

Assessment of the value of a meta-analysis

General considerations

Appraisal of a meta-analysis includes a critical evaluation of the research question, of the literature research, the study selection, the data abstraction, quality assessment of the studies included, and data analysis (Oxman and Guyatt 1991, Bhandari et al. 2004a) (Text box 1). In a meta-analysis that has been done well, all these steps are explicitly described and are reproducible. Lastly, as a prerequisite for understanding a meta-analysis, readers should understand some basic terminology (e.g. relative risk, odds ratios, weighted mean differences, standard mean differences, heterogeneity) and understand how to interpret the results in order to determine how to apply them to surgical practice (Text box 2).

Was the research question focused and clinically relevant?

Focused clinical questions such as “Does nailing or plating of proximal tibia fractures in young adults result in higher union rates?” are more likely to be applied to a specific clinical scenario than broad reviews about “aspects of surgical treatment of proximal tibia fractures”. Ideally, a meta-analysis should investigate a question that is well defined.

Text box 2. Glossary of terminology

95% confidence interval – Range of two values around the point estimate within which it is probable with 95% confidence that the true value lies, for the entire population of patients from which the study patients were selected. In theory, if the study were to be repeated 100 times under the exactly same conditions, 95 times the true value would lie within the 95% confidence interval.

Relative risk – The ratio of the risk of an event among an exposed population to the risk among the unexposed population, or vice versa. Example: the proportion of patients who develop an infection after nailing divided by the proportion of patients who develop an infection after plating.

Odds ratio – The odds of exposure to a risk factor in the population with a positive event to the odds of exposure to a risk factor in the population with no positive event, or vice versa. Example: the ratio of infected patients treated with a nail to infected patients treated with a plate divided by the ratio of uninfected patients treated with a nail to uninfected patients treated with a plate.

Number needed to treat – The number of patients who must be treated to prevent one bad outcome. It is the inverse of the absolute risk reduction between two groups under comparison.

Weighted mean difference – The average value after pooling results of individual studies. The contribution of each study to the mean difference is weighted by sample size.

Standard mean difference (effect size) – The difference in the outcome between two groups expressed as a multitude of standard deviations based on a pooled average standard deviation across both groups. The pooled standard deviation can be calculated as the root of the sum of the squared standard deviations of both groups, divided by 2. Example: group A has an average functional outcome score of 80 with a standard deviation of 12 and group B has an average functional outcome score of 86, also with a standard deviation of 12. The difference in the average functional outcome score is 6, which is equal to half the standard deviation. Thus, the effect size is 0.5.

The reader has to evaluate whether the meta-analysis he or she is reading deals with a question that is focused enough to be applicable to a real-life situation in his or her clinical practice. In order to do that, the meta-analysis must describe what population, what intervention, and what outcome parameters have been considered and what interventions are being compared.

Was the literature research systematic and reproducible?

A thorough systematic search is vital for a valid meta-analysis. The search must include multiple databases to avoid the risk of missing relevant studies. MEDLINE, EMBASE, and the Cochrane controlled trials register are the most relevant and the most frequently used databases available. While MEDLINE is a very large database with over 15 million citations covering the literature since 1966, EMBASE has a better coverage of European journals (Egger and Davey-Smith 2001). However, a search limited to these databases would still miss a lot of unpublished studies. Bhandari and co-workers found that only one-third of studies presented as abstracts at the American Academy of Orthopedic Surgeons meeting were followed by a full-text publication (Bhandari et al. 2002b). Despite controversies about inclusion of unpublished studies, mainly related to potentially lower methodological quality (Dickersin et al. 1987, Dickersin 1990, Cook et al. 1993), their omission can lead to a possible overrepresentation of studies with positive results (publication bias) (McAuley et al. 2000).

Authors of a meta-analysis should therefore make efforts to identify unpublished studies. Furthermore, it is more reassuring to the reader of a meta-analysis that all relevant studies have been identified if hand searches of bibliographies of included articles have been performed and experts in the area contacted. It is only possible for the reader to assess whether studies could have been missed if the search strategy is described in detail and is reproducible. Ideally, the literature search should be done in duplicate by two independent investigators, in order to limit random and systematic errors.

Publication bias

Publication bias refers to the higher probability of studies with positive results being published (Begg and Berlin 1989, Easterbrook et al. 1991, Dickersin and Min 1993, Stern and Simes 1997, Egger and Smith 1998). Many small studies with results that are not statistically significant are not

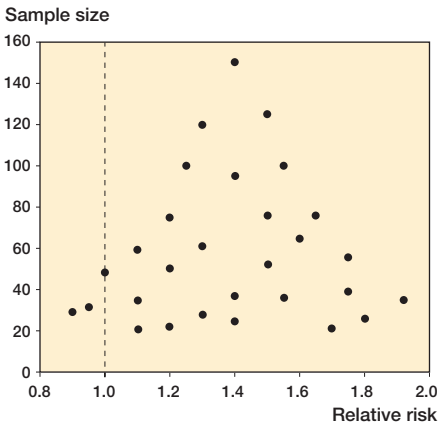

Figure 1. This figure demonstrates assessment of publication bias by plotting the relative risk point estimates of individual study results (black circles) against the sample size of the study. With decreasing sample size, study variability decreases and the point estimates of individual studies vary more widely. If no publication bias is present, this scatter plot is symmetrical and has the characteristic shape of a funnel. The dotted line symbolizes the line-of-no-effect (relative risk of 1).

