OVERVIEW

# Applying Evidence Standards to Rehabilitation Research

**Authors:**

Mark V. Johnston, PhD
Mark Sherer, PhD
John Whyte, MD, PhD

**Affiliations:**

From Outcomes Research, Kessler Medical Rehabilitation Research and Education Corporation, West Orange, New Jersey (MVJ); Physical Medicine and Rehabilitation, University of Medicine and Dentistry of New Jersey/New Jersey Medical School, Newark, New Jersey (MVJ); Neuropsychology, Methodist Rehabilitation Center, Jackson, Mississippi (MS); Neurology, University of Mississippi Medical Center, Jackson, Mississippi (MS); Moss Rehabilitation Research Institute, Albert Einstein Healthcare Network, Philadelphia, Pennsylvania (JW); and the Department of Rehabilitation Medicine, Jefferson Medical College, Thomas Jefferson University, Philadelphia, Pennsylvania (JW).

**Correspondence:**

All correspondence and requests for reprints should be addressed to Mark V. Johnston, PhD, KMRREC, Research (East), 1199 Pleasant Valley Way, West Orange, NJ 07079.

**Disclosures:**

Supported, in part, by the Henry H. Kessler Foundation, West Orange, New Jersey.

0894-9115/06/8504-0292/0
*American Journal of Physical Medicine & Rehabilitation*
Copyright © 2006 by Lippincott Williams & Wilkins

DOI: 10.1097/01.phm.0000202079.58567.3b

## ABSTRACT

Johnston M, Sherer M, Whyte J: Applying evidence standards to rehabilitation research. Am J Phys Med Rehabil 2006;85:292–309.

**Objective:** To describe evidence grading methods employed in the systematic reviews in this special series of articles. To provide an overview of results of these reviews to critique the quality of rehabilitation research. To identify issues in the application of evidence grading methods to rehabilitation.

**Design:** Conceptual review of evidence-based practice and evidence grading methods. English-language research studies on rehabilitation of persons with spinal cord injury, traumatic brain injury, and burn for the 5-yr period of 1999–2004 were reviewed using methods of the American Academy of Neurology supplemented by Cochrane criteria and summarized.

**Results:** Rehabilitation has a shortage of high-level studies. The number of level 1 treatment studies was quite limited (five in spinal cord injury, 15 in traumatic brain injury, 12 in burn rehabilitation), as was the number of level 2 studies (26, 4, and 2, respectively). Despite the large number of correlational studies published, the number of high-level (1 or 2) diagnostic and prognostic studies was surprisingly limited (34, 11, and 5, respectively). The rate of production of high-level studies is rapidly increasing. Problems were encountered in applying standard evidence criteria to complex issues encountered in some studies, suggesting limitations and issues in evidence grading methodology.

**Conclusions:** Rehabilitation needs more high-level studies. Some improvements in research methodology are relatively affordable (e.g., improved blinding), whereas others are expensive (e.g., large randomized controlled trials). Lower-level investigations reporting promising results need to be followed by more definitive, higher-level trials.

**Key Words:** Evidence-Based Practice, Evidence Grading Methods, Systematic Reviews, High-Level Studies, Rehabilitation Research

Few topics in rehabilitation are as important—or as controversial—as evidence-based practice (EBP). Sackett has defined evidence-based medicine as "the integration of best research evidence with clinical expertise and patient values,"[1, p. 1] a definition that applies also to clinical practices in rehabilitation delivered by allied health professionals. The term "evidence-based" itself evokes general support: who does not attempt to apply evidence to their practice? But EBP also involves application of certain standard methods of grading strength or level of evidence based on the research design employed and emphasizing randomized controlled trials (RCTs). The result of a formal evidence review is a judgment of whether a treatment is effective, ineffective, or harmful, or a diagnostic procedure is accurate, or whether there is insufficient evidence to make these determinations. Treatment guidelines, practitioner education, clinical policies, payments, and priorities for research hinge on this judgment. Given the stakes involved, it is not surprising that controversies emerge regarding the details and substance of EBP methodologies and how the results of evidence reviews are to be applied.

## OBJECTIVES

The accompanying systematic reviews applied standard EBP methods to evaluate the quality of research on the rehabilitation of persons with spinal cord injury (SCI), traumatic brain injury (TBI), and burn. In this process, much was learned—not only about the quality of rehabilitation research but also about the application of EBP methods to complex issues in rehabilitation. This article will:

- Provide an introduction to EBP and evidence grading methods.
- Summarize the methods employed in the reviews on SCI, TBI, and burn rehabilitation.
- Provide an overview of results of these reviews and what they say about the quality of rehabilitation research. Specific results for SCI, TBI, and burn rehabilitation are reserved for the actual reviews.
- Point out lessons learned and issues encountered in the application of standard EBP methods. We take the position that extant evidence grading methods are sound but are not without limitations or ambiguities, which deserve scrutiny, debate, and eventual resolution.
- Note implications for the development of rehabilitation research.

The issues discussed should help prevent misunderstanding and misapplication of EBP methods in rehabilitation. Throughout, we attempt to draw out definite, noncontroversial lessons, balancing findings on the limitations of rehabilitation research against demonstrated promise and accomplishments.

## EBP AND PRINCIPLES OF EVIDENCE GRADING

The terms "levels of evidence" or "strength of evidence" refer to systems for classifying scientific studies for validity threats and possible bias using a hierarchy based on the research design and the data presented. Evidence grading is based on expert knowledge of the typical validity threats associated with various research designs. This knowledge is applied to the particular question and study under review. The major purpose of rating level of evidence is to assess the likelihood of bias with regard to a study conclusion.[2]

In the past, it was common to read reviews based on a subjective weighting of the strength of studies. Results of such reviews, consequently, depend as much on the subjective opinion of the author as on evidence or data. Years ago, it was common to read reviews that summarized evidence by counting the number of studies reporting positive and negative results and basing the conclusion on the balance of the two. The problem with this approach is that some studies are vastly stronger than others. A single rigorous definitive study may reasonably overturn the conclusions of a dozen weak studies with major uncontrolled validity threats. Beginning in the early 1980s, systems to weight or evaluate the strength of studies began to be developed and applied to reviews of the medical literature.[1,3] A number of evidence evaluation methods have since been developed.[4] Some systems employ three levels of quality, whereas others employ eight or more. RCTs are at the top of the hierarchy of research design for therapeutic, treatment, or intervention studies, but many other criteria also need to be examined to grade the quality of a research study and its susceptibility to bias. Confidence intervals or magnitude of effect sizes, possible measurement biases, number of consistent studies, and many other factors are important.[5] In each evidence synthesis process, reviewers should select a level-of-evidence system designed for the general purpose or type of study, as criteria for therapeutic or treatment studies differ from those for diagnostic, prognostic, or predictive studies. (Throughout this work we will use the terms therapeutic, treatment, and intervention interchangeably, and the terms prognosis and prediction, as the terms are so logically similar. We use the term "level of evidence" to emphasize that evidence grading is hierarchical, but the term is essentially synonymous with "class of evidence.")

Standard methods of evidence grading are designed to answer clinically applicable, practical questions. They address the question, "What is the direct evidence for what works in practice?" Basic and laboratory science, descriptive research, measure development, and correlational research are necessary and may be tremendously insightful and valuable, but such work typically does not provide good evidence regarding what works in practice to alleviate health problems or improve the function and quality of life of patients or persons with disabilities. In sum, grading studies by level of evidence shows the clinical applicability and degree of certainty of research results one would expect, given the research design.

