# [ CLINICAL COMMENTARY ]

**J. TIMOTHY NOTEBOOM,** PT, PhD, SCS[1] • **STEPHEN C. ALLISON,** PT, PhD[2]
**JOSHUA A. CLELAND,** PT, PhD, OCS, FAAOMPT[3] • **JULIE M. WHITMAN,** PT, DSc, OCS, FAAOMPT[4]

# A Primer on Selected Aspects of Evidence-Based Practice Relating to Questions of Treatment, Part 2: Interpreting Results, Application to Clinical Practice, and Self-Evaluation

Critical appraisal using evidence-based practice (EBP) methods permits clinicians to make independent professional judgments about the validity, strength, and relevance of evidence. Independent judgments are necessary because the interpretations and conclusions of authors in published studies should not be accepted without close scrutiny by the reader. The EBP approach facilitates extraction of critical information from studies on treatment, including patient demographics and reported treatment effects. Because clinicians are attempting to apply results from current best evidence to clinical practice, a key question to be answered is, "Are the patients in this study similar to the patient I am managing?" Therefore, patient demographic data, such as age, diagnostic classification, level of impairment/dysfunction at baseline, and level of acuity, are just some of the characteristics that are typically reported in the methods or results section of the study. The clinician may even wish to determine if previously published prognostic studies have specific patient demographic data that may predict which patients are more likely to achieve a successful outcome regardless of what treatment is applied. If the clinician determines that patients from the study sufficiently resemble the patient of interest, then the clinician can proceed to a critical appraisal of the study design and results. The EBP approach identifies a finite set of key validity issues to consider and facilitates decisions about clinical meaningfulness of reported treatment effects. Critical appraisal enables a clinician to answer 3 questions[37] after the foreground question is posed and the best evidence is found: (1) Are the results valid?, (2) What are the results?, and (3) How can I apply the results to patient care? The first of these 3 questions was addressed in part 1 of this series. The remaining 2 questions will be addressed in this commentary.

• **SYNOPSIS:** The process of evidence-based practice (EBP) guides clinicians in the integration of individual clinical expertise, patient values and expectations, and the best available evidence. Becoming proficient with this process takes time and consistent practice, but should ultimately lead to improved patient outcomes. The EBP process entails 5 steps: (1) formulating an appropriate question, (2) performing an efficient literature search, (3) critically appraising the best available evidence, (4) applying the best evidence to clinical practice, and (5) assessing outcomes of care. This second commentary in a 2-part series will review principles relating to steps 3 through 5 of this 5-step model. The purpose of this commentary is to provide a perspective to assist clinicians in interpreting results, applying the evidence to patient care, and evaluating proficiency with EBP skills in studies of interventions for orthopaedic and sports physical therapy. *J Orthop Sports Phys Ther 2008;38(8):485-501. doi:10.2519/jospt.2008.2725*

• **KEY WORDS:** *critical appraisal, physical therapy, treatment effectiveness*

## STEP 3B. CRITICALLY APPRAISING THE LITERATURE: WHAT ARE THE RESULTS?

Readers should understand statistical analyses and the presentation of quantitative results when critically appraising an article.[26] While an extensive review of data analysis techniques is beyond the scope of this commentary, we will

[1]Associate Professor, Department of Physical Therapy, Regis University, Denver, CO. [2]Professor, Rocky Mountain University of Health Professions, Provo, UT; Associate Professor, Baylor University, Waco, TX. [3]Assistant Professor, Department of Physical Therapy, Franklin Pierce College, Concord, NH; Research Coordinator, Rehabilitation Services, Concord Hospital, Concord, NH. [4]Assistant Professor, Department of Physical Therapy, Regis University, Denver, CO; Faculty, Regis University Manual Therapy Fellowship, Regis University, Denver, CO. Address correspondence to Dr Tim Noteboom, Regis University, 3333 Regis Blvd, G-4, Denver, CO 80212. E-mail: noteboom@regis.edu

describe a number of statistical concepts and procedures commonly used in physical therapy literature. Bandy[6] conducted a 2-year review of the literature published in the journal *Physical Therapy* and identified 10 statistical procedures that were used in 80% of the articles reviewed. These were descriptive statistics, 1-way analysis of variance (ANOVA), *t* tests, factorial ANOVA, intraclass correlation, post hoc analyses, Pearson correlation, regression, chi-square, and nonparametric tests analogous to *t* tests.[6] In this commentary we will review some basic statistical concepts that we feel are important for readers performing critical appraisals. We will also discuss statistical methods used to identify between-group differences in clinical trials that use both continuous scale outcomes and dichotomous scale outcomes, with illustrations from orthopaedic and sports physical therapy literature.

## Reporting of Results in Treatment Studies

Results published in studies of physical therapy interventions typically include a summary of the findings from a wide variety of tests and measures that quantify the outcome variables selected by the authors to determine the effects of the intervention being studied. In some instances, such as with case reports or case series, raw data from each subject in the study may be presented. However, this approach is not realistic or warranted in studies with larger samples. More commonly, data are analyzed and reported as aggregated group results. Numerical indices are then used to describe attributes of the aggregated data. The mean or average is a measure that describes central tendency in a distribution of scores, and is most useful for variables that are on an interval or ratio scale.[69] If data exhibit outliers such that the value of the mean would be distorted, the median is often reported as the measure of central tendency. The median might also be preferred over the mean when sample sizes are so small that they may not represent the target population. For example, in a case series including 7 patients with hip osteoarthritis, MacDonald et al[57] reported medians rather than means for all baseline attribute variables and for all outcome variables. If data are from nominal or ordinal scales, the mode or median scores, respectively, are reported to describe central tendency.

A comparison of means is frequently used to make judgments about differences between different groups or across various time points in a study. However, means are incomplete descriptors of data because they give information only about central tendency. A more complete description of the data includes an indication of the variability in the distribution of scores (dispersion of the individual data points). The more variable the data, the more dispersed the scores will be. Among several available measures of variability, the SD is the statistic most frequently reported, together with the mean so that data are characterized according to both central tendency and variability.[69] Results are commonly reported as the mean ± SD. For example, Hall et al[42] compared headache index results at 4 weeks for their treatment group (31 ± 9) and their placebo group (51 ± 15), revealing a between-group mean difference of 20 points, with somewhat greater variance in the placebo group. If the median is used as the measure of central tendency, the range or interquartile range should be used to describe variability of the data, as the median may not always be the central value within the given range, especially when the data are nonparametric.

## Statistical Analyses Using Hypothesis Testing and *P* Values

Although descriptive statistics such as the mean and SD of a sample may be useful in comparing 2 different treatment groups or different time points for 1 group, such as pretreatment to posttreatment scores, clinicians also want to know whether observed sample differences represent true differences in the target population of patients. Therefore it is necessary to apply inferential statistical tests, such as the *t* test, ANOVA, or analysis of covariance (ANCOVA), to determine if between-group differences are statistically significant. These tests are examples of parametric tests, which are more robust tests in identifying significant differences in group means. However, there are assumptions that need to be met to apply parametric tests, which typically include normal distribution of data, equal variances across group data, and independence of data.[69] Alternately, when assumptions underlying parametric statistical tests are not met, nonparametric analogs of these tests should be used, although nonparametric tests are generally less powerful. For example, Hale et al[41] decided in their study on postural control in patients with chronic ankle instability to use Kruskal-Wallis tests and Mann-Whitney *U* tests instead of ANOVAs and *t* tests, because their outcomes data were not normally distributed.

The traditional approach for making the decision about statistical significance is hypothesis testing. Taking a comparison of means as one example, hypothesis testing attempts to determine with statistical methods whether differences between or among means are due to chance or are reflective of a true population difference in the target population. A central concept in hypothesis testing is the null hypothesis, one form of which states that there is no mean difference in the target population, thereby implying that any observed differences in sample means are due to chance. Therefore, if we reject the null hypothesis based on results of a statistical test, then we consider it unlikely that an observed difference is due to chance, and the difference is said to be statistically significant. However, statistical tests provide estimates of probability along a continuum, which is why researchers either express a specific threshold value or accept the default value (.05 or 5%) for statistical significance. This threshold probability is the alpha level, or $\alpha$, which indicates the maximum level of risk tolerance for falsely rejecting the null hypothesis (a type I error).[69] The alpha level, sometimes expressed as the "level of significance," is established by the researcher prior to data collection.

When the alpha level is .05, *P* values less than .05 permit rejection of the null hypothesis, leading us to infer that true mean differences exist in the target population. When *P* values are greater than .05, we conclude that the risk for committing a type I error exceeds our predetermined threshold (the alpha level). Therefore, when *P* is greater than .05, we do not consider observed differences to be statistically significant and conclude that such differences between groups may be due to chance. However, the set point for alpha is somewhat arbitrary and the *P* values can be influenced by sample size. Therefore, while researchers may set a specific alpha to accept or reject the null hypothesis, the savvy reader should still examine the results, confidence intervals (CIs), and sample size to determine whether or not a *P* value greater than .05 may be potentially meaningful.

For example, in a recent clinical trial[67] comparing 2 types of exercises for increasing strength in women with chronic neck pain, both groups increased strength from pretreatment to posttreatment ($P<.01$). In other words, within-group improvements were significant. However, comparisons of improvements between the 2 groups ("between-group differences") were not significantly different ($P = .97$). Based on the 2 *P* values for within-group and between-group differences, we can reject the null hypothesis for within-group improvements in the target population and conclude that each treatment group achieved statistically significant gains in strength from pretreatment to posttreatment assessment points. In contrast, we must conclude that the observed between-group difference in improvement was due to chance, attributable only to sampling error, and does not reflect a true difference in effectiveness of the exercise programs in the target population.

## Confidence Intervals

Confidence interval analysis is an essential skill for the evidence-based practitioner and will comprise an important part of almost every critical appraisal of evidence. Montori[62] and others[3,76] have argued that because *P* values are not helpful in providing clinicians with information about the magnitude of the treatment effect, other statistics should be used. In contrast to *P* values, CIs provide information on the magnitude of the treatment effect in a form that pertains directly to the process of deciding whether to administer a therapy to patients. Whereas a sample statistic is only a point estimate of the true population value, the CI is a range of values within which the population value is likely to be found at a given level of confidence.[35] Sim and Reid[76] have reported that because CIs focus attention on the magnitude and the probability of a treatment effect, they thereby assist in determining the clinical usefulness and importance (as well as the statistical significance) of the findings.[76] Most often the 95% CI is used. This is commonly interpreted to represent the range of values within which we can be 95% certain that the true population value actually lies.[3]

For example, Gerber et al[34] reported the mean visual analog scale (VAS) score for knee pain after 15 weeks of postoperative exercise training for the experimental treatment group: 0.77 cm (95% CI: 0.19 to 1.35 cm). At a 95% level of confidence we conclude that the true posttreatment population mean pain value for patients receiving this type of exercise training is no less than the lower limit of the CI (0.19 cm) and no greater than the upper limit of the CI (1.35 cm). Readers should note that not all values within the CI are considered equally likely to be the true population value. The point estimate from the sample (0.77 cm) is considered the single best estimate of the population parameter, with values becoming increasingly less likely when approaching either limit of the CI.[33] The convention of using a 95% CI is arbitrary, similar to setting the alpha level to .05.