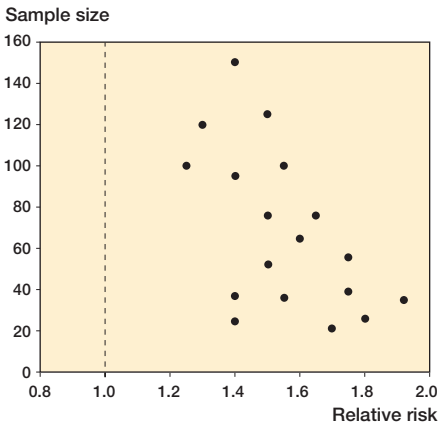

Figure 2. In contrast to Figure 1, in this figure the characteristic symmetrical funnel shape of the scatter plot is missing. Smaller studies with findings that are not statistically significant would normally be at the bottom-left side of the page, close to the line-of-no-effect (dotted line). An overrepresentation of smaller studies with significant findings on the bottom-right portion of the diagram, with few or no studies in the bottom-left part, indicates publication bias.

considered for publication—or not even submitted for publication by authors. The tendency of journal editors and reviewers to give preference to studies with positive results leaves many studies with negative results unpublished. This may bias the results of a meta-analysis if these studies cannot be identified elsewhere e.g. as abstracts of meetings. Consequently, inclusion of more studies with positive results can lead to overestimation of the treatment effect, and can possibly lead to a false-positive result. In one study that examined meta-analyses, exclusion of (originally included) unpublished studies resulted in a 15% increase in the treatment effect (McAuley et al. 2000).

Publication bias can be assessed visually by plotting the results of the individual studies on the x-axis (e.g. "odds ratio" or "relative risk" for dichotomous/binary variables or "weighted mean differences" or "standard mean differences" for continuous variables) against a measure of precision on the y-axis, such as standard error (Egger et al. 1997a). There is a correlation between sample size and standard error, so sample size can be used instead (Figure 1).

With increasing sample size, precision increases (smaller confidence intervals) and the point estimate of the investigated outcome is more likely to represent the true mean value in the population being investigated. Thus, the scattered dots are closer together at the top of the diagram, where the larger studies are displayed. Precision decreases with reduced sample size. By chance, results of small studies vary more widely and therefore are likely to be located more peripherally in the diagram representing over- or underestimations of a treatment effect. As a result of this, the dots are scattered further apart towards the bottom of the diagram where the smaller studies are displayed. If no studies have been omitted, the resulting scatter diagram is symmetrical and resembles an inverted funnel with its base at the bottom, signifying no publication bias (Figure 1). If small studies with negative results (underestimating the effect) are not published or not identified by the investigators of a meta-analysis, the scatter diagram will be asymmetrical at the bottom. This asymmetry indicates that smaller studies with negative findings are missing, either because they were not published or because they were not identified by the authors (Figure 2).

Funnel plot assessment is a qualitative analysis that can be subjective, as its interpretation is reviewer-dependent. Complementary statistical tests such as the rank correlation test have been

developed to quantify publication bias. However, the power of such a test depends on the number of studies included. In meta-analyses with less than 25 studies, its power is only moderate (Begg and Mazumdar 1994).

Language bias

Restricting the study selection to English-language journals is another potential source of bias. Egger at al. (1997b) found that the proportion of RCTs with positive results is almost double in English-language journals compared to German-language journals. Thus, if a meta-analysis excludes journals in languages other than English, an overestimation of the reported treatment effect may occur. However, another study found that the results would have been different in only 1 out of 36 consecutive meta-analyses (in 8 medical journals) if non-English articles had been included (Gregoire et al. 1995).

Was study selection systematic and reproducible?

A systematic and reproducible selection process is an important quality characteristic of a meta-analysis. This involves several steps. The first step involves screening the titles and abstracts of studies identified in the initial literature search for eligibility. In the second step, potentially eligible articles are retrieved and fully reviewed. The selection process should be documented, ideally in a flow diagram, with documentation of the number of studies excluded at each step and the reasons for exclusion (Moher et al. 1999)a (Figure 3). Having two investigators select studies independently and then resolving disagreements by consensus can help to limit selection bias.

Was the methodological quality of the studies assessed?

