Critical Appraisal of Research Evidence for Its Validity and Usefulness

Joy C. MacDermid, BScPT, PhD[a,b,*], David M. Walton, MScPT, PhD (c)[c], Mary Law, PhD[d]

KEYWORDS

• Relevance • Clinical research • Critical appraisal
• Evidence-based • Quality • Clinical recommendations

FIVE STEPS OF EVIDENCE-BASED PRACTICE

The five steps in the evidence-based practice (EBP) approach are:

Ask a specific clinical question.
Find the best evidence to answer the question.
Critically appraise the evidence for its validity and usefulness.
Integrate appraisal results with clinical expertise and patient values.
Evaluate the outcomes.

Step 3 in the EBP approach involves critical appraisal of the validity and usefulness of evidence, with the specific goal of identifying the highest quality evidence that applies to a given clinical question. Because evidence-based decision making requires using the best available evidence, quality and relevance judgments are important components in the process. In fact, this third step can be broken down into three sequential subcomponents: (1) determine whether the results of individual studies are true (internally valid); (2) determine whether the results apply to a given patient (generalizability/external validity); and (3) determine the nature and strength of recommendations based on synthesis of several individual evidence resources.

Critical Appraisal of Individual Study Quality (Internal Validity)

The importance of critical appraisal in EBP has led to the development of systems, processes, tools, and support systems for rating clinical research evidence. In fact, we now have systematic reviews of appraisal tools.[1] In addition, there has been an increased move toward having experts in critical appraisal perform this task. Clinicians are then able to ''pull-out'' preappraised forms of evidence, such as the PEDro Physiotherapy Evidence Database or OTSeeker. Most recently, there has been development of ''push-out'' approaches, where high quality, critically appraised evidence resources already rated by experts are sent directly to end users with specific information needs (eg, BMJ updates). This article focuses on how hand surgeons and therapists can access and apply ranking systems, critical appraisal tools,

J.C.M. is funded by a New Investigator Award, Canadian Institutes of Health Research. D.M.W. is funded by a Doctoral Fellowship, Canadian Institutes of Health Research. M.L. holds the John and Margaret Lillie Chair in Childhood Disability.

[a] Hand and Upper Limb Centre Clinical Research Laboratory, St. Joseph's Health Centre, 268 Grosvenor Street, London, Ontario, N6A 4L6, Canada
[b] School of Rehabilitation Science, McMaster University, Institute for Applied Health Sciences, 1400 Main Street West, 4th Floor, Hamilton, Ontario L8S 1C7, Canada
[c] The University of Western Ontario School of Physical Therapy, Room EC 1588, 1201 Western Road, London, Ontario, N6G 1H1, Canada
[d] School of Rehabilitation Science, McMaster University, 268 Grosvenor Street, Hamilton, Ontario, Canada
* Corresponding author. School of Rehabilitation Science, LB33, McMaster University, Institute for Applied Health Sciences, Room 429, 1400 Main Street West, 4th Floor, Hamilton, Ontario L8S 1C7, Canada
E-mail address: macderj@mcmaster.ca (J.C. MacDermid).

Hand Clin 25 (2009) 29–42
doi:10.1016/j.hcl.2008.11.003
0749-0712/08/$ – see front matter

and guides for making overall recommendations to provide guideposts on how research evidence can be transitioned into patient specific recommendations.

Critical appraisal first focuses on the internal validity of the study, or the extent to which the conclusions of the study are true within the particular context of the study. This process can be performed at various depths of analysis, such as quick classification systems or more detailed rating tools. Critical appraisal instruments range from very structured tools that contain specific questions and defined response categories, to more open-ended scales where the assessor makes guided subjective judgments on the quality of aspects of study design, using a framework provided by the assessment tool. Different critical appraisal tools are appropriate for different study designs. Hand surgeons and therapists should select different critical appraisal instruments depending on their clinical question, its associated study design, their familiarity with critical appraisal, personal preferences, accessibility of the literature, and a realistic balance between time commitment and depth of analysis.

Different depths of critical appraisal are also appropriate at different points in practice. For example, when needing to make quick decisions at the point of care, screening for specific randomized, controlled trials (RCTs) or presynthesized evidence may be the most expedient approach. The classic five levels of evidence will be useful for this purpose. In other cases, when planning to implement a new intervention into one's practice, there may be a significant learning curve and cost involved. Therefore, it would be important to delve more deeply into the study design to gain a more thorough understanding of issues that might affect the validity of the study conclusions, and the clinical interpretability or applicability across different patients. Furthermore, knowing the evidence about a specific planned intervention can guide its implementation. Clinicians who commit to learning and practicing detailed critical appraisal gain a greater appreciation of the issues that can compromise confidence in research studies. However, quick rating scales or even presynthesized evidence ratings have the advantage of being less time consuming than more traditional evaluation methods.

LEVELS OF EVIDENCE

The concept of ranking levels of evidence is based on the principle that certain study types have more rigor and these higher quality study designs provide more confidence to associated clinical decision-making. The "best" study design varies according to the type of study that is being conducted. For example, while the RCT is considered the best study design for detecting differences between intervention groups, for studies in prognosis a prospective cohort design with complete follow-up is the best design. The types of study designs that have been used often signify the state of knowledge about an intervention. Early in the development of an intervention, case series are the most common. Data from these designs are then used to develop RCTs. The classic "Sacketts" five levels of evidence are a broad ordinal tool but have had a tremendous impact. For example, many evidence reviews performed by the Cochrane Collaboration include either only RCTs or the two highest levels of evidence when conducting a systematic review.