EBP methodology grades the strength of recommendations to use (or not use) a procedure to the strength of the evidence for the benefit (or harm) produced by the procedure.

- Very strong evidence (e.g., two consistent level 1 studies) is needed to support a conclusion that treatment is "established as effective" for a patient problem and a recommendation that the treatment "should be employed" (a "level A" recommendation).[2, p. 39]
- Good but not conclusive evidence (e.g., one level 1 study and one level 2 study) is needed to support a conclusion that a treatment is "probably" effective and "should be considered" (a "level B" recommendation).
- Still weaker (equivocal) evidence may support only a recommendation that the procedure "may be considered" (a "level C" recommendation).

An evidence-based review may also result in the conclusion that evidence is insufficient to determine the effectiveness of a procedure. In that case, only recommendations for future research are clearly supported by the review.

## Evidence Grading Principles

There are many possible threats to the validity of inferences from a research study. In the sections below, we will first discuss the importance of specifying a clear question and general threats to study validity and later consider issues specific to treatment studies and then diagnostic/predictive studies.

### Specificity of the Treatment or Clinical Question Under Review

A clear rating of level of evidence is usually possible only if one is considering a clear question. As a rule, strength of evidence can be clearly graded only with regard to a particular, focused clinical question (American Academy of Neurology [AAN] again). (The Cochrane handbook notes that broad questions can also be reviewed[5]: the point here is that such reviews are more likely to encounter difficulties and to provide ambiguous results.) The applicable clinical population and presenting problem to the person need to be clearly defined. The treatment or intervention must be defined along with a measurable outcome of "clinical significance" or value to the person served. The result is a specific clinical meaningful statement, such as, "Therapy X improves outcome O by Y% in patients with accurately diagnosed problem A." Specific populations and outcomes need to be defined for a predictive study as well. A specific review question is usually needed to reach a definite conclusion. A study can easily provide high-level evidence regarding one question but only weak evidence regarding a somewhat different one.

An implication is that the question, "Does rehabilitation work?" is akin to question, "Does medicine work?" Such broad questions are so vague and heterogeneous that a definite answer is not possible. On the other hand, it is possible to review a broad group of studies to characterize the strength of methodology employed using the usual criteria employed in systematic evidence grading—the procedure employed here.

### Sample Size

It is often said that rehabilitation needs larger studies. Although it is true that rehabilitation studies have often been underpowered,[6] whether a study is "large" is not the principal methodological consideration. Sample size considerations are primarily handled by consideration of statistical *confidence intervals*, which are usually wide with small samples. With a small sample, a negative result will most likely be moot: one will not know whether the treatment does not work or whether sample size was too small to detect the effect. With very small samples (e.g., $<20$), sampling is likely to be unrepresentative, and analysis of subgroups and multiple outcomes is not possible or stable statistically. It is also harder to attain stability in use of conventional (asymptotic) statistical methods with small samples. However, a positive statistically significant result in a relatively small sample also implies a strong treatment effect. So small sample sizes can lead to downgrading of level of evidence, but small studies with positive results should not be simply dismissed. To do so would be to neglect studies with highly promising (though not proven) results. These principles apply to both therapeutic and predictive studies.

### Measurement Biases

Errors in measurement can invalidate or bias results of a research study, and all systems for grading level of evidence rate one or another aspect of measurement or ascertainment bias. There are several methods by which measurement bias can be eliminated or minimized, including blinded evaluation (or even complete allocation concealment throughout the study) and use of a completely objective outcome measure (e.g., survival *vs.* nonsurvival, weight, some laboratory test values).

### Attrition Biases

Differential losses between the treatment and control group over the course of the study are another important source of potential bias. The principle is that a study should account for attrition in a way that minimizes potential bias. The Oxford Center for Evidence Based Medicine and the AAN both use a rule of thumb: losses of >20% lead to downgrading of level of a study.[2,5]

### Consistency of Results

In a systematic evidence review on a clinically or practically important topic, consistency or homogeneity of results across multiple studies is an essential consideration.

## Principles for Grading Treatment Studies

### Randomization

As stated in the Cochrane handbook/Oxford Center, the randomized experiment is the best general purpose research design for testing the efficacy of a therapy or treatment.[5] In most extant evidence grading schemes, a therapeutic study must allocate participants randomly for it to be graded as a level 1 study. There are sound reasons for this strong preference for randomized design: randomization is the only general purpose research design that ensures the pretreatment equivalence of experimental and control group participants, assuming adequate sample size; selection biases (differences in severity between experimental and control groups) are probabilistically controlled so that differences in outcome to persons treated can be attributed to the treatment rather to differences in severity or nature of participants. Other validity threats (e.g., regression artifacts, history effects) can also be lessened by randomization.[7–9] Although there are other research designs that can also control for selection biases, they provide rigorous conclusions only in specific circumstances. We will discuss the significance of omission of these important but more complicated designs after summarizing direct results of the reviews.

### Other Validity Threats for Therapeutic Studies

Additional considerations for rating strength of evidence for therapeutic studies include[2,5]:

- A clearly or properly defined clinical group and outcome measure.
- Clearly defined intervention.
- Complete *allocation concealment* to avoid not only measurement biases but also *treatment contamination* and other possible biases.
- Use of intent-to-treat statistical analysis methods.
- Whether the study is prospective or retrospective.
- Narrow confidence intervals for results (a function of sample size but also of effect size, reliability of measurement, and experimental control).

## Prognostic, Diagnostic, and Screening Studies

Diagnostic, screening, and prognostic (or predictive) studies are also important. The logic of these types of studies is similar in that all involve identification and computation of stable predictive relationships, not proof of the efficacy of an intervention. A diagnostic study, for instance, can be considered to be a predictive study in which the outcome is a "gold standard" criterion for diagnosis. So we consider these together.

Criteria for rating level of evidence for predictive studies differ from those for treatment studies.[2,5] Randomized control is not directly relevant to evaluation of the quality of a prognostic or predictive inference. For questions of predictive accuracy, important elements in rating a study, include the following:[2, pp. 10-11]

- *Spectrum of participants.* Whereas well-designed therapeutic studies commonly involve great care in case selection, predictive studies are more valid if they include a "broad" spectrum of patients according to AAN criteria: one must have a sufficient sample size to do a prediction or diagnosis for the full range of patients seen in clinical practice or who have the problem in the community. Cochrane criteria emphasize that predictive studies should include a sample "representative" of that in the general population or that commonly seen in clinical practice.[5] Misleading estimates results can result from prognostic or diagnostic studies of narrowly defined, nonrepresentative cases. Base rates also greatly affect diagnostic and predictive accuracy.
- *Data quality and study design.* Retrospective studies typically have poorer data quality than prospective ones.
- *Measurement quality for both predictor and outcome variables.* Bias in measurement is typically

minimized by use of completely objective measurement procedures and/or by blinding so that the outcome variable is measured without knowledge of the predictor variable.
- *Attrition bias* is a significant threat to prognosis, diagnosis, and screening studies and to treatment studies. Losses to follow-up of >20% lead to downgrading these types of studies as well.[2,5]

Accuracy of prediction is the final consideration in assessing these types of studies. The preceding criteria help one to assess the believability or likelihood of bias of claimed predictive accuracy.