The quality of a meta-analysis is greatly influenced by the quality of the studies it attempts to summarize. The best meta-analysis will be of poor quality if it is based on poor-quality studies. Differences in study results between studies can be the result of differences in study methods (Moher et al. 1998). Methodologically less rigorous studies tend to overestimate treatment effects (Khan et al. 1996, Moher et al. 1998). When conducting a meta-analysis, it is therefore important to assess

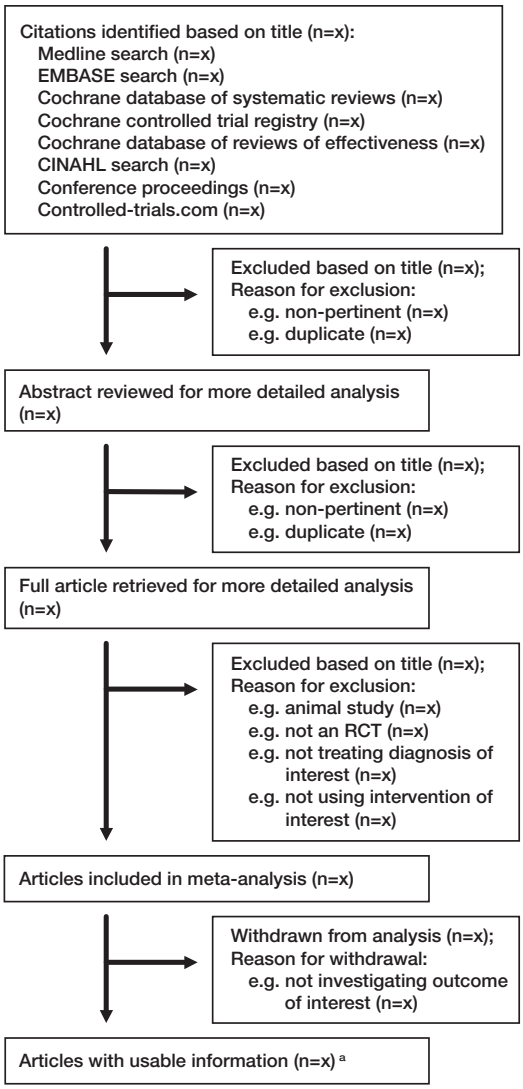

Figure 3. Flow diagram as required by the QUOROM guidelines, demonstrating the individual steps in the study selection process (Moher et al. 1999a).

the methodological quality of the studies. The majority of meta-analyses of orthopedic surgery-related topics have methodological limitations. In an assessment of 40 meta-analyses, 35 were found to have methodological flaws that could limit their validity (Bhandari et al. 2001). Lack of information on the methods used to retrieve and assess the validity of the primary studies was identified as the main deficiency.

Factors associated with bias in RCTs include lack of concealment of randomization; lack of blinding of physicians, patients, outcome assessors and data analysts, and failure to report reasons for excluding patients (Schulz et al. 1995, Juni et al. 1999, 2001, Moher et al. 1999b, Schulz 2000, Bhandari et al. 2002a, Fergusson et al. 2004). Several checklists and scales have been proposed for the assessment of study quality, which include those factors (Moher et al. 1995). While scales such as the Detsky scale can be useful (Detsky et al. 1992), many authors consider the scoring schemes to be imprecise and cutoff points can be arbitrary (Moher et al. 1995, Juni et al. 1999). The results of a meta-analysis can be biased if lower-quality studies are excluded. Presentation of the key methods of the studies included—e.g. in a table—is therefore preferable. However, if study quality scales are used it is useful to perform a sensitive analysis (i.e. subgroup analysis) to compare studies of different quality rather than excluding low-quality studies. Again, independent assessment of the methods in the studies that have been included, by two or more reviewers, can provide additional reassurance to the reader of a meta-analysis.

Blinding of outcome assessors

Unblinded outcome assessors can consciously or unconsciously bias results with personal preferences. A review of 32 RCTs published in the American Journal of Bone and Joint Surgery found that unblinded outcomes assessment was associated with a potential for exaggeration of the benefit of the effectiveness of a treatment by a factor of 3 (Poolman et al. 2007). Blinded assessment prevents detection bias. Since in surgical trials the surgeon cannot be blinded, methodological safeguards include blinding of the patients (if possible) and blinding of outcome assessors and data analysts.

Concealment of randomization

Randomization is considered concealed if the investigator cannot determine to which treatment group the next patient enrolled in the study will be allocated. Such knowledge might make the investigator prone to selectively apply exclusion and inclusion criteria, thereby biasing the results. For example, if randomization is based on date of birth or even or odd medical report numbers (clearly unconcealed), the investigator knows to which treatment the patient will be allocated before deciding to include or exclude the patient. This knowledge allows the investigator to be selective, depending on his or her own preferences—or simply his or her current mood; for example, the investigator might be in the mood to perform procedure A, but not procedure B. Since some exclusion/inclusion criteria are very subjective (e.g. "Do you think you could obtain a 1-year follow-up on this patient?"), selection bias can occur. In two separate studies, RCTs that used inadequate allocation concealment were found to result in an increased estimate of benefit by more than one-third (Schulz et al.1995, Moher et al. 1998). In an analysis of 29 orthopedic RCTs, 5 did not report the method of concealment (Li et al. 2005).