The "Classic" Levels of Evidence for Treatment Effectiveness

Because treatment effectiveness is one of the primary interests of clinicians, and the RCT is the ideal design for experimental evaluation of treatment effectiveness, the conduct of RCTs has expanded exponentially. Early evidence rating systems for treatment effectiveness designated RCTs as level 1 evidence. With the proliferation of RCTs emerged a new research methodology: the systematic review. The original levels of evidence developed at McMaster University were subsequently updated and are clearly presented on the Web site for the Oxford Center For Evidence-Based Medicine by David Sackett and colleagues (last updated May 2001, http://www.cebm.net/levels_of_evidence.asp). This rating system allows you to classify individual studies in broad categories or "levels" (see the article by Szabo and MacDermid elsewhere in this issue). Level 1 is the highest level of evidence that can be achieved for treatment effectiveness. Three potential situations are considered to be sufficiently rigorous to be labeled as level 1. Level 1a would consist of a systematic review of a number of RCTs, where the studies substantially agree with each other in terms of the direction and approximate size of the effects observed. A level 1b study would be an individual RCT where the size of the treatment effect was defined by a narrow confidence interval. A level 1c study is a very unusual circumstance in surgery or hand therapy, and is when an all-or-none phenomenon occurs in the absence of a randomized study. An example of a level 1c would be a study where an overwhelmingly dramatic change in outcomes can be demonstrated once a new treatment

becomes available. Cases where all patients die before an intervention is available, and some survive following introduction up of a new intervention, provide overwhelming evidence. For example, vaccination is widely accepted in practice although not based on RCT evidence. Level 1 studies are those that provide the highest internal validity (confidence that the study results are true), enhancing our confidence that if we select this intervention for our patients, we will be able to achieve similar outcomes. These same levels pertain to studies of treatment effectiveness (therapy) prevention, etiology, and harm.

Fig. 1 illustrates how the levels of evidence hinge on a critical element important in research design. As we lose a critical element of internal validity, we also lose confidence that we might achieve the reported outcomes by selecting these interventions for our patients. Randomization is the single most protective factor against biases within clinical studies, as it controls for known and unknown confounders (assuming adequate sample size). Level 1 is the only level that provides experimental data, the remaining levels being observational.

Level 2 studies differ from RCTs in that we do not implement randomization. The protection against potential biases and confounders is lessened.[2] The most positive aspect of a prospective cohort study is that it identifies patients before experiencing the outcome (treatment or exposure), and thereby reduces the potential for a spectrum of biases (eg, differential recruitment, ascertainment bias, recall bias). A number of additional elements of research design are important to ensure that research designs maximize their internal validity. These include the use of standardized outcome measures, adequate sampling, appropriate blinding,[3–5] rigorous follow-up, and proper statistical analysis, including adjustment for important potential confounders. A level 2a study is a systematic review of cohort (prospective) studies that agree with each other in terms of the direction and approximate size of the effects obtained. A level 2b study is a single, high quality cohort study (with greater than 80% follow-up). Follow-up is a critical element of quality, particularly in cohort studies where differential loss to follow-up might obliterate equality between groups, if it existed at the outset. Patients can drop out of studies because they experience overly favorable or unfavorable results compared with the remainder of the cohort. Thus, estimates of treatment effects may be over- or underestimated.

Level 3 studies for therapy are case-controlled designs. In a case-controlled study design, subgroups of patients are identified for study after their outcomes have been reached. Data collection about exposures, treatment options, and complications is retrospective. An example of such a design is a study of patients who did or did not return to work within 2 months following carpal tunnel surgery. Differences in these two groups of patients would be examined retrospectively to determine if treatment, or personal or work factors were associated with not returning to work. In case-controlled studies, we no longer have prospective data collection and are now subjected to additional sources of bias. For example, the authors experienced differential loss to follow-up in carpal tunnel surgery studies[6,7] where

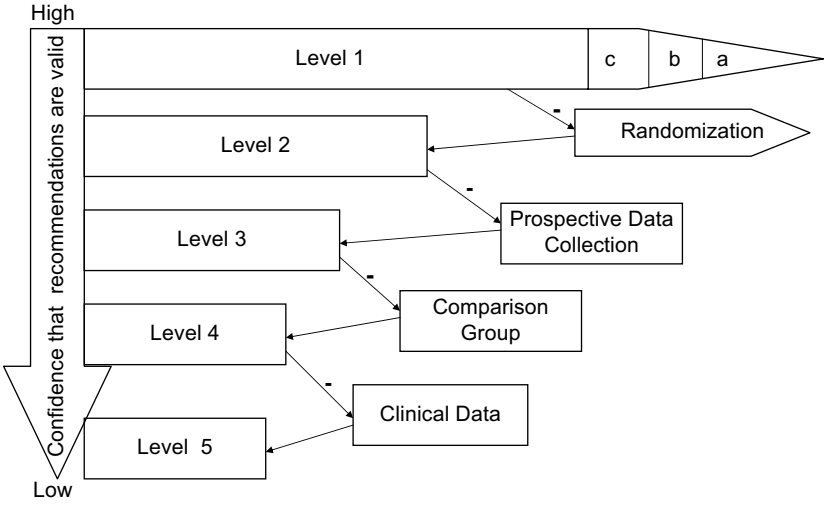

Fig. 1. Relationship between levels of evidence and key elements.

patients who were satisfied and had returned to work were reluctant to return for follow-up visits. Conversely, in another concurrent study on a different group of patients[8] (resection arthroplasty), the authors observed that patients who were dissatisfied were reluctant to return for a study visit determining final outcome status. The effect of dropouts on the estimated outcomes was different in these two cases. The reasons that specific subgroups of patients exist, are available for study, or provide specific outcomes data, are potentially related to the outcomes achieved (confounders) and can contaminate the observations.