The level of precision or imprecision expressed by CI width is affected by the sample size and the variance in the distribution of scores. Small sample sizes and greater variance result in wider CIs.[73] Wide CIs reflect imprecision in the data and uncertainty associated with the magnitude of the treatment effect.[33,44] In contrast, the narrower the width of the CI around the point estimate of the treatment effect, the more confident one can be that the true effect and its point estimate are similar, allowing the clinician to make more confident decisions from the data.

Although journals are increasingly requiring authors to report CIs, readers will often find published evidence with no CIs around the point estimates of treatment effects. Even when authors do report CIs they commonly fail to interpret them.[29] Readers performing critical appraisals of evidence can often compute CIs themselves given published details. A helpful and easy-to-use spreadsheet for computation of CIs (PEDro Confidence Interval Calculator) is freely downloadable from the PEDro website.[1] As an illustration, we can extract means, sample sizes, and SDs from a recent randomized controlled trial (RCT)[21] wherein authors found significantly better improvements ($P = .009$) in an experimental treatment group compared to a control group. Pretreatment to posttreatment improvements in shoulder internal rotation were $20° \pm 12.9°$ in the experimental group (n = 15) compared to $5.9° \pm 9.4°$ in the control group (n = 24). Although the authors did not report a 95% CI around the between-group difference, we can easily compute it using the PEDro Confidence Interval Calculator.[1] **FIGURE 1** shows results for this computation. From these results we see that the point estimate for the difference between mean group improvements was 14.1° in favor of the treatment group. The 95% CI does not include a zero difference, which is compatible with the statistically significant result ($P = .009$). Furthermore, we estimate the true population difference for mean improvement to be no less than 6.9° and no more than 21.3° favoring the treatment.

## Results for Continuous Scale Outcomes: Differences Between Means

If randomization in a RCT was effective in creating reasonably equivalent groups at baseline, the pretreatment group means

**To estimate a confidence interval for the difference between two means**

| | |
|---|---|
| Enter the mean of the control group here: | 5.9 |
| Enter the estimated population standard deviation for the control group here: | 9.4 |
| Enter the sample size (eg, number of subjects) for the control group here: | 24 |
| Enter the mean of the experimental group here: | 20 |
| Enter the estimated population standard deviation for the experimental group here: | 12.9 |
| Enter the sample size (eg, number of subjects) for the experimental group here: | 15 |
| Enter the required confidence interval (eg, 95%) here: | 95 |

**Result**

| | |
|---|---|
| The estimated difference between the two population means is: | –14.1 |
| The estimated CI is: | –21.34 to –6.86 |

**FIGURE 1.** Results from the PEDro Confidence Interval Calculator* for computation of a 95% confidence interval (CI) around a difference between 2 group means. Note that the sign of the difference and signs on upper and lower limits of the CI are arbitrary; differences and confidence limits must be interpreted in light of the nature of the scales and the relative outcomes between groups. From Physiotherapy Evidence Database (PEDro). Available at: http://www.pedro.fhs.usyd.edu.au/index.html. Accessed July 11, 2008.

for outcomes on continuous scales will be close to equal. Therefore, when group means are not meaningfully different at baseline, the magnitude of the between-group treatment effect, when statistically significant, can be most easily conceptualized as the posttreatment difference between group means for these outcome scales. However, clinicians should critically assess the within-group variability because variance that is much different between groups could be somewhat misleading. In cases where groups are not equivalent at baseline for important prognostic factors, ANCOVA methods can statistically adjust the posttreatment means to account for baseline differences.[69] For example, Rydeard et al,[72] found in a recent RCT that mean scores for the functional disability outcome were significantly different between groups at baseline in spite of randomization. Therefore, they used baseline functional disability outcome scores as a covariate in the statistical analyses, then found that the between-group difference in posttreatment means for functional disability, as adjusted by the ANCOVA method, was statistically significant between groups.

If the treatment under consideration (the experimental treatment) is more effective than the comparison (no treatment, placebo, or a competing treatment), the posttreatment experimental group mean will show greater improvement than the comparison group mean(s). For a scale on which a higher score is a better outcome (eg, muscle strength), the experimental group posttreatment mean will be greater than the comparison group mean if the treatment is effective. For a scale on which a lower score is a better outcome (eg, VAS for pain), the experimental group posttreatment mean will be less than the comparison group mean if the treatment is effective. The magnitude of this posttreatment between-group difference is a measure of the treatment effect and is sometimes called the raw effect size.[22] Computation of the raw between-group effect size is the simple subtraction of one group mean from another and is expressed in the relevant units of the outcome scale. Therefore, this point estimate of the raw effect size is conceptually intuitive and is crucial for deciding whether the magnitude of a statistically significant treatment effect is clinically meaningful. For example, Butcher et al[13] reported vertical jump takeoff velocity in a control group (2.29 ± 0.35 m/s) and in a trunk stability training group (2.38 ± 0.39 m/s) after 3 weeks of exercise, and found this difference to be statistically significant ($P<.05$). The raw between-group effect size is, therefore, 0.09 m/s (2.38 – 2.29 m/s). Knowing this value, the clinician can proceed to determine the clinical relevance of the treatment effect.

In contrast to using raw posttreatment scores to calculate the between-group effect size, authors will sometimes use change scores to represent average improvements over time by computing the difference between baseline, or pretreatment, means and posttreatment means. Between-group differences in average change scores are then computed to represent the magnitude of the between-group treatment effect. This approach was used by Johnson et al[49] when they reported that the improvement from baseline to posttreatment in shoulder external rotation in the experimental treatment group (31.3° ± 7.4°) was significantly better ($P<.001$) than that in the comparison treatment group (3.0° ± 10.8°).

Raw effect sizes are commonly transformed into unitless effect size indices, such as *d* for the *t* test and *f* for ANOVA, which are examples of standardized effect size indices.[22] The most common approach in rehabilitation research is to divide the raw effect size by the combined (pooled) SDs. This method has the benefit of accounting for both the magnitude of the treatment effect and the variability of the group means. For example, using values from the between-group comparison in the Butcher et al study[13] reported above, the raw effect size was .09 m/s, whereas the effect size index (*d*) was 0.24 (0.09 m/s divided by the pooled SD of 0.37 m/s). Effect size indices provide a general indication for relative magnitudes of treatment effects. For example, Cohen[22] characterized effect size indices for a comparison of 2 means as follows: 0.2, small; 0.5, medium; 0.8, large. Although unitless effect size indices are helpful for comparing the magnitude of effect sizes among studies using different outcomes measures, these transformed indices of treatment effect are not as intuitive or as helpful as raw effect sizes for making the crucial comparisons that allow clinicians to judge whether treatment effects exceed thresholds for clinical meaningfulness, as discussed below. However, if variance is much different between or among groups, raw effect sizes may be misleading. In addition, effect size indices can be useful for comparing treatment effects across more than one experiment. For these reasons, readers may wish to consider both raw effect sizes and the standardized effect size indices when critically appraising evidence for therapy.

### The Minimal Detectable Change and Minimal Clinically Important Difference Properties

Decisions about clinical meaningfulness of results involve judgments about thresholds distinguishing trivial effects from clinically important effects. Although any such judgment can be subject to debate and will depend on multiple contextual considerations and local circumstances, these judgments are essential in any critical appraisal of evidence. Because clinicians are frequently interested in identifying the amount of change over time, measurement properties such as minimal detectable change (MDC) are important to consider. Similar to other measures of reliability, such as standard error of measurement (SEM), the MDC is the smallest real difference, which represents the smallest change in score that likely reflects true change rather than measurement change alone.[74,77] For example, Stratford and colleagues[78] have reported that the Roland-Morris Questionnaire, a commonly used outcome measure for patients with low back pain, has an MDC of 4 points. Therefore, to be confident that 2 scores taken across time represent a true change the scores would need to be more than 4 points from each other. However, MDC only provides an indication of the minimum change that is detectable by the instrument, and not necessarily the amount of change that could be considered clinically meaningful to the patient. Jaeschke et al[46] defined the minimal clinically important difference (MCID) as "the smallest difference in score in the domain of interest which patients perceive as beneficial." There is a growing body of literature outlining methods for determining MCID values,[8,46] reporting MCIDs for specific scales, and using MCIDs to make judgments about clinical meaningfulness of treatment effects in clinical trials. Although published MCID values must be considered in the context of the varying methods and intended purposes for their derivations or estimations,[8] clinicians unfamiliar with specific scales will often find it helpful to be aware of published MCID values when critically appraising evidence. No single published value for a MCID can be applied uncritically in all circumstances or for all purposes.[59] Rather, a published MCID can provide an initial reference point when applying personal clinical expertise to make independent judgments about what distinguishes trivial from clinically important treatment effects in a local context. An illustrative patient scenario integrating patient values with published MCIDs to make patient-relevant judgments in a critical appraisal is given below in the section titled "Step 4. Incorporating Evidence Into Clinical Practice."

Published MCID values for selected outcome scales commonly used in orthopaedic and sports physical therapy are displayed in **TABLE 1**.

Although the definition of the MCID above suggests application to an individual patient, MCID values are commonly employed to make judgments about the clinical meaningfulness of averaged group treatment effects, both for within-group effects[25,53] and for between-group effects.[20,24,63] Indeed, Jaeschke et al[46] explicitly anticipated use of the MCID to make judgments both for individual and group differences. If the observed raw effect size is equal to or greater than the MCID, the treatment effect is considered clinically meaningful. Otherwise, the treatment effect is deemed trivial regardless of whether statistical significance is achieved. For example, Hyland et al[45] found in a RCT that the posttreatment pain VAS outcome in a calcaneal taping group ($2.7 \pm 1.8$) was significantly better ($P<.001$) than that of the control group ($6.2 \pm 1.0$). Inasmuch as the point estimate for the treatment effect was a posttreatment between-group difference of 3.5 cm favoring calcaneal taping, we can compare this value to a MCID for the pain VAS. If we accept a suggestion of 3.0 cm as a reasonable value for the MCID for the pain VAS,[55] we consider the treatment effect in the study sample to be a clinically meaningful benefit because the point estimate of the effect (3.5 cm) is greater than the MCID (3.0 cm).