In concealed forms of randomization, exclusion and inclusion criteria are applied first, and then the treatment is determined. This limits the possibility of selection bias. Most RCTs rely on sealed envelopes for treatment allocation (Li et al. 2005). While sealed envelopes are usually a concealed form of randomization, breach of concealment is possible—and even includes illicit opening and transillumination of sealed envelopes (Schulz 1995). Ideal forms of concealed randomization are central internet- and telephone-based randomizations.

Were the data analyzed and reported appropriately?

Single RCTs often lack the power to detect important differences because of limited sample sizes. The purpose of a meta-analysis is to overcome this limitation and increase the probability of detecting differences if they exist. Combining data from individual studies accomplishes this aim by increasing the sample size. To achieve higher sample sizes and to increase the validity and generalizibility of the results, individual studies as well as meta-analyses of them often have broad inclusion criteria. The validity of results is increased if the magnitude of the results is similar across different patient populations, interventions, surgeons' level of expertise, and ways of measuring outcomes. If different patient populations—e.g. of different ages and in different countries/continents—are investigated, the validity and generalizability of

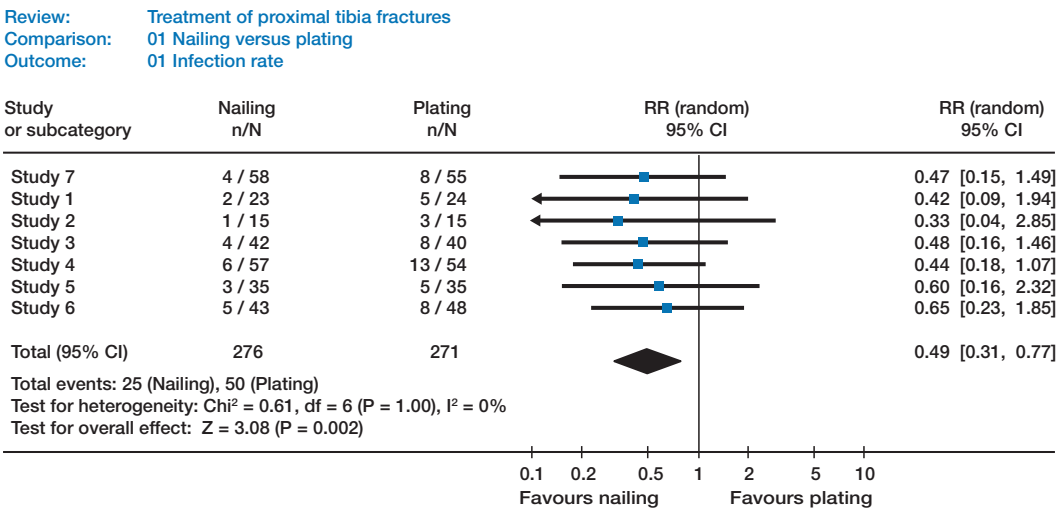

Figure 4. This figure shows a Forest plot, which is typically used to present the results of a meta-analysis. In this hypothetical example the outcome parameter is infection—a dichotomous outcome (present or absent). The relative risk of developing an infection when treated with a nail or a plate is displayed on the x-axis. Instead of relative risks, odds ratios could also be presented. The line-of-no-effect (vertical line) separates outcomes that favor nailing and plating. The squared blue boxes represent the point estimates and the horizontal lines represent the associated 95% confidence intervals for each study. This example demonstrates the power of a meta-analysis. Whereas the confidence intervals of all individual studies cross the line-of-no-effect (relative risk = 1) representing no significant differences, the confidence interval of the summary estimate (diamond shape) lies entirely to the left of the line-of-no-effect representing a significantly lower infection risk with nailing (p = 0.002). Overlapping confidence intervals of the individual studies and an I^2 value of 0% with a non-significant p-value (1.0) indicates homogeneity of the studies, and justifies presenting a summary estimate.

findings is higher, which means that these findings can be extrapolated to those different populations. However, there is a "down-side" to broad inclusion criteria. For example, surgeons may argue that femoral neck fractures in the elderly and in young adults are different problems with different treatment options; in other words, the two groups are not homogeneous and should not be combined.

Heterogeneity

Heterogeneity is defined as inconsistency in the treatment effect across primary studies (Deeks et al. 2001). While homogenous results do not exclude differences in patient populations and interventions between studies, in reverse heterogeneous results probably reflect some underlying differences in clinical and/or methodological aspects between studies. When substantial heterogeneity exists, pooling data from multiple trials and presenting a single summary estimate can be misleading (Thompson and Pocock 1991) and should be avoided. Inconsistencies in study results in a meta-analysis reduce the confidence in its conclusions. There are two basic ways of assessing heterogeneity: (1) by visually inspecting the point estimates and confidence intervals of a Forest plot, and (2) by performing a statistical comparison (Cooper and Rosenthal 1980). Results of individual studies are homogeneous when the point estimates are similar and the confidence intervals overlap (Figure 4). However, how similar is similar enough to justify pooling of studies? Statistical tests like the I^2 test exist to quantify heterogeneity (Higgins and Thompson 2002, Higgins et al. 2003).The I^2 statistic describes the percentage of total variation across studies that is attributable to heterogeneity rather than chance (Higgins et al. 2003). A value greater than 25% is considered to reflect low heterogeneity, 50% moderate, and 75% high heterogeneity. However, there is no definitive cut-off value at which no data pooling should be performed. When the associated p-value is small (< 0.1), it is unlikely that heterogeneity is due to chance alone. When the p-value is > 0.1, heterogeneity can still not necessarily be excluded as tests for heterogeneity tend to be underpowered for detecting small dif-