At the next level of evidence another critical element of research design is lost: the comparison group. Level 4 evidence for treatment effectiveness consists of a single group or case series. No matter how rigorously we evaluate their outcomes, we remain uncertain what would have happened to these patients if an alternate intervention had been selected. Despite this flaw, case series remain one of the most common study designs reported in hand surgery journals,[9] and have, in some cases, been able to provide sufficient evidence to change practice, particularly where harm is demonstrated (eg, silicone synovitis). Investigators commonly attempt to mitigate the inherent weakness of this study design by comparing their results with those reported in other case series. However, these comparisons are tenuous because such a wide range of factors affect outcomes across settings.

Finally, at the lowest level of evidence, we lose the most critical component of internal validity when it comes to clinical research-observations made on patients. Level 5 consists of expert opinion, physiology, bench (laboratory) research, or first principles (eg, theory, anatomy, physiology, biomechanics). Although bench research, theory, and foundational science are very useful in generating hypotheses about what clinical outcomes might be achieved in specific clinical interventions, it is only through testing these hypotheses on actual patients that we have substantive evidence of the actual impact on patients.

Thus, one can see that the levels-of-evidence system is an ordinal ranking scale that focuses on the most critical element of research design for intervention studies.

Levels of Evidence for Other Study Designs

Other categories of clinical research require different study designs. Optimal designs for different clinical questions are specifically outlined in the table provided at the Center for Evidence-Based Medicine and included in the article "Introduction to evidence-based practice" by Szabo and MacDermid elsewhere in this issue. These include: prognosis, diagnosis, differential diagnosis/symptom prevalence study, economic, and decision analyses. For example, a level 1b for a prognosis study is an individual inception cohort study with greater than 80% follow-up, where a clinical decision rule has been validated in a single population. Conversely, the optimal study design for a diagnostic-test study consists of a cohort study with good reference standards or a clinical-decision rule tested within one clinical center. Despite differences in the optimal study design across different types of clinical questions, certain consistencies are evident:

A systematic review of high-quality studies always provides the highest level of rigor.

An individual study using the optimal design for that type of clinical question is considered level 1.

Prospective data collection indicates higher study quality than retrospective data collection.

Expert opinion, bench research, conceptual frameworks/theories/first principles are always considered the lowest (level 5) evidence.

A variety of other rating systems have been proposed by different investigators. For example, different health service organizations have modified versions. Some of these organizations have used the term "levels of evidence" to refer to the overall state of evidence, whereas others use it for classifying individual studies. These systems may include different descriptors for five levels, the addition of different ranks (or subtypes), and even different labels.

While the intent of many of these investigators or organizations has been to simplify or customize processes to their needs, the existence of multiple systems provides an additional source of confusion. Despite this, there are many similarities across the different versions of the levels-of-evidence system. The authors prefer to use the classic five levels for ranking individual studies, as this is developed by leaders in the field, has been tested over many years, is relatively clear and comprehensive, is the most widely used system, and is easily accessible to the public (http://www.cebm.net/index.aspx?o=1025). The authors also choose to distinguish between the level of evidence of an individual study and the overall level of evidence that must be considered when making a recommendation. The latter

involves synthesis of multiple studies and sources of evidence and is discussed in the ''Grading recommendations'' section later in this article.

Critical Appraisal Tools

While it is important to understand the basic principles involved in critical appraisal, the use of tools to provide structure to the process can be invaluable. All three authors have developed critical appraisal tools and use them for teaching critical appraisal or conducting systematic reviews or meta-analyses. Critical appraisal forms developed by Law and colleagues in 1998 are examples of open-ended critical appraisal tools. There are versions for intervention/effectiveness studies and qualitative studies. These tools can be downloaded from the McMaster University Web site at http://www.srs-mcmaster.ca/Default.aspx?tabid=630. The form and associated guide lead the appraiser to consider various aspects of design through a series of open-ended questions. These questions are listed in **Boxes 1, 2**.

A second type of critical appraisal approach is used by MacDermid. These tools provide structure with quantitative (3-point) response categories that are associated with specific descriptors for each item. Specific scoring criteria for each item are provided in an accompanying interpretation guide. Forms are available for effectiveness, diagnostic tests, and psychometric (outcome measure) studies. The forms and associated guides are available from the lead author (JM) and the questions included on each scale are listed in **Boxes 3–5**. One author (DW) developed a critical appraisal tool for prognostic studies to conduct a meta-analysis of studies on risk of poor outcomes following whiplash. The tool used items from the literature and other scales to derive the criteria judged most appropriate for this specific context (items in **Box 6**).

A variety of critical appraisal tools have been developed, and there is no clear indication which of them is best. There is debate amongst methodologists about the relative benefit of using customized critical appraisal tools or generic ones when conducting systematic reviews. For the purposes of improving your critical appraisal skills, it is important to discuss and compare your results with those derived by others. For example, the text *Evidence-Based Medicine: how to practice and teach EBM*, now comes with a CD containing a variety of examples from different types of articles and different disciplines.[10] The items used in the PEDro scale are sometimes used for critical appraisal in other circumstances (http://www.pedro.fhs.usyd.edu.au/scale_item.html).

Many systematic reviews use the Jadad scale.[11] There are potential problems when using this scale to evaluate studies in hand surgery/therapy. First, two of the items relate to randomization, two relate to double blinding, and one relates to description of withdrawals and dropouts. Because hand surgery and hand therapy interventions do not easily lend themselves to double blinding, most studies fare poorly in quality ratings on this scale. For example, in a review of 2,169 published surgical trials in the *Journal of Bone and Joint Surgery* over a 10-year period, only 3% were randomized (n = 64). Of these, the overall mean study quality was 1.7/5.[12] Although a brief scale is preferable, there has been concern about the lack of comprehensive coverage of methodologic quality.[13] Second, there is a generalized concern about reliability of the scale,[12,14] especially among orthopedic surgeons.[12]

A systematic review addressed 120 different critical appraisal tools appearing in the literature.[1] This review found substantial variation between instruments in scope, structure, and scoring. **Table 1** provides additional Web sites that provide access to a variety of critical appraisal forms, and outline their purpose and number of items. Hand surgeons and therapists may wish to avoid scales designed only for use with RCTs (especially those that focus on blinding issues), as these will apply to only a small subset of the evidence currently available in the literature.