It is also important to ask whether the diagnosis or prognosis is of utility in practice. A more precise diagnosis is of little practical use if the condition is not treatable anyway, and elaborate predictive models may not be employed in practice. The reviews in this supplement, however, focus on the issue of general research quality. Additional work is needed to address questions of utility in practice and the integration of research results with the values of patients and persons with disabilities.

## METHODS EMPLOYED FOR THESE REVIEWS

### Aims Differ from Typical Evidence Reviews

The aim of the reviews in this supplement was to provide an overview of the state of the science based on published research on SCI, TBI, and burn rehabilitation in recent years. To do this, the methodological quality of more recent rehabilitation research was systematically reviewed. This aim differs from that of the typical evidence review, such as those performed by Cochrane Collaboration and the AAN, which are designed to evaluate the clinical applicability of research results to a specific clinical problem and to develop recommendations for clinical practice.[3] *The current reviews are not intended to support recommendations or guidelines for clinical practice.* To draw clinical or practice implications, older studies, and more recent ones, would have to be reviewed. One would expect also that a review designed to make practice recommendations would focus on a well-defined clinical question rather than on SCI, TBI, or burn rehabilitation as a whole.

The systematic reviews that follow this Overview provide a systematic grading of the quality of rehabilitation research, defining quality in terms of typical indicators of strength of evidence for clinical application. By using systematic, common criteria, personal and professional biases are lessened. These reviews provide a more standardized evaluation of rehabilitation evidence than many previously published reviews and textbooks (cf., National Institutes of Health review of TBI, cite textbooks of rehabilitation, textbooks of SCI, TBI rehabilitation, Birch-Davis), but they do not eliminate the personal judgment of expert authors. The reviews may be seen as a mix of traditional scholarly review techniques, which can be highly insightful but involve a degree of personal opinion, and modern evidence review techniques, which are more reliable but which can miss insights.

### Choice of Evidence Grading Methodology

Having decided to aim for a systematic EB review, the question arose, which method of grading the level or quality of evidence should be employed? There are several widely respected alternatives. Probably the most commonly employed method for judging level of evidence internationally is that of the Cochrane Collaboration.[5] In the United States, methods have been developed by the Agency for Health Care Policy and Research,[10] the United States Preventive Services Task Force,[11] and other organizations. For the current review, we chose to employ the grading methodology of the AAN because its *Clinical Practice Guideline Process Manual*[2] is clear and provides an optimal level of detail. Although very similar to Cochrane methods, AAN instructions are designed for application to clinical populations that are seen by referral (as opposed to primary care or population-based samples), which is the situation in which rehabilitation services are delivered. The AAN method has been employed to develop a number of highly regarded, clinically applicable evidence-based guidelines. The American Congress of Rehabilitation Medicine's Clinical Practice Committee has adopted the AAN manual as the method of first choice for evidence grading for interdisciplinary rehabilitation.

We encountered situations in which the AAN manual in itself was insufficient. In these cases, we consulted the *Cochrane Reviewers' Handbook*[5] and employed knowledge of principles underlying evidence grading.

### Specifics of AAN Grading Method Employed

In the AAN method, risk of bias—that is, level or class of evidence—is evaluated using a four-tiered rating. "In this scheme, studies graded class I are judged to have a low risk of bias; studies graded class II are judged to have a moderate risk of bias; studies graded class III are judged to have a moderate to high risk of bias; studies graded class IV are judged to have a very high risk of bias." [2, p. 12] Expert opinion (class 4 in the AAN hierarchy) has been the main basis of past guidelines in rehabilitation, but such evidence is formally excluded from the attached reviews.

AAN criteria for rating of level of evidence of therapeutic and prognostic studies are reproduced in Table 1.[2, pp. 39–40] AAN also provides criteria for diagnostic and screening studies, which are rather similar to diagnostic criteria.

In evidence synthesis, grading is with respect to a specific clinical problem or study inference. No particular conclusion was specified in the current review, which had a broader purpose, so grading here was applied more mechanically. Difficulties or ambiguities in rating occasionally resulted. Some studies, for instance, provided level 2 evidence for one aspect of the clinical problem but level 3 or 4 for another important aspect. In these circumstances, authors and raters used their best judgment. More definite ratings are logically possible only by specifying the clinical problem more precisely.

## Search Parameters

Each review describes its own search variables and inclusion/exclusion criteria. Because the aim was to characterize recent rehabilitation research most applicable to the United States, searches were limited to English-language scientific publications in the last 5 yrs (1999 through 2004). For all three reviews, MEDLINE and PsychInfo were searched exhaustively for these years. Some authors examined other databases, and all authors also employed their own knowledge and read reference lists of prominent studies to augment the search. The search, then, is comprehensive but not exhaustive.

Although the focus was primarily on studies in the 5-yr period, authors found themselves making statements referencing previous research, formally or informally. Pre-1999 studies were covered via extant review articles on the period and author knowledge, but they were not reviewed systematically.

Although authors strove to cover as much literature as they could in a short time, one cannot

**TABLE 1** American Academy of Neurology criteria for grading diagnostic and therapeutic studies

| Rating of Prognostic Article | Rating of Therapeutic Article |
|---|---|
| Class I: Evidence provided by a prospective study of a broad spectrum of persons who may be at risk for developing the outcome (e.g., target disease, work status). The study measures the predictive ability using an independent gold standard for case definition. The predictor is measured in an evaluation that is masked to clinical presentation and, the outcome is measured in an evaluation that is masked to the presence of the predictor. All patients have the predictor and outcome variables measured. | Class I: Prospective, randomized, controlled clinical trial with masked outcome assessment, in a representative population. The following are required:<br>(a) primary outcome(s) clearly defined<br>(b) exclusion/inclusion criteria clearly defined<br>(c) adequate accounting for drop-outs and cross-overs with numbers sufficiently low to have minimal potential for bias<br>(d) relevant baseline characteristics are presented and substantially equivalent among treatment groups or there is appropriate statistical adjustment for differences. |
| Class II: Evidence provided by a prospective study of a narrow spectrum of persons at risk for having the condition, or by a retrospective study of a broad spectrum of persons with the condition compared with a broad spectrum of controls. The study measures the prognostic accuracy of the risk factor using an acceptable independent gold standard for case definition.... measured in an evaluation that is masked..... | Class II: Prospective matched group cohort study in a representative population with masked outcome assessment that meets a–d above *or* an randomized controlled study in a representative population that lacks one criterion of a–d. |
| Class III: Evidence provided by a retrospective study in which either the persons with the condition or the controls are of a narrow spectrum. The study measures the predictive ability using an acceptable independent gold standard for case definition. The outcome, if not objective, is determined by someone other than the person who measured the predictor. | Class III: All other controlled trials (including well-defined natural history controls or patients serving as own controls) in a representative population, where outcome is independently assessed, or independently derived by objective outcome measurement. |
| Class IV: Any design where the predictor is not applied in an independent evaluation *or* evidence provided by expert opinion or case series without controls. | Class IV: Evidence from uncontrolled studies, case series, case reports, or expert opinion. |

Reproduced from the *American Academy of Neurology Clinical Practice Guideline Process Manual*.[2]

claim that all important topics have been covered. Simpler functional skill training and learning-based interventions, the bread-and-butter of traditional rehabilitation, were not fully covered, nor were relevant findings in long-term care, chronic care, or disease management, or studies of programs that include SCI, TBI, or burns among other diagnostic groups. Most studies of assistive technology (AT) and environmental interventions were not reviewed. This is a significant limitation, as such interventions are common and characteristic in rehabilitation.