If a reader is not sufficiently familiar with an outcome scale to make an intuitive judgment about clinical meaningfulness of a treatment effect size, and if no published MCID can be found for that outcome scale, it is often helpful to convert the effect size to a percent difference. Following the example from Hyland et al[45] above, the percent difference between groups in posttreatment pain VAS (10-cm scale) means was calculated as follows: $(6.2 - 2.7) \div 6.2 = 57\%$. Therefore, the mean pain VAS score for the treatment group was 57% lower (better) than that of the control group. Most clinicians would judge a 57% average reduction in pain to be clinically meaningful, even without being familiar with a particular pain outcome scale.

### Interpretations of Apparently Positive Trials: MCID, Effect Size, and CI Limits

A clinical trial is termed "positive" when the null hypothesis is rejected by formal hypothesis testing. In a positive trial, authors conclude that results are statistically significant and that the experimental treatment is more effective than the comparison. Guyatt et al[38] use the phrase "apparently positive trial" to communicate the idea that critical appraisal requires an evidence-based practitioner to look beyond statistical significance. Additional judgments must be made about clinical meaningfulness of the treatment effect

TABLE 1 Published Values for Minimal Clinically Important Differences (MCIDs) on Select Outcome Scales

| Outcome Scale | Suggested MCID* | Clinical Context | Published Study |
|---|---|---|---|
| 6-minute walk test | 54 m | Patients with chronic obstructive pulmonary disease | Wise and Brown, 2005[86] |
| 10-cm pain visual analog scale | 3.0 cm | Emergency room patients with acute pain | Lee et al, 2003[55] |
| 11-point numeric pain rating scale | 2 | Patients with chronic pain | Farrar et al, 2001[28] |
| American Shoulder and Elbow Surgeons Standardized Shoulder Form, patient self-report section | 6.4 | Patients with musculoskeletal shoulder pathologies | Michener et al, 2002[60] |
| Functional rating index | 9 | Patients with low back pain | Childs et al, 2005[15] |
| Gait speed | 0.10 m/s | Patients recovering from hip fracture | Palombaro et al, 2006[68] |
| General function score | 12 | Patients with chronic low back pain | Hagg et al, 2003[40] |
| Lower Extremity Functional Scale | 9 | Patients with lower extremity musculoskeletal dysfunction | Binkley et al, 1999[11] |
| Modified Low Back Pain Disability Questionnaire | 6 | Patients with low back pain | Fritz and Irrgang, 2001[31] |
| Neck Disability Index | 7.0 | Patients with cervical radiculopathy | Cleland et al, 2006[19] |
| Neck Disability Index | 5.0 | Physical therapy outpatients with musculoskeletal neck pain | Stratford et al, 1999[80] |
| Oswestry Disability Index | 10 | Patients with chronic low back pain | Hagg et al, 2003[40] |
| Patient-Specific Functional Scale | 2.0 | Patients with cervical radiculopathy | Cleland et al, 2006[19] |
| Quebec Back Pain Disability Scale | 15 | Patients with low back pain | Fritz and Irrgang, 2001[31] |
| Roland-Morris Back Pain Questionnaire | 2 (baseline, 0-8); 4 (baseline, 5-12); 5 (baseline, 9-16); 8 (baseline, 13-20); 8 (baseline, 17-24) | Patients with low back pain (duration, <6 wk) | Stratford et al, 1998[79] |
| SF-36 bodily pain subscale | 7.8 | Patients with hip or knee osteoarthritis | Angst et al, 2001[4] |
| SF-36 physical function subscale | 3.3 | Patients with hip or knee osteoarthritis | Angst et al, 2001[4] |
| SF-36 physical component summary | 2.0 | Patients with hip or knee osteoarthritis | Angst et al, 2001[4] |
| Simple shoulder test | 10 | Patients undergoing physical therapy treatment for shoulder pain of musculoskeletal, neurogenic, or undetermined origin | Michener & McClure, 2002[60] |
| Visual analogue scale (VAS) of back pain | 18 | Patients with chronic low back pain | Hagg et al, 2003[40] |
| Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) | 20% | Patients with hip or knee osteoarthritis | Barr et al, 1994[7] |
| Zung Depression Scale | 8 | Patients with chronic low back pain | Hagg et al, 2003[40] |

** Units are scale points unless otherwise indicated.*

and the level of precision in the point estimate of the effect size. These judgments are accomplished by comparing the raw effect size, with its accompanying CI, to the MCID. Even when we conclude that results are clinically meaningful because the point estimate for the raw effect size is greater than the MCID, we must recognize that the true size of the treatment effect may be more or less than the point estimate from sample data. The upper and lower limits of the 95% CI around that point estimate for the effect size give us an indication of just how small or how large the true treatment effect might be in the population of interest. Therefore, we consider the 95% CI to determine whether the MCID is within that interval. If the MCID is within the 95% CI, then we cannot rule out at a 95% level of confidence that the true population treatment effect might be trivial (less than MCID). On the other hand, if the raw effect size is greater than the MCID and the MCID is excluded from the 95% CI, then we are 95% confident that there is a clinically meaningful benefit of treatment in the population—even if the true magnitude of that benefit is at the limit of the CI suggesting the smallest benefit of treatment. Guyatt et al[38] characterize a positive trial in which the 95% CI excludes the MCID as "a definitive trial."

Following the example above from Hyland et al,[45] we can consider the raw point estimate of the treatment effect (3.5 cm on the pain VAS) in the context of its 95% CI and the MCID (3.0 cm). This study had a small sample of subjects in the 2 groups considered here: 10 patients in the control group and 11 patients in the calcaneal taping group. Entering those sample sizes and the posttreatment pain VAS means and SDs for the 2 groups into the PEDro Confidence Interval Calculator spreadsheet (**FIGURE 1**), we find that the 95% CI is 2.2 to 4.9. We conclude from this CI that the true treatment effect size in the target population is no less than 2.2 cm on the pain VAS, and no greater than 4.9 cm. Inasmuch as the MCID (3.0 cm) is not excluded by the 95% CI, we cannot rule out a trivial treatment effect in the target population. This is because the study results are compatible with true treatment effects as small as 2.3, 2.5, or 2.7 (etc), which are all smaller than the MCID and are therefore not clinically meaningful. This analysis does not change the fact that a statistically significant treatment effect was found favoring the experimental treatment, nor does it change the fact that the best estimate[33] of the population treatment effect (3.5 cm) is clinically meaningful. Rather, this illustration demonstrates the imprecision inherent in studies with small sample sizes and suggests that additional evidence with larger samples and correspondingly greater precision (less variability) is required before we consider this finding definitive.[38]

Adequate precision to rule out a trivial treatment effect in a positive trial is illustrated in a study of radial shock wave therapy for calcific tendinitis of the shoulder.[14] Posttreatment pain VAS scores (mean ± SD) were significantly better ($P$ = .004) in the treatment group (0.90 ± 0.99) than in the control group (5.85 ± 2.23). The between-group difference was 4.96 cm (95% CI: 4.23 to 5.67). If we accept the MCID value of 3.0 cm for the pain VAS,[55] we consider this study to be convincing evidence for a clinically meaningful benefit of treatment, inasmuch as the study results suggest an average treatment effect no less than 4.23 cm in the target population. In other words, the trial is definitive for this outcome, because the 95% CI around the point estimate for the treatment effect excludes the MCID.

### Interpretations of Apparently Negative Trials: MCID, Effect Size, and CI Limits

A clinical trial is termed "negative" when we fail to reject the null hypothesis. In a negative trial, authors conclude that results are not statistically significant and that the experimental treatment is no more effective than the comparison. Guyatt et al[38] use the phrase "apparently negative trial" to communicate the idea that critical appraisal requires an evidence-based practitioner to be wary of results from negative trials unless adequate statistical power can be demonstrated. The danger is that an underpowered trial might fail to find statistical significance in sample data even when there is a meaningful benefit of treatment in the target population (a type II error). Authors will frequently attempt to address this issue by revealing details of the statistical power analysis used to estimate the required sample size before the study was conducted. This approach is unsatisfying in part because a priori power computations require estimations of variance that may or may not reflect the observed variance in sample data. Guyatt et al[38] suggest a different method for determining whether a negative trial has sufficient statistical power. Here again we consider the 95% CI around the point estimate of the raw effect size, to determine whether the MCID is within that interval. If the MCID is within the 95% CI, then we cannot rule out at a 95% level of confidence that the true population treatment effect might be clinically meaningful (greater than MCID), even though the authors failed to reject the null hypothesis.[50] This circumstance would reveal inadequate statistical power in the study, suggesting that we should not accept any conclusion that the treatment is ineffective. On the other hand, if the MCID is excluded from the 95% CI, then we are 95% confident that there is no clinically meaningful benefit of treatment in the population, even if the true magnitude of the treatment effect is at the limit of the CI suggesting the largest between-group difference. Guyatt et al[38] characterize a negative trial in which the 95% CI excludes the MCID as "definitely negative." A reader critically appraising a negative trial in which the 95% CI around the treatment effect excludes the MCID can be confident that the failure to find a statistically significant difference is not attributable to a type II error. In other words, if precision in the study is sufficient for the 95% CI to exclude the MCID, the study has adequate statistical power to detect a clinically meaningful difference if one exists in the target population.