DO THE RESULTS APPLY TO MY PATIENT?

Once you decide that the conclusions within a given study are likely to be true, then you can move to the decision about relevance to your patient. You want to generalize results found within research studies for your patient. The basic question here is, ''were the patients/circumstances in the study sufficiently similar to mine that my patient could reasonably expect a similar outcome?'' You should know which aspects of your patient (disease, comorbidity, cultural, psychosocial, family, and so forth) will affect the outcomes of your test/intervention, and whether these were represented on the studied patients. Ideally, subgroup analyses within RCTs will highlight differential expectations for different subgroups.

You must also evaluate your own beliefs, skills, and circumstances to determine if they can reproduce the interventions studied in the literature. Critically evaluate your own expertise, equipment, staff, and setting: are there important differences,

Box 1
Critical review form—qualitative studies (Version 2.0)

Citation

Study purpose

1. Was the purpose and/or research question stated clearly? Outline the purpose of the study and/or research question.

Literature

2. Was relevant background literature reviewed?
3. Describe the justification of the need for this study. Was it clear and compelling?
4. How does the study apply to your practice and/or to your research question?
5. Is it worth continuing this review?

Study design

6. What was the design? Was the design appropriate for the study question? (ie, rationale) Explain.
7. Was a theoretic perspective identified? Describe the theoretic or philosophical perspective for this study: for example, researcher's perspective.
8. Describe the method(s) used to answer the research question. Are the methods congruent with the philosophical underpinnings and purpose?

Sampling

9. Was the process of purposeful selection described? Describe sampling methods used. Was the sampling method appropriate to the study purpose or research question?
10. Was sampling done until redundancy in data was reached? Are the participants described in adequate detail? How is the sample applicable to your practice or research question? Is it worth continuing?
11. Was informed consent obtained?

Data collection

12. Describe the context of the study. Was it sufficient for understanding of the "whole" picture?
13. What was missing and how does that influence your understanding of the research?
14. Do the researchers provide adequate information about data collection procedures (eg, gaining access to the site, field notes, training data gatherers)? Describe any flexibility in the design and data collection methods.

Data analyses

15. Describe method(s) of data analysis. Were the methods appropriate? What were the findings?
16. Describe the decisions of the researcher re: transformation of data to codes/themes. Outline the rationale given for development of themes.
17. Did a meaningful picture of the phenomenon under study emerge? How were concepts under study clarified and refined, and relationships made clear? Describe any conceptual frameworks that emerged.
18. Was there evidence of the four components of trustworthiness (credibility, transferability, dependability, confirmability)?
19. For each of the components of trustworthiness, identify what the researcher used to ensure each.
20. What meaning and relevance does this study have for your practice or research question?

Conclusions and implications

21. What did the study conclude? Were the conclusions appropriate given the study findings?
22. What were the main limitations of the study?
23. What were the implications of the findings for occupational therapy (practice and research)?

The full form and guide of this questionnaire (as well as an adapted word version) are available from: www.srs-mcmaster.ca/ResearchResourcesnbsp/CentreforEvidenceBasedRehabilitation/EvidenceBasedPracticeResearchGroup/tabid/630/Default.aspx

Courtesy of the Evidence-Based Practice Research Group, McMaster University, Hamilton, ON; with permission. Copyright © 1998.

and if so, how might these modify your plan or expectations? Expert surgeons may achieve excellent outcomes with a complicated procedure they have performed hundreds of times, but novice surgeons are unlikely to get similar outcomes. This is particularly true for more complex surgical skills, such as arthroscopic techniques. Similarly, some specialized rehabilitation therapies are highly effective with advanced training but similar outcomes may not be achieved without the same level of training and experience. This is particularly true for more complex technical skills, such as manual therapies or complicated orthotic devices.

Box 2
Critical review form—quantitative studies

Citation

Study purpose

1. Was the purpose and/or research question stated clearly? Outline the purpose of the study and/or research question.

Literature

2. Was relevant background literature reviewed?
3. Describe the justification of the need for this study.

Design

4. Describe the study design. Was the design appropriate for the study question? (eg, for knowledge level about this issue, outcomes, ethical issues, and so forth).
5. Specify any biases that may have been operating and the direction of their influence on the results.

Sample

6. Was the sample described in detail (who, characteristics, how many, how was sampling done?) If more than one group, was there similarity between the groups?
7. Was the sample size justified?
8. Describe ethics procedures. Was informed consent obtained?

Outcomes

9. Were the outcome measures reliable?
10. Were the outcome measures valid?
11. Specify the frequency of outcome measurement (ie, pre-, post-, follow-up), the outcome areas and list the measures that were used.

Intervention

12. Was the intervention described in detail?
13 Provide a short description of the intervention (focus, who delivered it, how often, setting). Could the intervention be replicated in practice?
14. Was contamination avoided?
15. Was cointervention avoided?

Results

16. Were results reported in terms of statistical significance?
17. What were the results? Were they statistically significant (ie, $P<.05$)? If not statistically significant, was study big enough to show an important difference if it should occur? If there were multiple outcomes, was that taken into account for the statistical analysis?
18. Were the analysis methods appropriate?
19. What was the clinical importance of the results? Were differences between groups clinically meaningful? (if applicable)
20. Did any participants drop out from the study? Why? (Were reasons given and were drop-outs handled appropriately?)

Conclusions and implications

21. What did the study conclude?
22. What are the implications of these results for practice? What were the main limitations or biases in the study?
23. Were the conclusions appropriate given the study methods and results?