Rehabilitation draws from many other fields, so it is difficult to provide a truly complete definition of relevant topics for evidence review.

## RESULTS OF REVIEWS

As in health care as a whole, the rate of publication of RCTs in rehabilitation has increased rapidly in recent years. According to MEDLINE indexing (Table 2), more RCTs have been done in SCI, brain injury, and burn rehabilitation in the last 5 yrs ($n = 55$) than in all previous years ($n = 40$). Prognostic or predictive studies, however, are still far more frequent (by a factor of 4–10) than controlled trials in these groups, whether the study is labeled as “rehabilitation” or not.

Another observation is that, although SCI rehabilitation is older than TBI rehabilitation, the rate of conduct of RCTs in TBI has increased in recent years so that the total number of RCTs in TBI rehabilitation now approaches that in SCI or burn rehabilitation. For all three diagnostic groups, the literature on general medical treatment is much greater than that labeled “rehabilitation.” The distinction between rehabilitation and other forms of care is not clear. Medical rehabilitation in practice involves numerous interventions from acute medical care and from primary care and chronic care. Preceding numbers vary depending on details of the search terms, but the pattern remains the same.

### Limited Number of High-Level Studies

Despite the rapid increase in the number of rehabilitation trials, the number of RCTs and other high-level studies remains quite limited in SCI, TBI, and burn rehabilitation. As shown in Table 3 below, only five level 1 therapeutic studies were found in SCI rehabilitation, 15 in TBI, and 12 in burn rehabilitation for the 5-yr period. But the problem is not simply one of a limited number of RCTs. Many studies were downgraded for reasons other than lack of a randomized control group.

The number of studies that could even be classified as level 2 was surprisingly limited. This is an important concern because level 2 studies permit a conclusion that a treatment is “probably” effective, which is commonly sufficient to justify the treatment in practice and certainly shows that a treatment is highly promising (assuming results of the trial are positive). The need is for more high-quality studies, both level 1 and 2.

**TABLE 2** Number of randomized controlled trials (RCTs) and prognostic studies in last 5 yrs and before by diagnostic group

| Diagnostic Groups | 1998 and Before | 1999–2004 | Totals |
|---|---|---|---|
| | RCTs[a] | | |
| SCI | 146 | 105 | 251 |
| Brain injuries | 130 | 123 | 253 |
| Burns | 262 | 195 | 457 |
| | RCTs in Rehabilitation[b] | | |
| SCI | 18 | 17 | 35 |
| Brain injuries | 7 | 22 | 29 |
| Burns | 15 | 16 | 31 |
| | Prognostic Studies[a] | | |
| SCI | 866 | 781 | 1647 |
| Brain injuries | 1698 | 915 | 2613 |
| Burns | 740 | 579 | 1319 |
| | Prognostic Studies in Rehabilitation[b] | | |
| SCI | 87 | 141 | 228 |
| Brain injuries | 112 | 150 | 262 |
| Burns | 16 | 35 | 51 |

SCI, spinal cord injury.

Source: MEDLINE (PubMed). Search: human, end of 2004 and all previous years, using MeSH Terms: spinal cord injury, brain injury (inclusion of trauma or traumatic did not change count), burn, rehabilitation, and prognosis; and publication type (randomized controlled trial, no limit for prognostic studies).

[a] Without restriction to “rehabilitation.”

[b] With restriction to “rehabilitation.”

**TABLE 3** Level of rehabilitation studies documented in reviews

| Level and Type of Study | Number |
|---|---|
| SCI | |
| Level 1 therapeutic studies | 5 |
| Level 2 therapeutic studies | 26 |
| Level 1 diagnostic/prognostic studies | 9 |
| Level 2 diagnostic/prognostic studies | 25 |
| Total RCTs reviewed | 32 |
| TBI | |
| Level 1 therapeutic studies | 15 |
| Level 2 therapeutic studies | 4 |
| Level 1 diagnostic/prognostic studies | 1 |
| Level 2 diagnostic/prognostic studies | 10 |
| Total RCTs reviewed | 32 |
| Burn | |
| Level 1 therapeutic studies | 12 |
| Level 2 therapeutic studies | 2 |
| Level 1 prognostic studies[a] | 0 |
| Level 2 prognostic studies[a] | 5 |
| Total RCTs reviewed | 17 |

SCI, spinal cord injury; RCTs, randomized controlled studies; TBI, traumatic brain injury.

Source: Evidence tables of accompanying reviews.

[a] Prognostic but not diagnostic studies were reviewed in burn evidence tables.

## Methodological Limitations and Lessons from Reviews

The reasons for downgrading research studies reviews provide important lessons for future rehabilitation research.

### Issues in Therapeutic Studies

Infrequent use of randomized assignment was a very common methodological weakness of therapeutic studies in rehabilitation, but only one. Both RCTs and other therapeutic studies were downgraded because of:

- Lack of blinded or objective outcome measures.
- Loss to follow up of >20%.

There were only a small number of well-controlled level 2 studies (which could be well-matched cohort studies) in TBI rehabilitation; SCI had many more level 2 studies.

*Sample size* was a particular problem. Most rehabilitation trials were small or limited in number of cases. In a simple two-group analysis of variance (without covariates), only a large treatment effect (Cohen's $f = 0.40$) is likely to be detected (with 80% power) in a sample size of 50 (25 in each cell).[12] To detect a more plausible, moderate treatment effect ($f = 0.25$), a sample size of 130 (65 in each cell) would be needed. Most rehabilitation trials increased experimental power by use of pretest covariates, but even so, only moderate or moderately large treatment effects could be detected, and then only for the primary outcome variable.

Limited statistical power is a particular problem in many rehabilitation studies because treatment effects in initial human trials are typically poorly understood. The precise outcome that is most responsive to the intervention and the subgroups that most benefit are often unclear before conduct of the research. Preliminary studies of small groups and even single case studies can be valuable to estimate the responsiveness of different outcome measures to treatment, and we recommend such studies. But larger samples permitting subgroup analysis are still needed to provide strong evidence regarding variations in patient responsiveness to interventions.

In sum, well-controlled studies remain few. More randomized trials are needed, and somewhat larger trials are needed. These trials should include measures and analyses necessary to advance understanding of what works best, for whom, and in what way. Limited funding was surely a major reason for small sample sizes. The paucity of multisite trials and limitations in participant accrual were also evident.

### Issues in Prognostic and Diagnostic Studies

Given the large total number of correlational studies in the rehabilitation literature, the small number of level 1 prognostic studies in rehabilitation was surprising. This is of particular concern because the Model Systems databases have such a longitudinal (and therefore correlational and predictive) structure. Reasons why prognostic or similar diagnostic or screening studies were downgraded include the following:

- Drop out rates of >20%.
- Use of potentially biased outcome measures without blinding, or failure to specify objectivity of outcome measure employed.
- Failure to properly specify the group to which the prediction applies. A common inclusion criterion in rehabilitation studies was admission to a comprehensive rehabilitation facility (e.g., analyses of the Model Systems database), the objective basis of which is unclear. Many predictive studies made no attempt to define a representative patient population or even studied a highly selected rather than a broad range of patients.

A rehabilitation prognosis study should be based on a sample of patients with objective indicators of need for rehabilitation. Many works would have been rated as strong (level 1 or 2) predictive studies if inclusion criteria were clearly defined in objective terms (e.g., all patients with impairment x and activity of daily living limitation y at z days after injury).