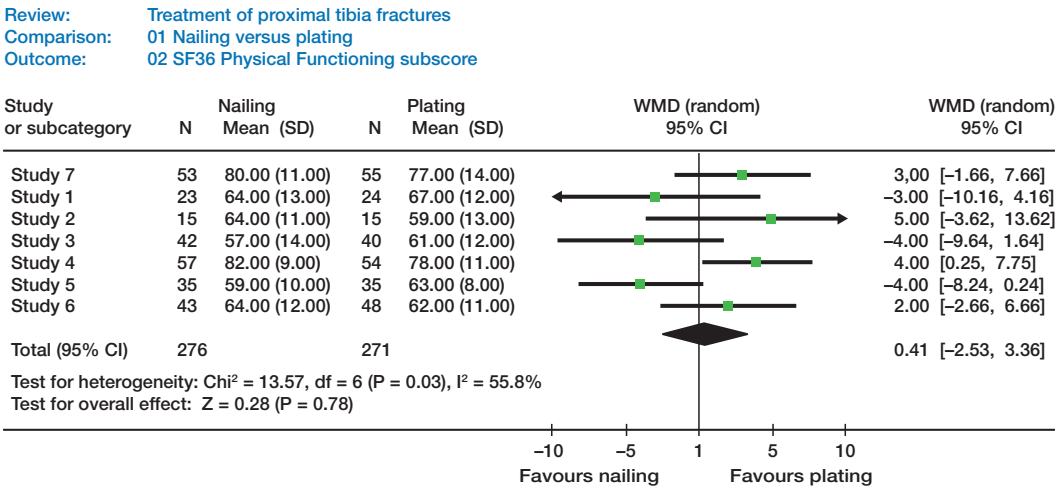

Figure 5. This figure is similar to the Forest plot in Figure 4 except that the outcome parameter is continuous (SF36 score) and the studies are heterogeneous. Continuous outcome parameter can either be presented as weighted mean differences (WMD) of the outcome parameter, as in this example, or as standard mean differences (SMD = effect sizes). This example shows several non-overlapping confidence intervals, which indicate heterogeneity of the studies. Heterogeneity is confirmed by a high I^2 value of 56% and a significant associated p-value (0.03). In the light of such a large significant heterogeneity, caution is warranted in interpreting the summary estimate (diamond shape).

ferences—especially if the number of the studies and their sample sizes are low. In turn, even when heterogeneity is low, reflected by a low I^2 value, a significant associated p-value has no meaning if the differences in result across studies are not clinically important. Authors of meta-analyses should look for explanations of heterogeneity by conducting so-called sensitivity analyses (i.e. subgroup analyses) based on a priori hypothesis. For example, if the authors believe results might differ by the patients' sex, males and females should be analyzed separately to see whether heterogeneity in those subgroups decreases. To diminish spurious false-positive findings, a priori hypotheses of potential sources of heterogeneity should be developed. Such hypotheses might include clinical aspects like sex, age, co-morbidities, differences in interventions, and also methodological aspects such as study quality or length of follow-up. I^2 values and associated p-values are generally displayed with each Forest plot (Figure 5).

Fixed-effect versus random effects model

When combining data, two statistical models can be used: (1) the fixed-effects model, and (2) the random-effects model. Fixed-effects models assume that the true effect of treatment is the same for every study, whereas random-effects models assume that an effect may vary across studies because of differences between studies. In other words, the choice of the model is based on the degree of heterogeneity between studies. If the studies differ in several aspects—and therefore a large amount of between-study variability (heterogeneity) exists—the use of a fixed-effects model is problematic. However, how much heterogeneity is too much—and can we really be sure how much heterogeneity exists between studies? Unfortunately, exact cut-off values are not defined in the literature and often the test for heterogeneity is underpowered with regard to detecting important differences. Thus, the fixed-effects model is generally used if there is no heterogeneity in a meta-analysis that includes a large number of studies, preferably with large sample sizes. Only then can the investigators and readers be confident that the test for heterogeneity had sufficient power to detect important differences. If there is any concern about heterogeneity, the random-effect has been advocated for data pooling (DerSimonian and Laird 1986). The random-effects model results in wider confidence intervals around the point estimates; however, it is a more conservative choice for the analysis, as it is not based on the assumption that studies are homo-

geneous. The reader should be suspicious about the results of a meta-analysis when I^2 values are high (> 50%) and the fixed-effects model has been used. We prefer to take the more conservative approach, and always use the random-effects model – even if studies are homogenous. However, the random-effects model also has its limitations. It has been criticized for its limitation in taking covariates into account (e.g. different study populations or co-treatments) to explain heterogeneity (Thompson and Pocock 1991, Lau et al. 1998).