The full form and guide of this questionnaire (as well as an adapted word version) are available from: www.srs-mcmaster.ca/ResearchResourcesnbsp/CentreforEvidenceBasedRehabilitation/EvidenceBasedPracticeResearchGroup/tabid/630/Default.aspx

Courtesy of the Evidence-Based Practice Research Group, McMaster University, Hamilton, ON; with permission. Copyright © 1998.

Grading Recommendations

While the levels-of-evidence system provides the user with a relative level of confidence in the results of individual study findings, making practice recommendations based on all available evidence in an area is often challenging for the novice evidence-based practitioner, as less attention has been directed at this process. The process involves examining multiple studies to make overall recommendations. Grades of A to D, which focused primarily on the level of evidence, have been commonly used (http://www.cebm.net/index.aspx?o=1025). One disadvantage of this system is that it focuses primarily on the nature of the evidence and does not consider other factors that would influence the strength of recommendations. Perhaps for this reason, a variety of systems and scales have been developed for grading recommendations. A further complication is that these recommendation scales have varied widely across different organizations, as organizations try to develop systems that meet their individual needs. For example, some have altered the terminology, some prefer visual indicators, and some include different conceptual components (eg, balance of risk and harm or costs) in the recommendation process.

Box 3
Critical appraisal of study design for psychometric articles evaluation items (outcome measure research)

Evaluation criteria

Study question

1. Was the relevant background research cited to define what is currently known about the psychometric properties of the measures under study, and the need or potential contributions of the current research question?

Study design

2. Were appropriate inclusion/exclusion criteria defined?
3. Were specific psychometric hypotheses identified?
4. Was an appropriate scope of psychometric properties considered?
5. Was an appropriate sample size used?
6. Was appropriate retention/follow-up obtained? (Studies involving retesting or follow-up only)

Measurements

7. Documentation: Were specific descriptions provided or referenced that explain the measures and its correct application/interpretation (to a standard that would allow replication)?
8. Standardized methods: Were administration and application of measurement techniques within the study standardized and did they consider potential sources of error/misinterpretation?

Analyses

9. Were analyses conducted for each specific hypothesis or purpose?
10. Were appropriate statistical tests conducted to obtain point estimates of the psychometric property?
11. Were appropriate ancillary analyses were done to describe properties beyond the point estimates (Confidence intervals, benchmark comparisons, standard error of measurement/minimal important difference)?

Recommendations

12. Were the conclusions/clinical recommendations supported by the study objectives, analysis and results?

Total score % (sum of subtotals/24*100) is based on criteria met from the rating guide and scored as 2,1, or 0 depending on compliance with standards.

This Box lists the criteria rating the quality of a study addressing the psychometric properties of an outcome measure. The full form and guide are available from the author or in the textbook *Evidence-Based Rehabilitation*.

For example, an overall rating of evidence across different studies is used by some medical groups and contains just four levels (obtained from http://www.cochranemsk.org/review/writing): platinum, gold, silver, and bronze. This system gives qualitative ratings based on number and quality (based on two or three key criteria), as listed below. Note that two of the four levels require RCTs. Within this system, case series are demoted to the lowest level of evidence, with expert opinion and bench research, and there is less differential between cohort/case-controlled designs. For hand surgery, this system is likely to rank most current evidence at the lowest level. The authors do not recommend this option for surgery, primarily because the authors believe a better approach is evolving to consensus (see the Grades Of Recommendation Assessment, Development and Evaluation Working Group or GRADE).

Platinum level

The platinum ranking is given to evidence that meets the following criteria as reported in a published systematic review that has at least two individual controlled trials, each satisfying the following:

Sample sizes of at least 50 per group. If they do not find a statistically significant difference, they are adequately powered for a 20% relative difference in the relevant outcome.

Blinding of patients and assessors for outcomes;

Handling of withdrawals greater than 80% follow-up—imputations based on methods such as last observation carried forward acceptable;

Concealment of treatment allocation.

Gold level

The gold ranking is given to evidence if at least one RCT meets all of the following criteria as reported:

Sample sizes of at least 50 per group. If they do not find a statistically significant difference, they are adequately powered for

Box 4
Criteria for evaluation of quality of an intervention study

Evaluation criteria

Study question

1. Was the relevant background work cited to establish a foundation for the research question?

Study design

2. Was a comparison group used?
3. Was patient status at more than 1 time point considered?
4. Was data collection performed prospectively?
5. Were patients randomized to groups?
6. Were patients blinded to the extent possible?
7. Were treatment providers blinded to the extent possible?
8. Was an independent evaluator used to administer outcome measures?

Subjects

9. Did sampling procedures minimize sample/selection biases?
10. Were inclusion/exclusion criteria defined?
11. Was an appropriate enrollment obtained?
12. Was appropriate retention/follow-up obtained?

Intervention

13. Was the intervention applied according to established principles?
14. Were biases due to the treatment provider minimized (ie, attention, training)?
15. Was the intervention compared with an appropriate comparator?

Outcomes

16. Was an appropriate primary outcome defined?
17. Were appropriate secondary outcomes considered?
18. Was an appropriate follow-up period incorporated?

Analysis

19. Was an appropriate statistical test(s) performed to indicate differences related to the intervention?
20. Was it established that the study had significant power to identify treatment effects?
21. Was the size and significance of the effects reported?
22. Were missing data accounted for and considered in analyses?
23. Were clinical and practical significance considered in interpreting results?

Recommendations

24. Were the conclusions/clinical recommendations supported by the study objectives, analysis and results?

Total Quality Score based on sum of above (2,1, or 0 per item) =/48.
Level of Evidence (Sackett) 1 □ 2 □ 3□ 4□ 5□.
This figure lists the criteria rating the quality of a study addressing effectiveness. It can be used for all levels of studies. The reviewer is also given a checkbox to mark the level according to the classic levels of evidence rating system. The items are scored 2,1,0. The full form and guide are available from the author or in textbook *Evidence-Based Rehabilitation*.