A particular problem was recurrently encountered: attempts to draw treatment (causal) inferences from correlational, pre–post, or longitudinal data sets. Many rehabilitation studies attempted to do this. Analyses of the resulting data have defined the nature of rehabilitation, clarified problems and expected recovery, and provided valuable information about factors related to recovery. But such studies are limited when evaluated from a strength-of-evidence perspective. Evidence standards clearly distinguish criteria for therapeutic studies from those for prognosis studies. According to standard evidence grading, such studies provide only a weak basis for inference of causation (level 3 usually). Longitudinal databases (especially if they lack a nontreatment comparison group) are a good research design for predictive studies but a weak design for study of therapeutic effectiveness.

A final observation on prognosis studies is in order. Many such studies failed to show that the prediction was clinically useful. Predictive relationships, although statistically reliable, were often not sufficiently precise to support clinical decisions. Often, no attempt was made to demonstrate the clinical utility of the prediction (e.g., to show that the prediction was more reliable than another method of comparable or greater expense). Accurate prediction is important for planning of clinical rehabilitation and for design of rehabilitation research.

## APPLICATION OF EVIDENCE STANDARDS TO REHABILITATION

Research is complex, and the true strength of evidence is not a simple matter. Although it is possible for reviewers to attain "reasonable to excellent agreement" using explicit rating schemes for study quality, disagreements and ambiguities occur.[13] In the attempt to apply evidence-based standards to rehabilitation, we too encountered occasional ambiguities and problems. More frequently, we encountered issues that are likely to become problematic in the future, as formal evidence syntheses increasingly affect the direction of rehabilitation research and clinical practice.

Most of the studies review criteria downgraded would be rated as weak by any set of recognized criteria, but there were a few (and only a few) cases in which a study was viewed subjectively as excellent but was downgraded because of review criteria. The considerations in the section to follow do not materially affect the overall evaluation of research quality in the literature reviewed, but they do suggest certain needs for broader and more sophisticated review methodologies in certain circumstances and do affect recommendations for the future of rehabilitation research.

### Technical Issues

#### Retrospective *vs.* Prospective Database Analyses

The terms "retrospective" and "prospective" were repeatedly ambiguous and controversial when applied to analyses of some databases (e.g., the Model Systems' databases). Some reviewers asserted that analyses of these databases are retrospective because one is analyzing previously collected data. Others asserted that the analysis was prospective because the databases were created to support research. Although the distinction between "retrospective" and "prospective" may not always be clear, the real issue is the quality of the data, whether major confounders have been controlled and whether the number of hypotheses tested has inflated alpha (type 1 error).[5,9,12,13] A priori specification of the hypotheses each database is designed to test would greatly alleviate these problems by permitting adjustment for the risks of repeat statistical testing[14] and by helping to ensure that the needed data are collected to test the hypotheses.

#### Grading the Quality of Cohort Studies

Extant evidence grading methods do not clearly distinguish well-matched comparison group studies with excellent control for case severity from cohort comparisons of unknown similarity. This is a potentially serious problem, as comparison with similar groups can provide evidence that a factor probably affects outcomes and that a treatment is "probably" effective—the usual criterion underlying a recommendation that a treatment "should be considered".[2, p. 39] Cohort studies can give results similar to RCTs.[15,16] However, judging similarity of groups is challenging in rehabilitation because comparability can be defined in terms of both severity of disease and function (activity/participation). Comparison with a similar cohort is one of the most practical research designs but difficult to grade in evidence reviews.[17] Considerations for appraising the quality of cohort studies and the comparability of comparison groups have been published.[18,19] Evaluation of the quality of cohort studies should also consider advances in causal modeling, propensity score modeling, and instrumental variable analysis.[20–22]

#### Grading Nonrandomized ABA and Individual Baseline Designs

Crossover, ABA, and interrupted time series designs are not explicitly distinguished from other studies lacking a comparison group (level 3 studies) in standard evidence grading manuals. Such studies can in fact provide much stronger evidence than simple pre–post research designs or other studies lacking a comparison group, provided certain criteria are met.[7,8,23,24] With interrupted time series, key criteria include evidence (not assumption) of baseline and continuing stability of the trend line and rapid effect onset.[7,8,25] (Crossover and ABA designs are particularly strong and powerful for smaller samples if combined with randomization).[9,24]

Individual baseline and n of 1 studies are not incorporated into standard evidence grading or are assigned to low levels (grade 3 or 4), regardless of design or context. But is this always so? The question is important because rehabilitation is in practice highly individualized, research funding is scarce, and individual baseline and case study designs are commonly done.

Individual baseline and n of 1 studies provide essential information in EBP, as one wants to know not whether a treatment tends to work on average for a group but whether the treatment works for the individual.[26,27] Studies of individuals and small case series can be optimal for exploring a new treatment, for titrating therapies, for documenting a promising variation in behavioral therapies, for enhancing knowledge of generalization of treatment to a new group, and to enhance understanding why some patients respond to a treatment of known (average) effectiveness whereas others do not, that is, for extending results of an RCT. Assertions that n of 1 reports somehow provide more reliable evidence that a treatment works than a randomized trial are not generally accepted, but rehabilitation needs methods to evaluate individual patient improvement.[28] Time series analysis can be applied to these small sample studies, provided there are a large number of measurement points before and after the intervention.[25] A more sophisticated consideration of the uses of well-designed n of 1 and case series investigations would be valuable.

## Grading Strong Quasi-Experiments

Extant evidence grading systems for medical and health studies do not mention certain strong quasi-experimental research designs and hence do not distinguish them from weaker, observational designs. For example, if treatment assignment can be made dependent on a cut point in a strictly quantitative scale and certain other assumptions are met (the "regression discontinuity" research design), strong inference of treatment effectiveness is possible.[7,8] Multiple interrupted time series studies can also provide strong evidence of effectiveness: assumptions include establishing baseline stability or trend (which requires many data points) and replication of the effect multiple times (to exclude contextual or history artifacts).[7,8] Evidence grading methods need to recognize the actual strength of such research studies, if and when they occur. The Department of Education's What Works Clearinghouse classifies a regression discontinuity study as providing level 1 evidence.[29] Other methods of grading evidence should also incorporate means to properly grade strong quasi-experimental research designs.

In the research literature examined for these reviews, no regression discontinuity studies were encountered, and the time series studies encountered suffered from major flaws (e.g., assumed rather than measured baseline stability). Consequently, limitations on the sophistication of evidence grading methods did not affect this overview of the state of the science. More sophisticated and sensitive evidence grading standards, however, would motivate the use of strong quasi-experimental research designs, which can be employed when randomization is infeasible or unethical.

## Study Typology and Review Criteria

The review criteria applied are based on established typing of studies as therapeutic (or interventional) or predictive (prognostic/diagnostic). However, criteria have been proposed for quality of decision assist and economic studies.[30,31] It is possible that some studies in rehabilitation are better considered as decision assist studies than as treatment effectiveness, diagnostic, or prognosis studies, as the study aims to help clinical professionals, people with disabilities, or policy makers to more rationally weigh alternatives or identify priorities. Expansion of review criteria to include such studies might result in higher and more appropriate grading of some investigations.