Authors in a recent RCT[23] found no statistically significant difference ($P$ = .33) for knee flexion range of motion outcomes among 3 groups: a control group receiving no time on a continuous passive motion (CPM) machine, a treatment group receiving CPM treatments of 35 minutes duration once daily, and another treatment group receiving CPM treatments of 2 hours duration once daily. The authors considered 10° to be the MCID for this outcome. **FIGURE 2** provides a graphical display of 95% CIs for each of the 3 between-group comparisons at the time of discharge from hospital. The 2 dotted vertical lines represent MCIDs of 10° favoring either of the 2 groups for each plotted comparison. The solid vertical line represents the null value (0°) for the between-group differences. The 95% CIs around each of the between-group effect sizes are represented by horizontal lines with vertical anchors at each end, reflecting upper and lower limits of the CIs. Each 95% CI includes the null value, suggesting no statistically significant differences—a finding consistent with results from the traditional hypothesis test ($P$ = .33). However, only 1 of the 3 95% CIs excludes the MCID. Therefore, statistical power in this study was adequate to rule out a clinically meaningful treatment effect in the target population for 1 between-group comparison (CTL-EXP1); but the study power was insufficient to

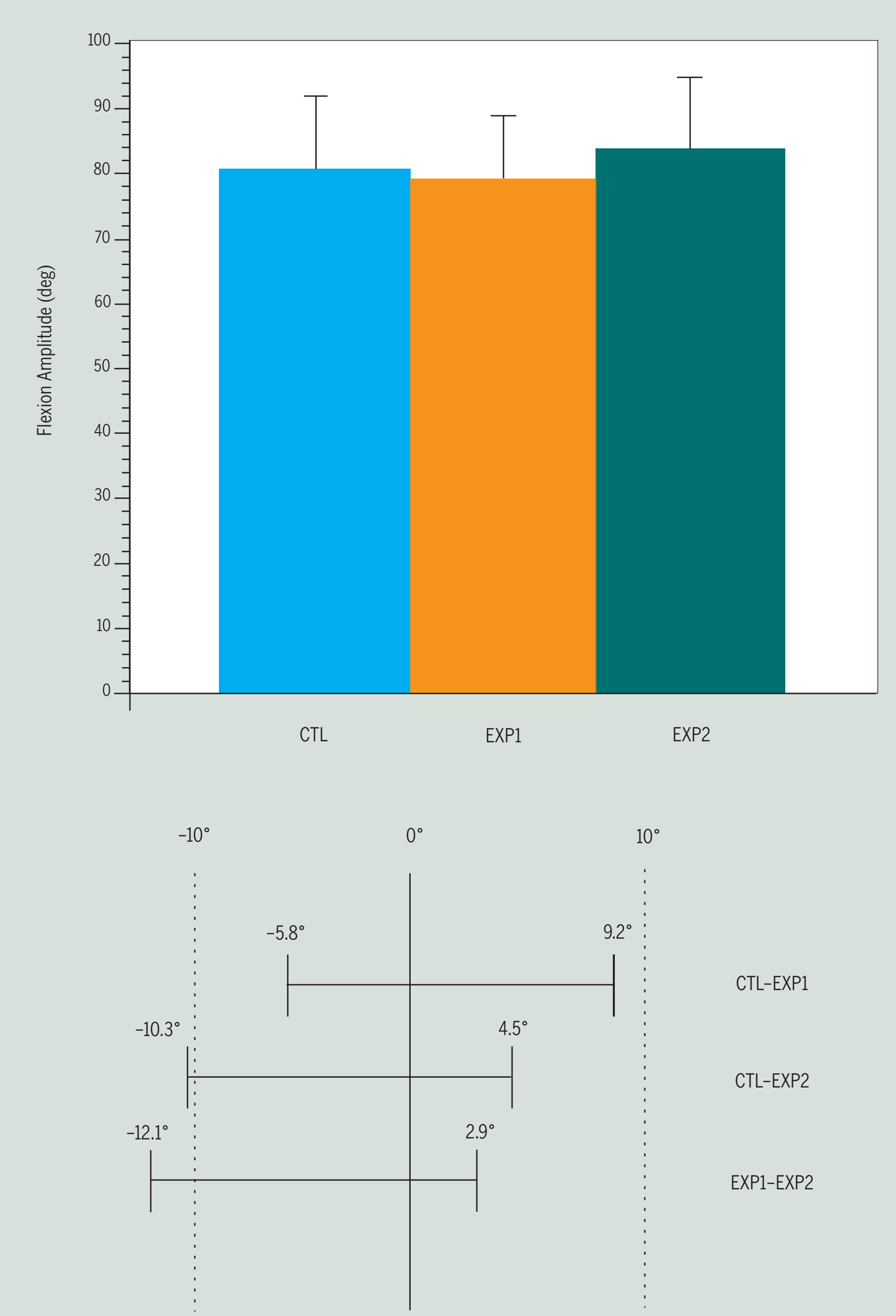


**FIGURE 2.** Results from a negative trial showing 95% confidence intervals in relation to the minimal clinically important difference. From Denis M, Moffet H, Caron F, Ouellet D, Paquet J, Nolet L. Effectiveness of continuous passive motion and conventional physical therapy after total knee arthroplasty: a randomized clinical trial. *Phys Ther.* 2006 Feb;86(2):174-85. Modified with permission.

rule out a small but potentially meaningful difference for 2 of the 3 between-group comparisons.

Adequate precision and statistical power are illustrated in a negative trial comparing arthroscopy to placebo arthroscopy in patients with knee osteoarthritis.[63] Authors determined the MCID for the pain subscale of the Arthritis Impact Measurement Scales (AIMS) to be 10 points. At the 6-week follow-up measurement, the average pain score for the arthroscopy with debridement group (49.9 ± 23.3) was not significantly different ($P$ = .85) from the placebo group (50.8 ± 23.2). The difference between means was 0.9 (95% CI: –7.7 to 9.4). Therefore, the largest treatment effect favoring arthroscopy in the target population consistent with results from this study would be 9.4 points on the AIMS pain subscale: a trivial difference. Given that the MCID was excluded from the 95% CI, we conclude that the study had adequate precision and sufficient statistical power to have found a clinically meaningful difference, if one existed, in the target population. This interpretation is the same as that expressed by the authors: "If the 95 percent confidence interval around the estimated size of the effect does not include the minimal important difference, one can reject the hypothesis that the arthroscopic procedures have a small but clinically important benefit."[63]

### Results for Dichotomous Outcomes: Risk Reduction and Number Needed to Treat

Although authors of clinical trials in physical therapy most often select continuous outcome variables, there are many important naturally dichotomous outcomes that should be included in studies of orthopaedic and sports physical therapy. Dichotomous outcomes are those that patients either experience or do not experience. Examples are recurrent dislocations, failure to achieve complete recovery, failure to return to competition, recurrence of low back pain, receiving injections, and subsequent surgery. Because the statistical methods for analyzing dichotomous outcomes quantify reduction in risk, dichotomous outcomes are usually operationalized as negative outcomes (numbers of patients who did have a recurrent dislocation, patients who were not able to return to sport, etc). Important continuous scale outcomes can be dichotomized using the MCID to report numbers of patients who achieve or fail to achieve clinically meaningful improvements in motion, strength, pain reduction, etc.[65] For ex-

| TABLE 2 | Examples of Number Needed to Treat (NNT) Values for Various Physical Therapy-Related Interventions* | |
|---|---|---|
| **Clinical Question** | **NNT** | **95% Confidence Interval** |
| How effective is early cardiac rehabilitation on health-related quality-of-life score in patients experiencing a cardiovascular incident? (Comparison treatment: usual care)[66] | 6 | 3 to 21 |
| How effective is vitamin D supplementation in preventing falls in ambulatory or institutionalized older adults? (Comparison treatment: calcium or placebo)[12] | 15 | 8 to 53 |
| How effective is a multidisciplinary intensive diabetes education program on improving glycemic control or decreasing diabetes-related distress in patients with diabetes? (Comparison group: standard care)[52] | 1.8 | 1.5 to 2.4 |
| How effective is adding 3 stretching sessions to a typical weekly infantry training program on reducing incidence of overuse injury in military basic trainees? (Comparison treatment: typical infantry training)[43] | 8 | 4.6 to 33.9 |
| How effective is range-of-motion exercise and joint mobilization on improving wrist extension following Colles fracture? (Comparison treatment: home exercise program)[83] | 2.3 | 2 to 17 |
| How effective is combined cervico-thoracic manipulation and exercise therapy in reducing headache frequency in patients with persistent headache? (Comparison treatment: self-care instruction)[51] | 1.9 | 1 to 3 |
| How effective is a stabilization program in decreasing pain and disability in patients with low back pain who are categorized as being hypermobile in the lumbar spine? (Comparison treatment: lumbopelvic manipulation)[32] | 1.6 | 1.2 to 10.2 |
| How effective is combined manual physical therapy, exercise, and unloaded treadmill walking on perceived recovery for patients with lumbar spinal stenosis? (Comparison treatment: flexion exercises and treadmill walking program)[84] | 2.6 | 1.8 to 7.8 |
| How effective is combined manual physical therapy and exercise for avoiding total knee arthroplasty surgery up to 1 year posttreatment? (Comparison treatment: placebo ultrasound)[25] | 7 | 4 to 105 |

** Results are presented without regard for levels of evidence or the extent to which validity threats were protected in the referenced studies. These factors varied widely among studies cited.*

ample, Clegg et al[18] dichotomized their primary outcome: the WOMAC pain subscale with raw scores ranging from 0 to 500. Authors dichotomized this scale by reporting percents of patients in each study group who achieved at least 20% improvement after treatment. This cut score is the MCID recommended by developers of the WOMAC.[7] Results for dichotomous outcomes can be reported as odds ratios[61] but are frequently reported as absolute risk reduction (ARR), relative risk reduction (RRR), and number needed to treat (NNT).[65]

Deyle et al[25] reported the number of patients who had knee replacement surgery by the time of a 1-year follow up in each of 2 groups with knee osteoarthritis. In the placebo group, 8 of 41 patients (20%) had surgery compared to only 2 of 42 patients (5%) receiving manual therapy and exercise. The ARR is the difference between these 2 proportions: 20% – 5% = 15% (95% CI: 1% to 28%). The RRR is the reduction in risk relative to that in the comparison group: (20% – 5%) ÷ 20% = 75% (95% CI: 5% to 100%). The NNT is computed by taking the reciprocal of the ARR: 1.0 ÷ 0.15 = 7 (95% CI: 4 to 105). Reporting results in this way reveals that, although the risk of needing surgery within 1 year was 20% in the placebo group, risk was reduced by 15% in absolute terms and by 75% in relative terms by providing manual therapy and exercise. The wide 95% CIs around the point estimates reveal considerable imprecision in the results. The principles discussed above for appraising "apparently" positive and negative trials apply equally to assessing dichotomous outcomes. However, rather than comparing the MCID to point estimates and associated CIs for mean effect sizes, a clinical judgment is required (depending on multiple considerations of context) to determine the minimal clinically important amount of risk reduction for comparison with the point estimates and associated CIs for ARR, RRR, and NNT. For example, if a clinician considers a 5% RRR for needing total knee arthroplasty to be clinically meaningful, the evidence from Deyle et al[25] would be considered "definitive." On the other hand, if a clinician judges a 30% RRR to be minimally clinically meaningful, the point estimate from Deyle et al[25] (75% RRR) would be considered promising; but the wide CI around that treatment effect would lead the clinician to seek additional evidence, perhaps from a larger trial with greater precision.