Understanding the results

Classically, results of meta-analyses are presented visually in so-called Forest plots. In order to understand a Forest plot, the reader should have some knowledge of statistical terms such as point estimates, confidence intervals (CIs), relative risks (RRs), relative risk reduction (RRR), odds ratios (ORs), weighted mean differences (WMDs), standard mean differences (SMDs), and number needed to treat (NNT) (Text box 3).

Consider the following hypothetical example with a dichotomous outcome measure: "Does nailing of proximal tibia fractures result in a lower infection rate than plating?"

Text box 2 illustrates the differences between odds ratios and relative risks, and between relative risks and absolute risks. When looking at the equation, the difference between odds ratios and relative risks is the denominator. In odds ratios, the event is divided by the number of no events (infections over no infections), whereas in relative risks the number of events is divided by the total number of cases. Practically speaking, this means that if the number of events (in this example, infections) is low, odds ratios are almost equal to relative risks, as demonstrated in scenario 1 (4% vs. 2%; RR: 0.5; OR: 0.49). Odds ratios and relative risks can differ greatly, if the event rates are high—as demonstrated in scenario 2 (40% vs. 20%; RR: 0.5; OR: 0.38). Generally, either odds ratios or relative risks are presented. Odds ratios are typically calculated in case-control studies where cases that developed an event (infection in our example) are identified first and then evaluated as to whether they were exposed to a factor (treated by a nail in our example) compared to cases that did not develop an infection. In relative risks, on the other hand, one starts out with the populations (patients treated with a nail or plate) and calculates the percentage of events (infection) in each population. Mathematically, when relative risks can be calculated, odds ratios can also be calculated. Except for case-control studies, we prefer to present relative risks rather than odds ratios—simply because they are easier to understand for non-statisticians and the corresponding relative risk reductions can be calculated.

In the example above, in both scenarios nailing of proximal tibia fractures reduces the relative risk of developing an infection by 50% compared to plating. This number is easy to understand, but its meaning can be misleading. It can be more meaningful to look at the differences in absolute risks or the number needed to treat (NNT), which is the inverse of the absolute risk difference. In scenario 1 of the example, the NNT is 50—which means that for every 50 fractures that are treated with a nail as opposed to a plate, one infection can be prevented.

Text box 3. Comparison of result measures for dichotomous outcomes

Scenario 1:

	Infected	Not infected	Total
Nailing	2	98	100
Plating	4	96	100

Incidence of infection after nailing: 2/100 = 2%
Incidence of infection after plating: 4/100 = 4%
Odds ratio (OR): (96/98) / (4/2) = (2/98) / (4/96) = 0.49
Relative risk (RR): (2/100) / (4/100) = 0.5
Relative risk reduction (RRR): 1 – 0.5 = 0.5 = 50%
Absolute risk reduction (AR): 4% – 2% = 2%
Number needed to treat (NNT): 1/(2%) = 50

Scenario 2:

	Infected	Not infected	Total
Nailing	20	80	100
Plating	40	60	100

Incidence of infection after nailing: 20/100 = 20%
Incidence of infection after plating: 44/100 = 40%
Odds ratio (OR) (60/80) / (40/20) = (20/80) / (40/60) = 0.38
Relative risk (RR): (20/100) / (40/100) = 0.5
Relative risk reduction (RRR): 1 – 0.5 = 0.5 = 50%
Absolute risk reduction (AR): 40% – 20% = 20%
Number needed to treat (NNT): 1/(20%) = 5

Scenario 2 has the same relative risk; however, the NNT is 5. It seems obvious that an intervention that results in a lower NNT can be interpreted as having a greater clinical meaning.

ORs, RRs, RRRs, and NNTs are measures used for dichotomous outcome variables (present or not present). When the outcome variable under investigation is continuous, such as a functional outcome score, the summarized estimate in a meta-analysis is presented either as a weighted mean difference (WMD) or as a standard mean difference (SMD). A weighted mean difference is simply the aggregated differences of the individual study differences weighted by their sample size. However, if data from different scoring systems—e.g. SF-36 and EQ5D functional outcome scores—with different ranges is summarized, it is clear that the data cannot be aggregated simply by calculating the average difference, because the scales are different. In those cases, the data can be transformed into standard mean difference (SMD), which is also known as effect size. The SMD describes a difference between two groups (e.g. nailing and plating) in multitudes of standard deviations. A pooled standard deviation across both groups is used for calculation of the SMD. SMDs can be thought of as the percentile standing of the average experimental participant (nailing) relative to the average control participant (plating). An effect size of 0 indicates that the mean of the experimental group (e.g. nailing) is at the fiftieth percentile of the control group (e.g. plating). An effect size of 1 indicates that the mean of the experimental group is at the eighty-fourth percentile of the control group. This is based on a normal distribution of values, which is typically displayed as a Gaussian curve. Cohen (1988) defined effect sizes as small (0.2), medium (0.5), and large (0.8).