Courtesy of Joy C. MacDermid, BScPT, PhD, London, ON.

a 20% relative difference in the relevant outcome.

Blinding of patients and assessors for outcomes;

Handling of withdrawals greater than 80% follow-up—imputations based on methods such as last observation carried forward acceptable;

Concealment of treatment allocation.

Silver level

The silver ranking is given to evidence if a systematic review or randomized trial does not meet the above criteria. Silver ranking would also include evidence from at least one study of nonrandomised cohorts who did and did not receive the therapy or evidence from at least one case-controlled study. A randomized trial with a ''head-to-head'' comparison of agents is considered silver level ranking unless a reference is provided to a comparison of one of the agents to placebo showing at least a 20% relative difference.

Bronze level

The bronze ranking is given to evidence if there is at least one high-quality case series without controls (including simple before and after studies in which the patient acts as their own control) or if it is derived from expert opinion based on clinical experience without reference to any of the

Box 5
Items for evaluating quality of diagnostic tests

Evaluation criteria

1. Was there an independent, blind comparison with a reference standard test?
2. Was the reference standard/true diagnosis selected a gold standard or reasonable alternative?
3. Was the reference standard applied to all patients?
4. Did the actual cases include an appropriate spectrum of severity?
5. Were the noncases patients who might reasonably present for differential diagnosis?
6. Did the noncases include an appropriate spectrum of patients with alternate diagnoses?
7. Did the study have an adequate sample size?
8. Was the description of the test maneuver described insufficient detail to permit replication?
9. Were exact criteria for interpreting the test results provided?
10. Was the reliability of the test procedures documented?
11. Were the number of positive and negative results reported for both cases and noncases?
12. Were appropriate statistics (sensitivity, specificity, likelihood ratios) presented?
13. If the test required an element of examiner interpretation were the qualifications and skills of the examiner described (if n/a leave blank)
14. Were the training, skills and experience of the examiner appropriate to the test conducted? (if n/a leave blank)

foregoing (for example, argument from physiology, bench research or first principles).

In fact, there is little consistency across these rating systems and limitations have been noted.[15] This lack of consistency can make it difficult for the inexperienced evidence-based practitioner to understand how they should deal with multiple pieces of information. This is particularly problematic when groups try to develop evidence-based clinical practice guidelines where the strength and wording of the recommendations are a key output. In response to this concern, an international group, the GRADE Working Group, focused on development of a system that could be used to grade the quality of evidence and strength of

Box 6
Items included in a rating scale for determining the quality of prognostic (cohort) studies

Sampling

1. Were sample characteristics clearly stated?
2. Were the characteristics of the refusers stated and were differences between refusers/acceptors investigated?
3. Was the source population described?
4. Were the subjects recruited within a reasonably narrow time-frame?

Methodology

5. Was the exposure to the prognostic factor(s) captured using valid and reliable instruments?
6. Were the investigators who captured outcome blinded to the presence/absence of prognostic factors?
7. Did follow-up occur at the same point post-injury for all subjects?
8. If the patients received intervention during the study, was it standardized, or was the effect of intervention statistically controlled for?
9. Is the attrition rate acceptable?
10. Is there evidence that subjects lost to follow-up were similar on baseline characteristics to those who completed the study?

Analysis

11. Are appropriate univariate crude estimates presented?
12. Are appropriate multivariate analysis techniques employed?
13. Is the sample size large enough for the number of variables investigated?
14. Have the authors controlled for important confounders, either through stratification or statistical covariation?
15. Was data manipulation appropriate?

Results

16. Were the results for prognostic factors presented in a clear and understandable fashion?
17. Were the results for the main outcomes presented in a clear and understandable fashion?

This table contains items included in the critical appraisal tool developed by Walton and Pretty for a meta-analysis on prognostic factors in acute whiplash. The full form and user's guide are available from DW. The inter-rater reliability for the tool overall was 0.81, ranging from 0.44 (Q14) to 1.00 (Q3,9,10,12,17). (Joy C. MacDermid, PhD, unpublished data, 2008.)

Table 1
Online critical appraisal/recommendations tools

Web Site URL	Type of Studies Evaluated	Number of Items
Appraisal of Guidelines for Research & Evaluation[a] http://www.agreecollaboration.org/	Clinical practice guidelines	23
Best Evidence Topics http://www.bestbets.org/links/BET-CA-worksheets.php	Diagnostic test	29
	Economic analysis	34
	Prognosis	37
	Systematic review	33
	Qualitative research	40
	Clinical practice guidelines	32
Center for Evidence Based Emergency Medicine http://www.ebem.org/analyse.html	Treatment effectiveness	13
	Prognosis	10
	Diagnostic test	11
	Systematic review	12–15
Center for Evidence Based Medicine, Oxford http://www.cebm.net/critical_appraisal.asp	Treatment effectiveness	11
	Prognosis	10
	Diagnostic test	8
	Economic analysis	14
	Systematic review	10
	Clinical practice guidelines	18
Center for Health Evidence[a] http://www.cche.net/usersguides/main.asp	Treatment effectiveness	12
	Diagnostic test	9
	Prognosis	9
	Clinical practice guideline	10
	Economic analysis	10
	Qualitative research	8
Critical Appraisal Skills Program[a] http://www.phru.nhs.uk/Pages/PHD/resources.htm	Diagnostic test	12
	Qualitative study	10
	Economic analysis	10
	Systematic review	10
Evidence Based Medicine, Alberta[a] http://www.med.ualberta.ca/ebm/ebm.htm	Treatment effectiveness	11
	Prognosis	10
	Diagnostic test	10
	Economic analysis	10
	Systematic review	11
	Clinical practice guidelines	11
Evidence Based Medicine, Duke http://www.mclibrary.duke.edu/subject/ebm?tab=appraising&extra=worksheets	Treatment effectiveness	12
	Prognosis	9
	Diagnostic test	9
	Qualitative study	7
	Economic analysis	10
	Systematic review	10
	Clinical practice guidelines	4
Health Care Practice Research & Development Unit http://www.fhsc.salford.ac.uk/hcprdu/critical-appraisal.htm	Treatment effectiveness	51
	Qualitative study	44
	Economic analysis	68
McMaster—School of Rehabilitation Science[a] http://www.srs-mcmaster.ca/ResearchResourcesnbsp/CentreforEvidenceBasedRehabilitation/EvidenceBasedPracticeResearchGroup/tabid/630/Default.aspx	Treatment effectiveness	15
	Qualitative study	27