*Criteria for quality of measurement studies are not addressed at all by Cochrane, AAN, or other medical evidence standards.* Rehabilitation aims to improve function and quality of life, and so sensitive, reliable, and valid measures of these constructs are required before valid treatment studies can be undertaken. "Measurement Standards for Interdisciplinary Medical Rehabilitation"[32] provides a basis for evaluating such works. These standards elucidate key principles, and additional work would be needed to develop a simple but reasonably valid method of rating the quality of measures.

Finally, the appropriateness of evidence-based medicine criteria for AT and environmental interventions, which are common in rehabilitation but not reviewed here, is particularly questionable and will be discussed below.

## Problems in Applying Evidence Grading Standards to Rehabilitation

In this section we consider underlying issues raised by the process of applying evidence grading methods to rehabilitation.

### Applying Evidence Grading Standards to Medical Interventions

Standard evidence grading methods applied well to "medical" procedures and problems, that is to interventions whose proximal objective is to reduce biological pathologies or impairments[33] or to improve body function and structure.[34] For such procedures, questions of the applicability of evidence grading methods themselves were few, rather technical, and similar to those encountered in grading evidence throughout medical care. In sum, it is hard to see how evidence criteria for medical procedures in rehabilitation differ from those for other medical populations.

This not to say that important issues were not encountered, such as:

- Heterogeneity of patient response to treatment.[35] Reporting of group average responses in terms of a single primary outcome is often insufficient.
- Side effects and long-term outcomes were often unclear.
- RCTs are not only few but may have been conducted on groups or in circumstances that differ from those encountered in clinical rehabilitation so that the applicability of results to rehabilitation populations is limited or unclear.
- As noted previously, grading of the quality of some study designs (e.g., cohort studies and ABA designs) was sometimes unclear or problematic.

These issues and problems in evidence grading are not unique to rehabilitation but are shared by other areas of medical practice.

### Applying Evidence Grading to Nonmedical Interventions

More frequent difficulties were encountered in applying evidence grading methods to stereotypically nonmedical interventions, including behavioral and psychosocial intervention, AT, and environmental interventions. These interventions may or may not affect pathology or body impairment, but they do involve and affect activity/participation and quality of life.

#### *Activity-Based and Psychosocial Interventions*

Behavioral, psychosocial, activity, exercise, and/or educational interventions differ from pharmaceuticals in that they work through the attention and action of the person rather than through an external agent. The person's motives, values, and thoughts and the situational and environmental factors are typically involved. These interventions are typically multifaceted, often context-dependent,[36] and vary depending on the responsiveness of the individual. Personal factors can be involved, including the development of relationship of the person with the therapist, and it can be difficult to distinguish the treatment from characteristics of the person delivering it.[37,38] Goals of behavioral and psychosocial interventions may be multiple, individualized, dependent on culture, and difficult to compare across participants.[37,38] Because it is so easy to modify the communicative and activity elements of the interventions, generalization of treatment is difficult and cannot be ensured by even a definitive RCT—a situation very different from drug studies, where it is commonly assumed that the drug will work in a similar way in other persons with the same diagnostic configuration. Even if the RCT proves that the psychosocial treatment works in one setting, further work is likely to be needed to foster and ensure generalization.

There are potential solutions to these problems. Although a substantial labor, it is possible to manualize activity and psychosocial interventions once the essential features that produce a positive effect are known, that is, to define the treatment protocol objectively. The treatment can be defined in terms that are relative to the person's stage of readiness, exercise tolerance, goals, and/or responsiveness rather than in simple absolute terms such as number of repetitions. If the treatment is proven to work, presumably in a RCT or other rigorous study, it will be worthwhile to develop a training program. Methods of assessing the fidelity of treatment delivery can be developed to ensure that the treatment is properly applied and generalizes.[39] Needed variations in treatment effectiveness in new settings and circumstances can and should be evaluated with additional studies, analogous to "phase 4" drug studies (although one might question whether funding will ever be sufficient to evaluate these questions of generalization using RCT methodologies). Finally, multiple outcomes can be measured and multiple effects can be tested, controlling for the biasing effects of repeated testing (that is, for the false discovery rate) by modern statistical procedures.[14]

#### *Assistive Technology and Environmental Interventions*

AT and environmental interventions present major challenges to standard evidence grading methods, rooted as they are in the assessment of drugs. AT and environmental interventions differ greatly from a drug. The working of the assistive device or environmental modification is prototypically visible to the person, rather than hidden in the body: the person can typically observe whether the device meets a current functional need or not.[40] Questions that the person might ask include long-term durability, utility, and alternatives. One does not do an RCT to test whether a wheelchair improves mobility of paraplegic persons any more than one would do an RCT to test the efficacy of parachutes to reduce mortality among people jumping from aircraft.[40]

Nonetheless, randomized trials are not completely irrelevant. Many forms of technology require substantial learning and adaptation to use effectively. A range of "extraneous" variables may modulate response to the intervention in the short term and affect activities and quality of life in the long term. RCTs are needed to test whether a program of improved equipment provision and training increases community participation (cf., the RCT by Mann et al.[41] showing that the delivery of needed AT reduces functional decline and institutionalization among fragile community living elders). The matter is complex. Frameworks for evaluation of AT[42] and for health technology assessment[43] are being developed. Improved guidelines for grading studies of AT and environmental interventions—and their match with individuals[44]—need to be developed.

### Multidisciplinary Programs, Policy, and Rehabilitation Services Research

Rehabilitation in practice involves provision of multiple interventions—pharmacologic, nursing, exercise, equipment, psychological, and others. There is sparse literature of RCTs on such multifaceted, rehabilitative programs (e.g., geriatric units,[45] comprehensive stroke units,[46] and brain injury units[47]), which was not reviewed in this supplement. A central logical problem is that insistence on a completely conventional or univocal definition of the intervention would lead to absence of study and absence of any good information—hardly a desirable result. RCTs are unlikely ever to be done to study the effectiveness or cost-effectiveness of most multifaceted, multidisciplinary programs, whole service delivery systems, or policy changes.

How can we study such complex programs with sufficient rigor to guide improvements in program effectiveness and cost-effectiveness? Past studies of randomized allocation to a well-defined guideline[48] or intervention system suggest that RCTs can sometimes be done. Times series and regression discontinuity research designs can be applied.[8,49] However, correlational methods are most commonly employed in the study of systems of care (health services research) and will often be the only feasible method. Questions of the quality and appropriateness of quasi-experimental and correlational research become critical.

### Implications of Results for Practice

Although the aim here was to grade quality of research rather than to draw practice recommendations, the question of practice recommendations should not be completely ignored: guiding clinical practice is the essential purpose of EBP reviews[3]—and indeed of research as a whole.

A review of the evidence relevant to many topics in rehabilitation would technically support only level B or C recommendations: rehabilitative intervention may possibly be effective and should be considered. Such a pallid recommendation is not satisfactory, given the magnitude of the human problem of disablement. Writers of guidelines in rehabilitation have too often had to base their recommendations on weak evidence—level 3 or level 4 studies—supplemented by expert experience and opinion. More and better evidence is needed regarding the effectiveness of alternative rehabilitative interventions.

## Wider Considerations: Food for Thought

Synthesis of evidence and its interpretation and application raise a number of questions and issues that deserve attention.

### Sparse Evidence

The main problem confronting rehabilitation from an EBP viewpoint is that of sparse evidence: the number of well-controlled trials is insufficient to provide definite guidance for most clinical decisions. This problem is not uncommon in EBP endeavors throughout health care,[1,3] but the fact that other fields have a similar problem should not lessen attention to the problem in rehabilitation.