Typically, the results of a meta-analysis are displayed in Forest plots for each outcome parameter (Figures 4 and 5). Individual studies are plotted sequentially on the y-axis with a summary estimate at the bottom. The x-axis shows one of the outcome measures described above. Point estimates are represented by square boxes. The weight of a study is reflected by the size of the square. The point estimates are accompanied by a line which represents their associated 95% confidence interval. A vertical midline called the line-of-no-effect divides the diagram into a part that favors the alternative intervention (e.g. nailing) and a part that favors the control intervention (e.g. plating). A confidence interval that crosses the line-of-no-effect indicates a statistically non-significant difference, whereas a confidence interval that does not cross the midline indicates a significant difference for either the alternative or control intervention, depending on whether it is located at the left side or the right side of the midline. The aggregate summary estimate is represented as a diamond shape at the bottom of the diagram.

Translating the results into clinical practice

When translating the results of a meta-analysis into clinical practice, it is above all important to consider whether the reported outcome parameters were clinically important and measured objectively. Also, the reader should not forget that what might be important to clinicians is not necessarily important to patients. Patients may value certain risks and benefits differently from clinicians. When looking at the actual results, readers should not just look at statistical significance but they should see the results in the light of their clinical importance. A statistically significant difference has no meaning if it is clinically irrelevant; any difference can be statistically significant if the sample size is large enough. On the other hand, when non-significant results are presented, the reader should not assume that there is no difference but rather look at the limits of the confidence interval. Regardless of statistical significance, the main question should always be: *"Does what I consider to be a clinically relevant value lie within or outside of the reported confidence interval?"*

In conclusion, meta-analysis can provide a useful tool to help in clinical decision making, especially in orthopedic surgery where primary studies of small sample size are the mainstay, provided that the meta-analysis is performed according to strict methodological rules as described above.

Antman E M, Lau J, Kupelnick B, Mosteller F, Chalmers T C. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction. Jama 1992; 268 (2): 240-8.

Baumgaertner M R, Higgins T F. Femoral neck fractures. In: Rockwood and Green's fractures in adults (Eds Buchholz R W, Heckman J D). Philadelphia, Lippincott Williams and Wilkins 2001: 1579-634.

Begg C B, Berlin J A. Publication bias and dissemination of clinical research. J Natl Cancer Inst 1989; 81 (2): 107-15.

Begg C B, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. Biometrics 1994; 50 (4): 1088-101.

Bhandari M, Morrow F, Kulkarni A V, Tornetta P, 3rd. Meta-analyses in orthopaedic surgery. A systematic review of their methodologies. J Bone Joint Surg (Am) 2001; 83 (1): 15-24.

Bhandari M, Richards R R, Sprague S, Schemitsch E H. The quality of reporting of randomized trials in the Journal of Bone and Joint Surgery from 1988 through 2000. J Bone Joint Surg (Am) 2002a; 84 (3): 388-96.

Bhandari M, Devereaux P J, Guyatt G H, Cook D J, Swiontkowski M F, Sprague S, Schemitsch E H. An observational study of orthopaedic abstracts and subsequent full-text publications. J Bone Joint Surg (Am) 2002b; 84 (4): 615-21.

Bhandari M, Devereaux P J, Swiontkowski M F, Tornetta P, 3rd, Obremskey W, Koval K J, Nork S, Sprague S, Schemitsch E H, Guyatt G H. Internal fixation compared with arthroplasty for displaced fractures of the femoral neck. A meta-analysis. J Bone Joint Surg (Am) 2003; 85 (9): 1673-81.

Bhandari M, Devereaux P J, Montori V, Cina C, Tandan V, Guyatt G H. Users' guide to the surgical literature: how to use a systematic literature review and meta-analysis. Can J Surg 2004a; 47 (1): 60-7.

Bhandari M, Montori V M, Devereaux P J, Wilczynski N L, Morgan D, Haynes R B. Doubling the impact: publication of systematic review articles in orthopaedic journals. J Bone Joint Surg (Am) 2004b; 86 (5): 1012-6.

Cohen J Statistical power analysis for the behavioural sciences. Edited, Hillsdale, NJ, Lawrence Earlbaum Associates, 1988.

Cook D J, Guyatt G H, Ryan G, Clifton J, Buckingham L, Willan A, McIlroy W, Oxman A D. Should unpublished data be included in meta-analyses? Current convictions and controversies. Jama 1993; 269 (21): 2749-53.

Cooper H M, Rosenthal R. Statistical versus traditional procedures for summarizing research findings. Psychol Bull 1980; 87 (3): 442-9.

Deeks M J, Altman D, Bradburn M J Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis. In Systematic reviews in health care: meta-anlysis in context, pp. 286. Edited by M Egger,G D Smith,D G Altman, 286, London, BMJ Publishing Group, 2001.

DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986; 7 (3): 177-88.