The NNT is defined as the number of patients who would need to be treated on average to prevent 1 bad outcome or achieve 1 desirable outcome in a given period of time.[54] Therefore, when a low NNT is associated with a treatment, this indicates that relatively few patients need to receive this treatment in order to avoid 1 bad outcome. Therefore, NNT values are used as a measure of treatment effectiveness and are helpful in cost-benefit calculations. However, it should be noted that NNT values alone are not sufficient to determine if an intervention approach should be implemented. Patient values and preferences, the severity of the out-

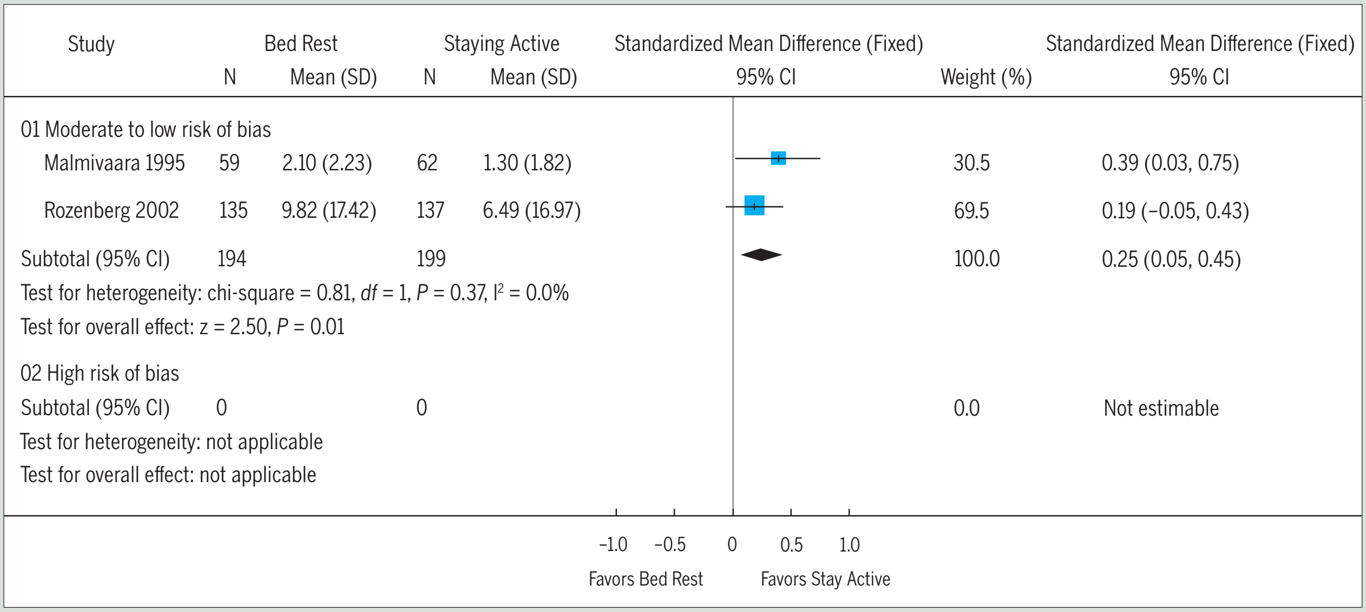


**FIGURE 3.** Forest plot demonstrating presentation of results from a meta-analysis in a systematic review (Hagen, 2004). From Hagen KB, Hilde G, Jamtvedt G, Winnem M. Bed rest for acute low-back pain and sciatica. Cochrane Database of Systematic Reviews. 2004; Issue 4. Art. No.: CD001254. Reproduced with permission. Copyright © 2004 John Wiley & Sons Ltd.

come that would be avoided, and the cost and side effects of the intervention are important determinants that should be considered when making treatment decisions. Thus, the threshold NNT will almost certainly be different for different patients and there is no simple answer to the question of when an NNT is sufficiently low to justify a treatment. **TABLE 2** lists several physical therapy-related interventions with associated outcomes, NNTs, and 95% CI values.

### Synthesized Results From Multiple Clinical Trials: Systematic Reviews

In spite of language used above to characterize results from a qualifying clinical trial as "definitive," a single trial will rarely provide final or completely conclusive evidence for treatment effectiveness. This is why multiple RCTs with consistent results provide stronger evidence than a single RCT. Consequently, a single systematic review with homogeneity of results from multiple RCTs provides a higher level of evidence (level 1a) than a single RCT with good protections against validity threats (level 1b).[2]

Systematic reviews are at the top of the evidence hierarchy because they typically use meta-analysis methods, when appropriate, to synthesize evidence from multiple single clinical trials.[38] In this way, results from the overall body of best evidence, filtered and selected by explicit methodological quality criteria, are synthesized to provide an overall estimate of treatment effectiveness. Meta-analysis methods allow pooling of sample sizes from among included studies, resulting in substantial advantages: (1) increased statistical ability to detect significant treatment effects, and (2) enhanced precision in estimates of effect sizes, reflected in narrower CIs around point estimates.

Results of meta-analyses are typically presented in forest plots. A forest plot representing the simplest case from a meta-analysis based on only 2 original studies is shown in **FIGURE 3**. Note that results from individual trials are represented by point estimates (squares in this example), with horizontal lines representing the CIs. Effect sizes for continuous scale outcomes in a meta-analysis are transformed to a normalized scale, such as a weighted mean difference (WMD) or a standardized mean difference. For dichotomous outcomes, effect sizes in a meta-analysis are typically reported as relative risk or odds ratios. The null value for the treatment effect is represented as a central vertical line in a forest plot. Point estimates for effect sizes plotted on one side of the vertical reference line favor the experimental treatment; points plotted on the other side of the line favor the comparison. If the CI around the point estimate crosses the vertical line, results are not statistically significant because those results are consistent with a zero treatment effect in the target population. **FIGURE 3** illustrates results from a systematic review[39] for an outcome of pain intensity at 12 weeks, comparing results obtained in patients treated with bed rest compared to patients who stayed active. Two RCTs were included in the meta-analysis: one with a statistically significant effect favoring the recommendation to stay active and one with no statistically significant difference between groups. Without meta-analysis the overall accumulation of evidence might appear to be equivocal, with one study suggesting benefit and another suggesting no benefit. The synthesized result pooling data in a meta-analysis from both studies is represented by the diamond shape labeled "subtotal" in **FIGURE 3**. This meta-analysis result from

**TABLE 3** — GRADING SCHEME FOR TREATMENT RECOMMENDATIONS IN A CLINICAL PRACTICE GUIDELINE

| Grade of Recommendation* | Clarity of Risk/Benefit | Methodological Strength of Support Evidence | Implications |
|---|---|---|---|
| 1A | Clear | RCTs without important limitations | Strong recommendation; can apply to most patients in most circumstances without reservation |
| 1B | Clear | RCTs with important limitations (inconsistent results, methodological flaws†) | Strong recommendation, likely to apply to most patients |
| 1C+ | Clear | No RCTs directly addressing the question, but results from closely related RCTs can be unequivocally extrapolated, or evidence from observational studies may be overwhelming | Strong recommendation; can apply to most patients in most circumstances |
| 1C | Clear | Observational studies | Intermediate-strength recommendation; may change when stronger evidence is available |
| 2A | Unclear | RCTs without important limitations | Intermediate-strength recommendation; best action may differ depending on circumstances or patients' or societal values |
| 2B | Unclear | RCTs with important limitations (inconsistent results, methodological flaws) | Weak recommendation; alternative approaches likely to be better for some patients under some circumstances |
| 2C | Unclear | Observational studies | Very weak recommendations; other alternatives may be equally reasonable |

*Abbreviation: RCT, randomized controlled trial.*
** Since grade B and C studies are flawed, it is likely that most recommendations in these classes will be level 2. The following considerations will bear on whether the recommendation is grade 1 or 2: the magnitude and precision of the treatment effect, patients' risk of the target even being prevented, the nature of the benefit and the magnitude of the risk associated with treatment, variability in patient preferences, variability in regional resource availability and health care delivery practices, and cost considerations. Inevitably, weighing these considerations involves subjective judgment.*
*† These situations include RCTs with both lack of blinding and subjective outcomes where the risk of bias in measurement of outcomes is high, RCTs with large loss to follow-up.*
*Adapted with permission from Guyatt G, Hayward R, Richardson WS, et al. Moving from evidence to action. In Guyatt G, Rennie D.* User's Guide to the Medical Literature: Essentials of Evidence-Based Practice. *Chicago: American Medical Association, 2002.*

aggregated evidence reveals a statistically significant benefit in favor of the recommendation to stay active: quite a different conclusion from the equivocal judgment suggested by a simple count of positive trials versus negative trials.

### Synthesized Results From Multiple Clinical Trials: Clinical Practice Guidelines

Clinical practice guidelines integrate synthesized evidence with broader cultural, societal, and patient-interest considerations. Results in practice guidelines come in the form of recommendations supported by specified levels of evidence. Readers performing a critical appraisal of a practice guideline should determine the method used by panel members to grade treatment recommendations, and then consider the relative strength of each recommendation. A common scheme for grading recommendations in clinical practice guidelines is reproduced in **TABLE 3**.

## STEP 3C. CRITICALLY APPRAISING THE LITERATURE: HOW CAN I APPLY THE RESULTS TO PATIENT CARE?

The final question in a critical appraisal of evidence involves a series of deliberate judgments about the relevance and applicability of the evidence to a specific patient in the context of a specific clinical setting. An evidence-based practitioner will need to decide whether the patient under consideration is sufficiently similar to the patients in the study or group of studies for the results to be relevant. For example, the clinician should determine whether the patients enrolled in the study were similar to his/her own patient, including the inclusion and exclusion criteria, age, gender, race, sociodemographics, stage of illness, comorbidity and disability status, and prognosis. Next, the practitioner must integrate patient values, preferences, and expectations in shared decision making when selecting a particular treatment. Also, the evidence will be relevant to a given patient only if outcomes measured in the clinical trial are consistent with the individual patient's goals. Consideration must be given to whether the treatment as structured in the research study is acceptable to the patient. Many issues must be considered, such as anticipated frequency and duration of patient visits, cost of the treatment, possible discomfort or other adverse effects of the intervention of interest and of competing interventions (such as injections, surgery, or other noninvasive interventions), and how consistent the treatment is with patient expectations. This final question also prompts the practitioner to integrate personal clinical expertise. Some treatments require specialty skills or specific equipment that may not be currently available and may not be obtainable in a reasonable amount of time to help a particular patient.