(*continued on next page*)

Table 1
(*continued*)

Web Site URL	Type of Studies Evaluated	Number of Items
Quality of Reporting of Meta-analyses http://www.consort-statement.org/QUOROM.pdf	Systematic review	17
School of Health and Related Research (ScHARR), University of Sheffield[a] http://www.shef.ac.uk/scharr/sections/ir/links	Systematic review Qualitative study	10 10
GRADE Working Group http://www.gradeworkinggroup.org/	Documents and free software supporting use of GRADE approach to making recommendations	—

[a] Has guide to interpretation.

recommendations in a method that balanced simplicity and clarity.[15–23] A number of articles have subsequently been published on their consensus of how to rate quality of evidence and strength of recommendations.

The GRADE system classifies evidence in one of four levels: high, moderate, low, or very low. Evidence is considered high quality if further research is very unlikely to change our confidence in the estimate of the effects. Moderate quality is present if further research is likely to have an important impact on our confidence in the estimate of the effect and may actually change the estimate. If further research is very likely to have an important impact on our confidence in the estimate of the effect and is likely to change the estimate, then it is considered low quality. If the estimated effect is very uncertain, the quality of evidence is considered very low quality.

Evidence based on RCTs starts off as high quality evidence but can decrease to a lower level if there are significant study limitations, inconsistency of results, indirectness of the evidence, imprecision, or reporting bias. Observational studies start off as low quality that can be graded upwards if the magnitude of the treatment effect is very large, or if all plausible sources of confounding have been identified and controlled, or if there is evidence of a dose-response gradient. Because various groups have in the past conveyed quality of evidence in different formats, there are nonspecific recommendations for using letters, numbers, symbols, and words to communicate grades of evidence. For example, high quality evidence can be labeled using the word "high," or alternatively, the number "1," the letter "A," the full darkened circle symbol "●" or the star approach "★ ★ ★ ★."[24] There have been suggestions that clinicians respond more favorably to symbols than they do to numbers or letters.[25]

There are four factors that influence the strength of the recommendation. Of these, the quality of evidence is one factor. The higher the quality of evidence, the more likely a strong recommendation is warranted. Strong recommendations tend to use statements that are in that category of "definitely should do" or "definitely should not do." Secondly, the balance between desirable and undesirable effects is considered. The larger the difference between desirable and undesirable effects, the more likely a strong recommendation is warranted. The narrower this gradient, the more likely a weak recommendation is warranted. For example, in hand surgery there has been considerable controversy about the relative role of endoscopic versus open carpal tunnel release. Despite numerous RCTs and systematic reviews and meta-analyses, there remains controversy. Part of this controversy is related to the fact that even in well-designed studies, the differential effects are narrow. Therefore, regardless of the quality of evidence, only weak recommendations should be made on use of these two interventions.

Weak recommendations tend to be more in the category of "probably should (or should not) do," and allow more latitude for the individual practitioner to consider local circumstances as potentially outweighing the small differential in potential effectiveness. For example, a more appropriate recommendation for carpal tunnel surgery might be that surgeons should examine a summary of evidence comparing endoscopic and open carpal tunnel release and be aware that there is strong evidence of small differential

outcomes between the two procedures. Differential outcomes include the potential for faster return to work and slightly higher (but still low) risk of complications with endoscopic procedures, as well as small differential recovery in physical impairments. Surgeons should be prepared to consider and discuss with patients their own experience, expertise, and circumstances, and the patient's values and preferences to choose the best surgical option.

A third factor that influences the strength of recommendation concerns patient values and preferences. The more variable or uncertain values and preferences are, the more likely a weak recommendation should be used. Finally, costs and resource allocation can be considered. The higher the cost of an intervention (particularly when considering cost:benefit ratio) the less likely a strong recommendation is warranted.

The GRADE system offers two grades of recommendation—strong and weak—and an option for no specific recommendation when the trade-offs are equally balanced or uncertain. When the desirable effects of intervention clearly outweigh the undesirable effects (or clearly do not), then strong recommendations are warranted. When trade-offs are less certain, the quality of evidence is lower, values are uncertain, or resource use is a concern, then the relative desirability may be less certain. Thus, the grades are 1 (strong), 2 (weak), or 0 (no specific recommendation).

Practitioners may find it helpful to record their evidence-based conclusions in short one-page summaries. The Center for Evidence-Based Medicine provides a free downloadable "CAT-maker" that calculates the number needed-to-treat from study data and allows one to summarize the available evidence on a one-page summary in a standardized fashion that can be stored in a personal file of evidence reviews (http://www.cebm.net/index.aspx?o=1216). Once the hand surgeon or therapist is clear in his or her own mind about the overall balance of the research evidence and has formulated that into a recommendation that they are comfortable with, they can then move to the process of integrating their view with the patients. The process of incorporating patient-centered care or shared decision-making with the evidence and clinical experience has been highlighted in other articles in this issue. A critical analysis of the literature and formulation of clear recommendations makes it easier to communicate more effectively with patients during these discussions. See Appendix 1 for useful Web links regarding the ideas discussed in this article.