Scientifically, absence of evidence does not prove absence of effectiveness. A treatment can only be proven to be ineffective by a large RCT providing small confidence intervals, showing no worthwhile effects on outcomes. When only small, exploratory studies exist, the strictly scientific conclusion is to withhold judgment. In practice, however, a decision must be made by clinicians, policy makers, or patients; and in science, hypotheses are tested and treated as confirmed or disconfirmed. So the problem is serious. We cannot solve the problem, but we can point out essential considerations.

#### *Balancing Types of Error*

There are two types of error that can occur in drawing a conclusion: type 1 error—the error of stating that a hypothesis is true when it is not—and type 2 error—the error of stating that a hypothesis is false when it is actually true. Both are equally real and important. Rigor in the sense of absence of type 1 error can always be increased by simply increasing standards, but without an increase in the sensitivity or power of the experiment, type 2 error will be increased. To increase rigor simply by insisting on narrow standards that ignore or vastly expand type 2 error would not be a contribution. The probability of type 2 error is significant in a field like rehabilitation, which has a substantial historical development, is widely accepted as effective, but which has mostly only small research trials. Evidence grading standards need to balance type 1 and type 2 errors, that is, to balance rigor and sensitivity.

### *Funding Bias*

The extent of evidence on any given topic will depend on funding for the needed studies. If publication bias—the tendency not to publish negative results—possibly affects the research literature,[50,51] surely funding bias does too. Research on some types of interventions (e.g., drugs) is comparatively well funded because there is a well-developed method of financing. There is comparatively little funding for other types of interventions (e.g., activity therapies and psychosocial interventions that cannot be patented, pragmatic clinical trials designed to test the cost-effectiveness of alternative strategies for multidisciplinary rehabilitation). Reviews have often concluded that there is insufficient evidence for many rehabilitation interventions. Does scarcity of evidence for rehabilitative interventions indicate an absence of effectiveness—or absence of funding for the needed clinical trials?

### *Optimal Trial Design*

Another consideration is that the presumptions that underlie certain features of clinical trial design for new drugs may not apply to many rehabilitation interventions. It is appropriate, for instance, to presume that a new drug should not be employed until it is proven to be effective and safe in an RCT. After all, even with extensive animal studies, a drug may not work in humans and may have unexpected side effects, and the patient cannot directly observe the biological mechanism of action. But are such presumptions appropriate to rehabilitation? Functionally based rehabilitative interventions (the activity and physical therapies, psychology) differ from drugs in that major aspects of process—functional gain or whether adaptive devices are meeting goals—can be observed by client and therapist. Especially in subacute and chronic populations, the client's satisfaction with functional progress is often the key consideration. Functional therapies tend to be relatively safe, and in any case, the safety of functional interventions is highly contextual and may be better studied using observational, field methodologies than a narrowly focused RCT. Vocational rehabilitation interventions are also highly contextual, depending on details of the particular job, employer, job market, and so on. Such considerations do not call into question the applicability of the RCT to rehabilitation, but they do raise the issue that optimal trial choice and design for functional therapies may differ from those for drug trials. A complete emphasis on internal validity justified by an assumption that generalization to practice will occur, univocal definition of the treatment, and the idea that every new treatment needs to be tested in an RCT—such presumptions may not be optimal for study of many functional therapies in rehabilitation. These considerations make it difficult to design trials that are relevant to clinical practice in rehabilitation and provide a disincentive for the development of clinically valid trials.

There are practical and ethical limitations to RCTs in rehabilitation.[52] Because rehabilitation is an accepted service, it would not be considered ethical to conduct a study with a control group that denies participants access to rehabilitative services. It is also very difficult to engender enthusiasm and funding for study of "old" interventions. Insistence on RCTs as the only acceptable method for providing evidence for treatment effectiveness serves as a barrier to investigation of these interventions using feasible quasi-experimental research designs and well-controlled cohort studies. Although such research may not always provide the highest level of evidence, it can answer the question of whether a treatment is probably effective. A reasonable probability of benefit is often an adequate level of certainty for patients and clinicians to agree to treatment. Provision of funding for application of these alternative designs for rehabilitation interventions that otherwise will not be investigated would result in an increase in the evidence base for these interventions.

## Inference from Multiple Lines of Evidence

Evidence grading methods only consider sources of information that are published in scientific journals, directly related, and well documented. By considering only directly related scientific studies, evidence grading methodologies provide a highly defensible answer to questions of strength of evidence. Scientific theories and findings, however, are often confirmed or disconfirmed, not by a single test, but by a convergence of evidence from different sources.[49,53] Clinicians, and their clients too, draw treatment conclusions on the basis of a multiplicity of information sources. Two sources of information each showing a 90% probability of effectiveness—*if* independent and about the same intervention—would conjointly indicate a 99% probability. In sum, there are situations in which consideration of multiple lines of evidence might alter practice recommendations. Inclusion of multiple lines of evidence could logically improve the sensitivity of evidence grading methods, but reliable, widely accepted methods of including such complexity have yet to be developed.

## Generalization and the Role of Theory

Generalization of treatment is a *major* issue in the application of research to practice. If a study recruits the same type of patients seen in clinical practice, provides the same treatment as would be delivered in practice, and assesses outcomes using the same criteria for success as one would employ in practice, generalization to similar patients may be reasonably (but not without risk) presumed. The problem is that RCTs may involve highly selected patients who receive atypically specialized treatment measured by outcomes that differ from those of most concern to the patient (that is, the study design is optimized to test efficacy of treatment in principle rather than effectiveness in practice). Clinicians (and persons with disability) need to know whether treatments shown to be efficacious in clinical trials will work for their patients with similar problems.

In our reviews, extrapolation of treatments known to be effective in other similar populations to rehabilitation samples was not explicitly or comprehensively reviewed. This is a major problem in evaluation of the evidence base of rehabilitation: in practice, rehabilitation involves the organized application of numerous interventions that are "known" or accepted as effective in other medical and healthcare disciplines. Diagnosis and treatment of depression, for instance, may vary for people with rehabilitative conditions (e.g., SCI, stroke, or TBI), but it would hardly be reasonable to claim that pharmacologic and cognitive-behavioral treatments are completely without evidence or applicability in rehabilitation.

"Extrapolations from Level 1 studies" to a similar population or problem are treated as a level 2 by the Oxford Cochrane site.[30] This is a very reasonable position, one not employed in many evidence reviews and inconsistently employed in others. But what constitutes a similar population or problem? How far should extrapolations go?

An important form of this question in rehabilitation is whether criteria for generalization for behavioral and psychosocial interventions are the same as those for more purely medical or physiologic interventions in rehabilitation. No doubt the effectiveness of learning-based interventions, for instance, may be altered by diagnostic factors, but is it reasonable to assume that the effectiveness is always bounded by—and therefore, effectively determined—by physical diagnosis? Difficulty of the skill to be learned, its meaning to the person, attention, and opportunities for repeated practice, and for feedback, would seem to be essential concerns for educational, learning-based, physical therapy and occupational therapy interventions.[54–56] Rehabilitation has a long history of research demonstrating the importance of function as the central concern.