Critical appraisal is an essential skill for an evidence-based practitioner. Although applying the principles outlined above for critical appraisal may be difficult to master initially, the process becomes much easier with practice. Critical appraisal using these principles is the best method to facilitate independent professional judgments about the validity, strength, and relevance of evidence for therapy. A checklist to organize key judgments during a critical appraisal for a RCT is included in **APPENDIX A**.

## STEP 4. INCORPORATING EVIDENCE INTO CLINICAL PRACTICE

Once it has been determined through critical appraisal that a particular study or group of studies provides valid, applicable evidence that a treatment yields clinically meaningful benefits, the clinician should integrate the evidence into clinical practice. If a given patient is reasonably similar to those in the study, a clinician should be able to integrate valid evidence with considerable confidence. However, any given patient will have a unique set of prognostic attributes. Clinicians must recognize that treatments typically are not uniformly effective inasmuch as reported results are for average treatment effects.[10] This is another reason why the clinician must integrate the best available evidence with clinical expertise and the goals, values, and expectations of the patient when determining which interventions are preferable for a particular individual.

Many perceived barriers may prevent successful integration of EBP into physical therapist practice.[47,58] One barrier is excessive reliance on clinical expertise which can be associated with failure to acknowledge and incorporate current best evidence into clinical practice. Expertise in physical therapist practice has been described as possession of professional values, decision-making processes, communication styles or skills, specialty certifications, and years of practice in physical therapy.[47,75] A study by Childs and colleagues[16] found that experienced physical therapists with orthopaedic or sports certifications demonstrate greater knowledge in managing musculoskeletal conditions than therapists without specialty certification. Despite these findings, one cannot infer that patients cared for by expert clinicians will achieve superior outcomes when compared to the outcomes of patients treated by novice clinicians.[71,85] In fact, it has been demonstrated that expert clinicians are often resistant to changing their practice behaviors even when their treatment approaches have been disproven.[5] Hence, while clinical expertise is important, it is insufficient to assure optimal outcomes. Reliance on clinical experience without including knowledge and application of evidence to clinical care is inconsistent with the principles of EBP.[16,85] Therefore, seeking and incorporating the best available evidence should be an integral part of the clinical decision-making process.

Instituting behavior change among practicing clinicians is one of the foremost barriers to successful integration of EBP.[17,36,85] While some clinicians are quick to adopt change, many others are unfortunately resistant to change and rely predominantly on their clinical experience rather than incorporating evidence into their practice.[9] Although the volume and quality of emerging evidence in many areas of physical therapist practice is mounting rapidly, we acknowledge that there are still many areas where evidence is sparse and inconclusive. In these instances, rather than waiting for the "perfect evidence," clinicians should act on the research evidence that is currently available and follow up by using patient-centered outcomes tools to determine those interventions which are effective for a particular patient and those which are not.[70] Critical appraisals for lower-level evidence, such as cohort studies, case series, and case reports, can be performed using the same principles outlined above and in part I of this series. However, it becomes immediately apparent when appraising lower-level evidence that unprotected validity threats in these types of studies permit substantial bias and severely limit confidence in reported results. The hierarchy of evidence does not exclude expert opinion (level 5 evidence); but opinion should be considered best evidence only with specific knowledge that higher-level evidence does not yet exist. Finally, it should be recognized that the results from higher levels of evidence, such as systematic reviews, might conclude that there is currently insufficient evidence to support one intervention option over another. In these instances, treatment decisions based on clinician expertise and experience (although these are lower forms of evidence in most evidence hierarchies) may in fact be the most appropriate form of guidance to inform clinical decision making.

To illustrate how knowledge of current best evidence, combined with critical appraisal skills, can guide clinical decision making, consider the case of a 74-year-old female with a history of spinal stenosis and cardiovascular disease who indicated that she developed her most recent bout of low back pain after injuring her back while playing with her great granddaughter 3 weeks previously. Her Modified Low Back Pain Disability Index was 20% and she indicated that her goals were to complete household activities without making her back pain worse and to be able to play with her great granddaughter in 2 weeks. The most impressive findings from the physical exam include generalized stiffness and loss of motion in both hips and lumbar spine in flexion. In consultation with the patient, you indicate that her goals seem realistic and that you wish to reassess her Modified Low Back Pain Disability score in 2 weeks and expect her to demonstrate at least a 6-point change. Your intervention strategy includes patient education, joint mobilization to the hips and lumbar spine, and implementation of a body weight-supported walking program.

This patient case illustrates several important issues. Although this patient has 2 potentially negative prognostic factors—a history of recurrent back pain and cardiovascular disease—her modified low back pain disability score of 20 indicates a mild level of disability. Because the MCID for the Modified Low Back Pain Disability Questionnaire is 6 points,[31] this is chosen as the quantitative goal that seems to best match those described by the patient. The intervention strategy is based on a recently published clinical trial by Whitman and colleagues[84] that used a program of patient education, body-weight supported treadmill training, and joint mobilization to the spine and hip joints. The typical subjects in the clinical trial were women with an average age of 69 years and a baseline Modified Oswestry score of 36, which seem to closely match the characteristics of this patient. In addition, the average modified low back pain disability score reduction at 6 weeks of the intervention program was approximately 10 points. Therefore, the goal of a 6-point change in 2 weeks seems realistic. As discussed in a previous section, however, MCIDs that are established based on group data can be misleading if applied to individual patients. Therefore, a more conservative approach of establishing goals that exceed the MCID threshold might be a better guideline to ensure that self-report measures represent true clinically important change.

## STEP 5. EVALUATING PERFORMANCE ON STEPS 1 THROUGH 4

Although most of this commentary addresses critical appraisal of evidence, this fifth and final step in the process of achieving successful implementation of EBP is arguably the most important. Self-assessment of practice begins as a student in the form of self-observation and judgmental processing and should continue through one's professional career.[64] The skills of self-awareness assist clinicians in identifying personal strengths as well as limitations.[27] It is with reflective practice that physical therapists will refine their efficiency with integrating the best available evidence into clinical practice.

Recognition of personal and professional limitations can be difficult and may result in avoidance of the issues, regardless of the internal drive and motivation of the therapist.[27] Developing competence in the EBP process will require clinicians to acknowledge times of uncertainty and the need for gathering information. Competence includes self-awareness on behalf of the therapist and the ability to recognize personal limitations, which can be very difficult. Straus et al[81] have developed a series of questions (**APPENDIX B**) to facilitate introspective self-evaluation for the evidence-based practitioner. Therapists should reflect subjectively on their ability to proceed through the first 4 steps, but should also assess patient outcomes objectively and formally in the context of best available evidence. Physical therapists should use reliable and valid outcome measures for every patient they see in clinical practice to ascertain if true and clinically meaningful changes in patient status occurred (ie, did patient improvements exceed the outcome scale's MDC and MCID scores). The data obtained through the use of valid and reliable outcomes tools, along with the self-evaluation of effectiveness and efficiency with the 4 steps, will enhance clinical practice. Clinicians may find it helpful to read one of the several case studies or case series where clinicians provide detailed description of applying current best evidence in managing patients with a variety of conditions. For example, MacDonald and colleagues[57] reported on the management of a series of patients with hip dysfunction who responded positively to novel manual therapy interventions. Similarly, Cleland et al[21] and Waldrop[82] have published case series that apply recently developed clinical prediction rules to patients.

As proposed by Flynn and colleagues,[30] the use of minimal data collection forms that include key examination findings and appropriate patient-centered outcome measures will allow students as well as practicing clinicians to monitor their individual clinical performance. With this information, clinicians can compare average patient improvements in clinical settings to average patient improvements in the current best evidence (ie, peer-reviewed, published literature), while accounting for differences between clinical and research settings and contexts. It is ultimately through these quality measurement processes and accountability to EBP principles that therapists become clinicians of excellence.[9]

## SUMMARY

Determining the source, validity, strength, and relevance of evidence for treatment decisions requires successful integration of the EBP process. The goal of EBP is to improve efficiency and assist clinicians in selecting interventions that will maximize patient outcomes rather than erroneously selecting interventions with little or no demonstrated effectiveness.[56]

The identification of appropriate foreground questions, performing literature searches, critically analyzing the best available evidence, applying the best evidence to clinical practice, and ultimately assuring the proficiency of the process will ultimately lead to optimal care for our patients. Developing proficiency in the 5-step process to EBP requires strong dedication and effort from students as well as practicing therapists, and at times can be quite challenging. However, as healthcare providers, therapists should approach the challenge of successful integration of EBP with enthusiasm, as the overall goal is to provide the best quality of care and maximize positive outcomes for their patients. They should embrace and not retreat from the challenge of integrating the best available evidence, clinical expertise, and patient values into clinical decisions for each

## REFERENCES


1. Physiotherapy Evidence Database (PEDro). Available at: http://www.pedro.fhs.usyd.edu.au/calculator.html . Accessed July 11, 2008
2. Levels of evidence and grades of recommendations. Available at: http://www.cebm.net/levels_of_evidence.asp. Accessed July 11, 2008.
3. Altman DG. Confidence intervals. In: Straus SE, Richardson WS, Glasziou P, Haynes RB, eds. *Evidence-based Medicine: How to Practice and Teach EBM.* Edinburgh, UK: Churchill Living-