APPENDIX 1: USEFUL WEB LINKS

http://bmjupdates.mcmaster.ca/index.asp (BMJ updates): A free push service (through McMaster University and BMJ) that delivers customized appraised research by email in the content areas and frequency you request.

http://davidmlane.com/hyperstat/index.html: Free online statistics textbook.

http://nilesonline.com/stats: Easy reading statistics textbook.

http://statpages.org: Free online statistics calculations.

http://www.cebm.net/critical_appraisal.asp (Center for Evidence-Based Medicine, Oxford): This site provides a database of critically appraised topics and tools.

http://www.consort-statement.org (Homepage): The CONSORT statement lays out a number of guidelines for conducting good RCTs, which are essential for sound systematic reviews. The homepage has more detailed information and updates on current work.

http://www.otseeker.com (OTseeker): This is a searchable database that provides abstracts and ratings of RCTs and systematic reviews relevant to occupational therapy.

http://www.pedro.fhs.usyd.edu.au (PEDro): This is a searchable Physiotherapy Evidence Database that provides bibliographic details, abstracts, and ratings of RCTs, systematic reviews, and evidence-based clinical practice guidelines in physiotherapy.

http://www.sportsci.org/resource/stats/index.html: An excellent primer or refresher to many aspects of statistics, complied and created by New Zealander William Hopkins.

REFERENCES

1. Katrak P, Bialocerkowski AE, Massy-Westropp N, et al. A systematic review of the content of critical appraisal tools. BMC Med Res Methodol 2004;4:22.
2. Mamdani M, Sykora K, Li P, et al. Reader's guide to critical appraisal of cohort studies: 2. Assessing potential for confounding. BMJ 2005;330(7497):960–2.
3. Montori VM, Bhandari M, Devereaux PJ, et al. In the dark: the reporting of blinding status in randomized controlled trials. J Clin Epidemiol 2002;55(8):787–90.

4. Devereaux PJ, Bhandari M, Montori VM, et al. Double blind, you are the weakest link–goodbye!. Equine Vet J 2005;37(6):557–8.
5. Devereaux PJ, Bhandari M, Montori VM, et al. Double blind, you have been voted off the island! Evid Based Ment Health 2002;5(2):36–7.
6. MacDermid JC, Richards RS, Roth JH, et al. Endoscopic versus open carpal tunnel release: a randomized trial. J Hand Surg [Am] 2003;28(3):475–80.
7. Boyd KU, Gan BS, Ross DC, et al. Outcomes in carpal tunnel syndrome: symptom severity, conservative management and progression to surgery. Clin Invest Med 2005;28(5):254–60.
8. Bain GI, Pugh DM, MacDermid JC, et al. Matched hemiresection interposition arthroplasty of the distal radioulnar joint. J Hand Surg [Am] 1995;20(6):944–50.
9. Amadio PC, Higgs P, Keith M. Prospective comparative clinical trials in the journal of hand surgery (American). J Hand Surg [Am] 1996;21(5):925–9.
10. Sackett DL, Straus SE, Richardson WS, et al. Evidence-based medicine. How to practice and teach EBM. 2nd edition. Toronto: Churchill Livingstone; 2000.
11. Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials 1996;17(1):1–12.
12. Bhandari M, Richards RR, Sprague S, et al. Quality in the reporting of randomized trials in surgery: is the Jadad scale reliable? Control Clin Trials 2001; 22(6):687–8.
13. Bhogal SK, Teasell RW, Foley NC, et al. The PEDro scale provides a more comprehensive measure of methodological quality than the Jadad scale in stroke rehabilitation literature. J Clin Epidemiol 2005;58(7):668–73.
14. Clark HD, Wells GA, Huet C, et al. Assessing the quality of randomized trials: reliability of the Jadad scale. Control Clin Trials 1999;20(5):448–52.
15. Atkins D, Eccles M, Flottorp S, et al. Systems for grading the quality of evidence and the strength of recommendations I: critical appraisal of existing approaches The GRADE Working Group. BMC Health Serv Res 2004;4(1):38.
16. Atkins D, Best D, Briss PA, et al. Grading quality of evidence and strength of recommendations. BMJ 2004;328(7454):1490.
17. Atkins D, Briss PA, Eccles M, et al. Systems for grading the quality of evidence and the strength of recommendations II: pilot study of a new system. BMC Health Serv Res 2005;5(1):25.
18. Guyatt G, Baumann M, Pauker S, et al. Addressing resource allocation issues in recommendations from clinical practice guideline panels: suggestions from an American College of Chest Physicians task force. Chest 2006;129(1):182–7.
19. Guyatt G, Vist G, Falck-Ytter Y, et al. An emerging consensus on grading recommendations? ACP J Club 2006;144(1):A8–9.
20. Guyatt G, Gutterman D, Baumann MH, et al. Grading strength of recommendations and quality of evidence in clinical guidelines: report from an American College of Chest Physicians task force. Chest 2006;129(1):174–81.
21. Guyatt GH, Oxman AD, Kunz R, et al. Going from evidence to recommendations. BMJ 2008; 336(7652):1049–51.
22. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008; 336(7650):924–6.
23. Guyatt GH, Oxman AD, Kunz R, et al. Incorporating considerations of resources use into grading recommendations. BMJ 2008;336(7654):1170–3.
24. Schunemann HJ, Best D, Vist G, et al. Letters, numbers, symbols and words: how to communicate grades of evidence and recommendations. CMAJ 2003;169(7):677–80.
25. Akl EA, Maroun N, Guyatt G, et al. Symbols were superior to numbers for presenting strength of recommendations to health care consumers: a randomized trial. J Clin Epidemiol 2007;60(12): 1298–305.