Detsky A S, Naylor C D, O'Rourke K, McGeer A J, L'Abbe K A. Incorporating variations in the quality of individual randomized trials into meta-analysis. J Clin Epidemiol 1992; 45 (3): 255-65.

Dickersin K. The existence of publication bias and risk factors for its occurrence. Jama 1990; 263 (10): 1385-9.

Dickersin K, Chan S, Chalmers T C, Sacks H S, Smith H, Jr. Publication bias and clinical trials. Control Clin Trials 1987; 8 (4): 343-53.

Dickersin K, Min Y I. NIH clinical trials and publication bias. Online J Curr Clin Trials 1993; Doc No 50.

Easterbrook P J, Berlin J A, Gopalan R, Matthews D R. Publication bias in clinical research. Lancet 1991; 337 (8746): 867-72.

Egger M, Smith G D. Bias in location and selection of studies. Bmj 1998; 316 (7124): 61-6.

Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. Bmj 1997a; 315 (7109): 629-34.

Egger M, Zellweger-Zahner T, Schneider M, Junker C, Lengeler C, Antes G. Language bias in randomised controlled trials published in English and German. Lancet 1997b; 350 (9074): 326-9.

Egger M, Davey-Smith G Principles of and procedures for systematic reviews. In Systematic reviews in health care: meta-anlysis in context, pp. 26. Edited by M Egger,G D Smith,D G Altman, 26, London, BMJ Publishing Group, 2001.

Fergusson D, Glass K C, Waring D, Shapiro S. Turning a blind eye: the success of blinding reported in a random sample of randomised, placebo controlled trials. Bmj 2004; 328 (7437): 432.

Gregoire G, Derderian F, Le Lorier J. Selecting the language of the publications included in a meta-analysis: is there a Tower of Babel bias? J Clin Epidemiol 1995; 48 (1): 159-63.

Higgins J P, Thompson S G. Quantifying heterogeneity in a meta-analysis. Stat Med 2002; 21 (11): 1539-58.

Higgins J P, Thompson S G, Deeks J J, Altman D G. Measuring inconsistency in meta-analyses. Bmj 2003; 327 (7414): 557-60.

Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. Jama 1999; 282 (11): 1054-60.

Juni P, Altman D G, Egger M. Systematic reviews in health care: Assessing the quality of controlled clinical trials. Bmj 2001; 323 (7303): 42-6.

Khan K S, Daya S, Jadad A. The importance of quality of primary studies in producing unbiased systematic reviews. Arch Intern Med 1996; 156 (6): 661-6.

Lau J, Ioannidis J P, Schmid C H. Summing up evidence: one answer is not always enough. Lancet 1998; 351 (9096): 123-7.

Li P, Mah D, Lim K, Sprague S, Bhandari M. Randomization and concealment in surgical trials: a comparison between orthopaedic and non-orthopaedic randomized trials. Arch Orthop Trauma Surg 2005; 125 (1): 70-2.

McAuley L, Pham B, Tugwell P, Moher D. Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses? Lancet 2000; 356 (9237): 1228-31.

Moher D, Jadad A R, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. Control Clin Trials 1995; 16 (1): 62-73.

Moher D, Pham B, Jones A, Cook D J, Jadad A R, Moher M, Tugwell P, Klassen T P. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 1998; 352 (9128): 609-13.

Moher D, Cook D J, Eastwood S, Olkin I, Rennie D, Stroup D F. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of Reporting of Meta-analyses. Lancet 1999a; 354 (9193): 1896-900.

Moher D, Cook D J, Jadad A R, Tugwell P, Moher M, Jones A, Pham B, Klassen T P. Assessing the quality of reports of randomised trials: implications for the conduct of meta-analyses. Health Technol Assess 1999b; 3 (12): i-iv, 1-98.

Oxman A D, Guyatt G H. Validation of an index of the quality of review articles. J Clin Epidemiol 1991; 44 (11): 1271-8.

Patsopoulos N A, Analatos A A, Ioannidis J P. Relative citation impact of various study designs in the health sciences. Jama 2005; 293 (19): 2362-6.

Poolman R W, Struijs P A, Krips R, Sierevelt I N, Marti R K, Farrokhyar F, Bhandari M. Reporting of outcomes in orthopaedic randomized trials: does blinding of outcome assessors matter? J Bone Joint Surg (Am) 2007; 89 (3): 550-8.

Schulz K F. Subverting randomization in controlled trials. Jama 1995; 274 (18): 1456-8.

Schulz K F. Assessing allocation concealment and blinding in randomized controlled trials: why bother? ACP J Club 2000; 132 (2): A11-2.

Schulz K F, Chalmers I, Hayes R J, Altman D G. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. Jama 1995; 273 (5): 408-12.

Stern J M, Simes R J. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. Bmj 1997; 315 (7109): 640-5.

Thompson S G, Pocock S J. Can meta-analyses be trusted? Lancet 1991; 338 (8775): 1127-30.

Wright J G, Swiontkowski M F, Heckman J D. Introducing levels of evidence to the journal. J Bone Joint Surg (Am) 2003; 85 (1): 1-3.