stone; 2005.
4. Angst F, Aeschlimann A, Stucki G. Smallest detectable and minimal clinically important differences of rehabilitation intervention with their implications for required sample sizes using WOMAC and SF-36 quality of life measurement instruments in patients with osteoarthritis of the lower extremities. *Arthritis Rheum.* 2001;45:384-391. http://dx.doi.org/10.1002/1529-0131(200108)45:4<384::AID-ART352>3.0.CO;2-0
5. Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction. *JAMA*. 1992;268:240-248.
6. Bandy W. Use of statistics in physical therapy over a 2-year period 2000-2002: implications for educators. *J Phys Ther Ed.* 2006;17:67-70.
7. Barr S, Bellamy N, Buchanan WW, et al. A comparative study of signal versus aggregate methods of outcome measurement based on the WOMAC Osteoarthritis Index. Western Ontario and McMaster Universities Osteoarthritis Index. *J Rheumatol.* 1994;21:2106-2112.
8. Beaton DE, Boers M, Wells GA. Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. *Curr Opin Rheumatol.* 2002;14:109-114.
9. Berwick DM. Disseminating innovations in health care. *JAMA*. 2003;289:1969-1975. http://dx.doi.org/10.1001/jama.289.15.1969
10. Bhandari M, Haynes RB. How to appraise the effectiveness of treatment. *World J Surg.* 2005;29:570-575. http://dx.doi.org/10.1007/s00268-005-7915-9
11. Binkley JM, Stratford PW, Lott SA, Riddle DL. The Lower Extremity Functional Scale (LEFS): scale development, measurement properties, and clinical application. North American Orthopaedic Rehabilitation Research Network. *Phys Ther.* 1999;79:371-383.
12. Bischoff-Ferrari HA, Dawson-Hughes B, Willett WC, et al. Effect of Vitamin D on falls: a meta-analysis. *JAMA*. 2004;291:1999-2006. http://dx.doi.org/10.1001/jama.291.16.1999
13. Butcher SJ, Craven BR, Chilibeck PD, Spink KS, Grona SL, Sprigings EJ. The effect of trunk stability training on vertical takeoff velocity. *J Orthop Sports Phys Ther.* 2007;37:223-231. http://dx.doi.org/10.2519/jospt.2007.2331
14. Cacchio A, Paoloni M, Barile A, et al. Effectiveness of radial shock-wave therapy for calcific tendinitis of the shoulder: single-blind, randomized clinical study. *Phys Ther.* 2006;86:672-682.
15. Childs JD, Piva SR, Fritz JM. Responsiveness of the numeric pain rating scale in patients with low back pain. *Spine*. 2005;30:1331-1334.
16. Childs JD, Whitman JM, Sizer PS, Pugia ML, Flynn TW, Delitto A. A description of physical therapists' knowledge in managing musculoskeletal conditions. *BMC Musculoskelet Disord.* 2005;6:32. http://dx.doi.org/10.1186/1471-2474-6-32
17. Choudhry NK, Fletcher RH, Soumerai SB. Systematic review: the relationship between clinical experience and quality of health care. *Ann Intern Med.* 2005;142:260-273.
18. Clegg DO, Reda DJ, Harris CL, et al. Glucosamine, chondroitin sulfate, and the two in combination for painful knee osteoarthritis. *N Engl J Med.* 2006;354:795-808. http://dx.doi.org/10.1056/NEJMoa052771
19. Cleland JA, Fritz JM, Whitman JM, Palmer JA. The reliability and construct validity of the Neck Disability Index and patient specific functional scale in patients with cervical radiculopathy. *Spine*. 2006;31:598-602. http://dx.doi.org/10.1097/01.brs.0000201241.90914.22
20. Cleland JA, Glynn P, Whitman JM, Eberhart SL, MacDonald C, Childs JD. Short-term effects of thrust versus nonthrust mobilization/manipulation directed at the thoracic spine in patients with neck pain: a randomized clinical trial. *Phys Ther.* 2007;87:431-440. http://dx.doi.org/10.2522/ptj.20060217
21. Cleland JA, Whitman JM, Fritz JM, Palmer JA. Manual physical therapy, cervical traction, and strengthening exercises in patients with cervical radiculopathy: a case series. *J Orthop Sports Phys Ther.* 2005;35:802-811. http://dx.doi.org/10.2519/jospt.2005.2077
22. Cohen L. *Statistical Power Analysis for the Behavioral Sciences.* 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
23. Denis M, Moffet H, Caron F, Ouellet D, Paquet J, Nolet L. Effectiveness of continuous passive motion and conventional physical therapy after total knee arthroplasty: a randomized clinical trial. *Phys Ther.* 2006;86:174-185.
24. Deyle GD, Allison SC, Matekel RL, et al. Physical therapy treatment effectiveness for osteoarthritis of the knee: a randomized comparison of supervised clinical exercise and manual therapy procedures versus a home exercise program. *Phys Ther.* 2005;85:1301-1317.
25. Deyle GD, Henderson NE, Matekel RL, Ryder MG, Garber MB, Allison SC. Effectiveness of manual physical therapy and exercise in osteoarthritis of the knee. A randomized, controlled trial. *Ann Intern Med.* 2000;132:173-181.
26. Domholdt E. *Physical Therapy Research.* 2nd ed. Philadelphia, PA: W.B. Saunders Co; 2000.
27. Epstein RM. Mindful practice. *JAMA*. 1999;282:833-839.
28. Farrar JT, Young JP, Jr., LaMoreaux L, Werth JL, Poole RM. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. *Pain*. 2001;94:149-158.
29. Fidler F, Thomason N, Cumming G, Finch S, Leeman J. Editors can lead researchers to confidence intervals, but can't make them think: statistical reform lessons from medicine. *Psychol Sci.* 2004;15:119-126.
30. Flynn TW, Wainner RS, Fritz JM. Spinal manipulation in physical therapist professional degree education: A model for teaching and integration into clinical practice. *J Orthop Sports Phys Ther.* 2006;36:577-587. http://dx.doi.org/10.2519/jospt.2006.2159
31. Fritz JM, Irrgang JJ. A comparison of a modified Oswestry Low Back Pain Disability Questionnaire and the Quebec Back Pain Disability Scale. *Phys Ther.* 2001;81:776-788.
32. Fritz JM, Whitman JM, Childs JD. Lumbar spine segmental mobility assessment: an examination of validity for determining intervention strategies in patients with low back pain. *Arch Phys Med Rehabil.* 2005;86:1745-1752. http://dx.doi.org/10.1016/j.apmr.2005.03.028
33. Gardner MJ, Altman DG. *Statistics With Confidence.* London, UK: BMJ Books; 2005.
34. Gerber JP, Marcus RL, Dibble LE, Greis PE, Burks RT, Lastayo PC. Safety, feasibility, and efficacy of negative work exercise via eccentric muscle activity following anterior cruciate ligament reconstruction. *J Orthop Sports Phys Ther.* 2007;37:10-18. http://dx.doi.org/10.2519/jospt.2007.2362
35. Greenfield ML, Kuhn JE, Wojtys EM. A statistics primer. Confidence intervals. *Am J Sports Med.* 1998;26:145-149.
36. Grimshaw JM, Shirran L, Thomas R, et al. Changing provider behavior: an overview of systematic reviews of interventions. *Med Care.* 2001;39:II2-45.
37. Guyatt G, Rennie D. *User's Guide to the Medical Literature: Essentials of Evidence-Based Practice.* Chicago, IL: American Medical Association; 2002.
38. Guyatt G, Walter S, Cook D, Jaeschke R. Therapy and understanding the results: confidence intervals. In: Guyatt G, Rennie D, eds. *User's Guide to the Medical Literature: Essentials of Evidence-Based Practice.* Chicago, IL: American Medical Association; 2002:
39. Hagen KB, Hilde G, Jamtvedt G, Winnem M. Bed rest for acute low-back pain and sciatica. *Cochrane Database Syst Rev.* 2004;CD001254. http://dx.doi.org/10.1002/14651858.CD001254.pub2
40. Hagg O, Fritzell P, Nordwall A. The clinical importance of changes in outcome scores after treatment for chronic low back pain. *Eur Spine J.* 2003;12:12-20. http://dx.doi.org/10.1007/s00586-002-0464-0
41. Hale SA, Hertel J, Olmsted-Kramer LC. The effect of a 4-week comprehensive rehabilitation program on postural control and lower extremity function in individuals with chronic ankle instability. *J Orthop Sports Phys Ther.* 2007;37:303-311. http://dx.doi.org/10.2519/jospt.2007.2322
42. Hall T, Chan HT, Christensen L, Odenthal B, Wells C, Robinson K. Efficacy of a C1-C2 self-sustained natural apophyseal glide (SNAG) in the management of cervicogenic headache. *J Orthop Sports Phys Ther.* 2007;37:100-107.
43. Hartig DE, Henderson JM. Increasing hamstring flexibility decreases lower extremity overuse injuries in military basic trainees. *Am J Sports Med.* 1999;27:173-176.
44. Herbert RD. How to estimate treatment effects

from reports of clinical trials. II: Dichotomous outcomes. *Aust J Physiother.* 2000;46:309-313.
45. Hyland MR, Webber-Gaffney A, Cohen L, Lichtman PT. Randomized controlled trial of calcaneal taping, sham taping, and plantar fascia stretching for the short-term management of plantar heel pain. *J Orthop Sports Phys Ther.* 2006;36:364-371. http://dx.doi.org/10.2519/jospt.2006.2078
46. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials.* 1989;10:407-415.
47. Jensen GM, Gwyer J, Shepard KF. Expert practice in physical therapy. *Phys Ther.* 2000;80:28-43; discussion 44-52.
48. Jette DU, Bacon K, Batty C, et al. Evidence-based practice: beliefs, attitudes, knowledge, and behaviors of physical therapists. *Phys Ther.* 2003;83:786-805.
49. Johnson AJ, Godges JJ, Zimmerman GJ, Ounanian LL. The effect of anterior versus posterior glide joint mobilization on external rotation range of motion in patients with shoulder adhesive capsulitis. *J Orthop Sports Phys Ther.* 2007;37:88-99. http://dx.doi.org/10.2519/jospt.2007.2307
50. Jones B, Jarvis P, Lewis JA, Ebbutt AF. Trials to assess equivalence: the importance of rigorous methods. *BMJ*. 1996;313:36-39.
51. Jull G, Trott P, Potter H, et al. A randomized controlled trial of exercise and manipulative therapy for cervicogenic headache. *Spine*. 2002;27:1835-1843; discussion 1843.
52. Keers JC, Groen H, Sluiter WJ, Bouma J, Links TP. Cost and benefits of a multidisciplinary intensive diabetes education programme. *J Eval Clin Pract.* 2005;11:293-303. http://dx.doi.org/10.1111/j.1365-2753.2005.00536.x
53. Koumantakis GA, Watson PJ, Oldham JA. Trunk muscle stabilization training plus general exercise versus general exercise only: randomized controlled trial of patients with recurrent low back pain. *Phys Ther.* 2005;85:209-225.
54. Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. *N Engl J Med.* 1988;318:1728-1733.
55. Lee JS, Hobden E, Stiell IG, Wells GA. Clinically important change in the visual analog scale after adequate pain control. *Acad Emerg Med.* 2003;10:1128-1130.
56. MacDermid JC. An introduction to evidence-based practice for hand therapists. *J Hand Ther.* 2004;17:105-117. http://dx.doi.org/10.1197/j.jht.2004.02.001
57. MacDonald CW, Whitman JM, Cleland JA, Smith M, Hoeksma HL. Clinical outcomes following manual physical therapy and exercise for hip osteoarthritis: A case series. *J Orthop Sports Phys Ther.* 2006;36:588-599. http://dx.doi.org/10.2519/jospt.2006.2233
58. Maher CG, Sherrington C, Elkins M, Herbert RD, Moseley AM. Challenges for evidence-based physical therapy: accessing and interpreting high-quality evidence on therapy. *Phys Ther.* 2004;84:644-654.
59. Make B. How can we assess outcomes of clinical trials: the MCID approach. *COPD*. 2007;4:191-194. http://dx.doi.org/10.1080/15412550701471231
60. Michener LA, McClure PW, Sennett BJ. American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form, patient self-report section: reliability, validity, and responsiveness. *J Shoulder Elbow Surg.* 2002;11:587-594. http://dx.doi.org/10.1067/mse.2002.127096
61. Moiler K, Hall T, Robinson K. The role of fibular tape in the prevention of ankle injury in basketball: A pilot study. *J Orthop Sports Phys Ther.* 2006;36:661-668. http://dx.doi.org/10.2519/jospt.2006.2259
62. 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
63. Moseley JB, O'Malley K, Petersen NJ, et al. A controlled trial of arthroscopic surgery for osteoarthritis of the knee. *N Engl J Med.* 2002;347:81-88. http://dx.doi.org/10.1056/NEJMoa013259
64. Musolino GM. Fostering reflective practice: self-assessment abilities of physical therapy students and entry-level graduates. *J Allied Health.* 2006;35:30-42.
65. Newman D, Allison SC. Risk and physical therapy? *J Orthop Sports Phys Ther.* 2007;37:287-289. http://dx.doi.org/10.2519/jospt.2007.0106
66. Oldridge N, Perkins A, Marchionni N, Fumagalli S, Fattirolli F, Guyatt G. Number needed to treat in cardiac rehabilitation. *J Cardiopulm Rehabil.* 2002;22:22-30.
67. O'Leary S, Jull G, Kim M, Vicenzino B. Specificity in retraining craniocervical flexor muscle performance. *J Orthop Sports Phys Ther.* 2007;37:3-9. http://dx.doi.org/10.2519/jospt.2007.2237
68. Palombaro KM, Craik RL, Mangione KK, Tomlinson JD. Determining meaningful changes in gait speed after hip fracture. *Phys Ther.* 2006;86:809-816.
69. Portney LG, Watkins MP. *Foundations of Clinical Research: Applications to Practice.* Upper Saddle River, NJ: Prentice Hall Health; 2000.
70. Reinertsen JL. Zen and the art of physician autonomy maintenance. *Ann Intern Med.* 2003;138:992-995.
71. Resnik L, Hart DL. Using clinical outcomes to identify expert physical therapists. *Phys Ther.* 2003;83:990-1002.
72. Rydeard R, Leger A, Smith D. Pilates-based therapeutic exercise: effect on subjects with nonspecific chronic low back pain and functional disability: a randomized controlled trial. *J Orthop Sports Phys Ther.* 2006;36:472-484. http://dx.doi.org/10.2519/jospt.2008.2669
73. Sackett DL, Straws SE, Richardson WS, Rosenberg W, Haynes RB. *Evidence-Based Medicine: How to Practice and Teach EBM.* London, UK: Harcourt Publishers Limited; 2000.
74. Schmitt JS, Di Fabio RP. Reliable change and minimum important difference (MID) proportions facilitated group responsiveness comparisons using individual threshold criteria. *J Clin Epidemiol.* 2004;57:1008-1018. http://dx.doi.org/10.1016/j.jclinepi.2004.02.007
75. Shepard KF, Hack LM, Gwyer J, Jensen GM. Describing expert practice in physical therapy. *Qual Health Res.* 1999;9:746-758.
76. Sim J, Reid N. Statistical inference by confidence intervals: issues of interpretation and utilization. *Phys Ther.* 1999;79:186-195.
77. Stratford PW, Binkley FM, Riddle DL. Health status measures: strategies and analytic methods for assessing change scores. *Phys Ther.* 1996;76:1109-1123.
78. Stratford PW, Binkley J, Solomon P, Finch E, Gill C, Moreland J. Defining the minimum level of detectable change for the Roland-Morris questionnaire. *Phys Ther.* 1996;76:359-365; discussion 366-358.
79. Stratford PW, Binkley JM, Riddle DL, Guyatt GH. Sensitivity to change of the Roland-Morris Back Pain Questionnaire: part 1. *Phys Ther.* 1998;78:1186-1196.
80. Stratford PW, Riddle DL, Binkley JM, Spadoni G, Westaway MD, Padfield B. Using the Neck Disability Index to make decisions concerning individual patients. *Physiother Can.* 1999;51:107-112.
81. Straus SE, Richardson WS, Glasziou P, Haynes RB. *Evidence-Based Medicine.* 3rd ed. Edinburgh, UK: Elsevier/Churchill Livingstone; 2005.
82. Waldrop MA. Diagnosis and treatment of cervical radiculopathy using a clinical prediction rule and a multimodal intervention approach: a case series. *J Orthop Sports Phys Ther.* 2006;36:152-159. http://dx.doi.org/10.2519/jospt.2006.2056
83. Watt CF, Taylor NF, Baskus K. Do Colles' fracture patients benefit from routine referral to physiotherapy following cast removal? *Arch Orthop Trauma Surg.* 2000;120:413-415.
84. Whitman JM, Flynn TW, Childs JD, et al. A comparison between two physical therapy treatment programs for patients with lumbar spinal stenosis: a randomized clinical trial. *Spine*. 2006;31:2541-2549. http://dx.doi.org/10.1097/01.brs.0000241136.98159.8c
85. Whitman JM, Fritz JM, Childs JD. The influence of experience and specialty certifications on clinical outcomes for patients with low back pain treated within a standardized physical therapy management program. *J Orthop Sports Phys Ther.* 2004;34:662-672; discussion 672-665. http://dx.doi.org/10.2519/jospt.2004.1535
86. Wise RA, Brown CD. Minimal clinically important differences in the six-minute walk test and the incremental shuttle walking test. *COPD*. 2005;2:125-129.

## APPENDIX A

### CHECKLIST FOR CRITICAL APPRAISAL OF A RANDOMIZED CONTROLLED TRIAL

| | Yes | No | Can't Tell | Not Applicable |
|---|---|---|---|---|
| **Are the results valid?** | | | | |
| Was a randomization procedure explicitly reported? | ___ | ___ | ___ | ___ |
| Was group assignment concealed from those enrolling patients? | ___ | ___ | ___ | ___ |
| Were groups reasonably homogenous at baseline? | ___ | ___ | ___ | ___ |
| Were the patients blinded to the treatment they received? | ___ | ___ | ___ | ___ |
| Were treating clinicians blinded to group membership? | ___ | ___ | ___ | ___ |
| Were data collectors blinded to group membership? | ___ | ___ | ___ | ___ |
| Was the follow-up period sufficiently long? | ___ | ___ | ___ | ___ |
| Did any patients drop out or switch group assignment? | ___ | ___ | ___ | ___ |
| If there were dropouts or switchover patients, was an intention-to-treat analysis performed? | ___ | ___ | ___ | ___ |
| Was the overall research experience equivalent for groups, other than the treatment(s) of interest? | ___ | ___ | ___ | ___ |
| **What are the results?** | | | | |
| Are the treatment effects statistically significant (a positive trial)? | ___ | ___ | ___ | ___ |
| In a positive trial, is the treatment effect size clinically meaningful (equal to or larger than the MCID*)? | ___ | ___ | ___ | ___ |
| In a positive trial, does the 95% confidence interval around the point estimate of the treatment effect exclude the MCID? | ___ | ___ | ___ | ___ |
| In a negative trial, does the 95% confidence interval around the point estimate of the treatment effect exclude the MCID? | ___ | ___ | ___ | ___ |
| **How can I apply the results to patient care?** | | | | |
| Is my patient sufficiently similar to patients in the treatment group? | ___ | ___ | ___ | ___ |
| Are the outcomes measured in the study relevant to my patient's goals? | ___ | ___ | ___ | ___ |
| Is the treatment compatible with my patient's values, preferences, and expectations? | ___ | ___ | ___ | ___ |
| Are the anticipated benefits worth the costs and potential for any adverse effects? | ___ | ___ | ___ | ___ |
| Do I have the clinical skills and any required equipment to provide the treatment? | ___ | ___ | ___ | ___ |

*Abbreviation: MCID, minimal clinically important difference.*

## APPENDIX B

### SELF-EVALUATION QUESTIONS FOR EVIDENCE-BASED PRACTITIONERS*

**Self-evaluation in asking answerable questions**

1. Am I asking any clinical questions at all?
2. Am I asking well-formulated questions:
   - Two-part questions about "background" knowledge?
   - Four- or three-part questions about "foreground" diagnosis, management, etc?
3. Am I using a "map" to locate my knowledge gaps and articulate questions?
4. Can I get myself "unstuck" when asking questions?
5. Do I have a working method to save my questions for later answering?

**A self-evaluation in finding the best external evidence**

1. Am I searching at all?
2. Do I know the best sources of current evidence for my clinical discipline?
3. Have I achieved immediate access to searching hardware, software, and the best evidence for my clinical discipline?
4. Am I finding useful external evidence from a widening array of sources?
5. Am I becoming more efficient in my searching?
6. Am I using truncations, Booleans, MeSH headings, thesaurus, limiters, and intelligent free text when searching MEDLINE?
7. How do my searches compare with those of research librarians or other respected colleagues who have a passion for providing best current patient care?

**A self-evaluation in critically appraising the evidence for its validity and potential usefulness**

1. Am I critically appraising external evidence at all?
2. Are the critical appraisal guides becoming easier for me to apply?
3. Am I becoming more accurate and efficient in applying some of the critical appraisal measures (such as likelihood ratios, NNTs, and the like)?
4. Am I creating any appraisal summaries?

**A self-evaluation in integrating the critical appraisal with clinical expertise and applying the result in clinical practice**

1. Am I integrating my critical appraisals into my practice at all?
2. Am I becoming more accurate and efficient in adjusting some of the critical appraisal measures to fit my individual patients (pretest probabilities, NNT/f, etc.)?
3. Can I explain (and resolve) disagreements about management decisions in terms of this integration?

**A self-evaluation of changing practice behavior**

1. When new evidence suggests a change in practice, am I identifying barriers to this change?
2. Have I carried out any check, such as audits of my diagnostic, therapeutic, or other EBM performance?

** Reproduced with permission from Straus SE, Richardson WS, Glasziou P, Haynes RB.* Evidence-Based Medicine: How to Practice and Teach EBM. *3rd ed. Edinburgh, UK: Churchill Livingstone; 2005. © 2005 Elsevier.*