A final question is whether criteria for quality of generalization studies should be the same as for effectiveness studies that attempt to test a new treatment process. Although a large RCT could provide a more rigorous answer to the questions of generalization, the expense of such RCTs will ordinarily reserve their use for higher priorities. If a generalization study assessed treatment fidelity and whether the new, similar patient group responded in ways that are similar to those of the preceding randomized efficacy trial, would that not often be sufficient to establish likely generalization of effectiveness? Studies of heterogeneity of individuals' responses to treatment also raise the issue of type of research design and level of evidence needed. Similar issues arise in prediction research, in which more precise prediction is required for outcomes of high social or economic importance. In sum, good evidence is needed, but is it reasonable to demand the same standards of rigor for every clinical question? The question of sufficiency of evidence is important, and it could be valuable to define criteria for appropriate or optimal, rather than simply maximal, research designs.

### *Role of Theory*

Science has always been based on theory and direct observation.[53] Theories need to be tested—validated, confirmed, disconfirmed, or altered—but generalization should occur (or not occur) according to a theoretical understanding of the problem and how the intervention affects it. Data can test a hypothesis in a sample from a well-defined population but do not by themselves provide a basis for estimating where broader generalization is most likely to occur.[8] Standard evidence grading methods, however, do not explicitly take theory into account. This is a limitation, as generally in science, confirmation of theory-driven hypotheses is given stronger evidentiary weight than confirmation of atheoretical (empirically derived) hypotheses.

In medicine, core theories are biological. Physicians and medical research panels commonly focus on biological plausibility or mechanism as the central consideration for whether they believe in an intervention. Similarly, in the usual evidence review using methods from evidence-based medicine, the bounds of the literature search are defined by diagnostic criteria.

However, psychosocial processes are also involved in rehabilitation, and psychosocial theories differ from biological theories. Rehabilitation concentrates on reduction of disability and improvement of quality of life. Diagnostic factors undoubtedly affect function, but pathology and disease alone do not determine disability or quality of life.[56] Rehabilitation involves psychosocial interventions and, in practice, the simultaneous application of psychosocial and medical interventions. Treatment objectives in rehabilitation may be at the level of disease, impairment, activity/participation, and/or quality of life,[57] and articulated biopsychosocial theories need to be explicated to specify the nested levels of rehabilitative interventions.[58]

If it were possible to develop criteria for validated or sound theory, ratings of strength of evidence could be developed for generalization of treatments that go beyond current ratings. Attempts have been made to define characteristics of a sound theory,[59] but definite criteria for validated theories remain elusive. Nonetheless, theories provide important guidance regarding the likely generalizability of research findings.

## Logic of Research Development

Much of the rehabilitation research literature describes and delimits problems, identifies correlates or factors involved in disablement or recovery, or suggests possibly effective elements of interventions. Other studies develop methods of assessment or measurement. Evidence-based grading standards properly classify such works as lower level because they do not establish clinical effectiveness. But are they "low quality" (bad) research—or are they better characterized as preparatory?

The logic of scientific discovery involves consideration, testing, and confirmation or disconfirmation based on multiple lines of evidences or sources of information.[7,22,53] After proof of principle, considerable research and development are typically needed to quantify, confirm, and understand the discovery and to test its effectiveness, generalizability, and value in practice. Extensive laboratory studies are typically required to develop new drugs or other therapeutic agents before moving to widespread clinical trials. The development of improved behavioral, educational, and psychosocial interventions also requires sustained developmental work. One cannot do a valid therapeutic study unless one can quantify those aspects of function or quality of life targeted by the intervention: preparatory research is required to develop outcome measures that meet standards for validity and reliability.[32]

## Phases of Research Development

In research to develop new pharmaceutical agents, the need for phases of research development, moving from smaller preliminary studies to larger more definitive ones, is well accepted. Typically four phases are defined (after basic nonhuman research is done to establish promise). These phases may be summarized as follows:

- Phase 1: Initial or preliminary human studies to estimate safety, dosage, and to see if the drug seems to have desired positive effects in humans.
- Phase 2: Clinical trials with small- to moderate-size groups. These provide informative but typically not definitive results. Even if a statistically stable positive effect is found, side effects, dosage, long-term effects, the optimal treatment group, exclusions, and outcome measures may need to clarified or modified, motivating the need for a more definitive trial.
- Phase 3: Large (theoretically) definitive trials to establish effectiveness and safety.
- Phase 4: After-market monitoring and studies. These are in fact often needed to understand unexpected effects and generalization to the wider populations seen in practice.

A similar logic of research development can be applied to rehabilitation studies, which are hardly simpler than drug studies. For behavioral and psychosocial or AT studies, the phases may need some redefinition: preliminary work may be done in clinical practice and in the laboratory. (Progress does not proceed only from the laboratory to practice: scientific knowledge in health care is also advanced by astute observation and analysis of events in practice.) Basic and phase 1 research may encompass measure development and correlational studies and also qualitative research, case studies, and case series. In rehabilitation, too, small RCTs would be advisable before large ones to ensure that the intervention is indeed promising and to identify dimensions and limitations of efficacy: clinical trials involve complex considerations, and good phase 2 studies commonly reveal something unexpected. Assuming results of a small trial are promising, it may be worthwhile to mount a definitive trial (phase 3) of the rehabilitative intervention.

In sum, much of the scientific literature in rehabilitation can be aptly characterized as preparatory; that is, it develops a background of understanding and quantification needed for the conduct of more definitive clinical trials. Rehabilitation research need not only study new interventions but also develop and test the rich variety of interventions whose effectiveness is suggested but unproven by past research.

## CONCLUSIONS

Systematic methods for review of strength of research evidence can and should be applied to rehabilitation. Extant methods for evidence review are based on sound and widely recognized rules-of-thumb for what typically constitutes stronger *vs.* weaker research designs, given the question at issue. We recommend that future reviews, textbooks, and practice recommendations in rehabilitation employ systematic review methods. This recommendation is not without qualification. Limitations to standard evidence grading methods were identified. Expert opinion can provide insights that a mechanical categorization of studies misses. Research is complex, and many complex issues arise in evidence grading. Nonetheless, to justify rehabilitation to the world—and to clarify the objective basis for rehabilitative interventions—systematic evidence reviews will have to be done.

Although the pace of accumulation of evidence is increasing, we continue to be confronted by the problem of sparse evidence. Rehabilitation needs to obtain firmer evidence—not of whether "rehabilitation works," which is too vague a question, but of what works, in what intensity, for whom, and under what circumstances—to provide a valued outcome. Although it is true that RCTs are expensive, require substantial preparatory work, and are not the only way of obtaining good evidence, other forms of research are also expensive but do not usually provide the same degree of rigor or respect. RCTs are accepted as the gold standard for evidence of therapeutic efficacy throughout medical care, and studies of new interventions in medical rehabilitation will in practice need to meet this standard.

At the same time, it should be recognized that RCTs are unlikely ever to be done in some key research areas, such as the evaluation of large multifaceted systems of care. Thus, to require RCTs before viewing the evidence as sufficient to guide decisions would be to ignore pressing issues of clinical service delivery and policy. A variety of high-quality nonrandomized research designs, including strong quasi-experimental research and well-controlled cohort studies, are needed to provide the needed information.

Scarcity of RCTs was far from the only limitation found in the rehabilitation research reviewed. Few level 2 treatment studies were found, and the number of high-level prognosis studies was small. Absence of blinding—which is not very expensive to implement—and high losses to follow-up were significant problems. Small samples sizes have led to uncertain results and lack of understanding of variations in participant response to treatment—a major limitation because individualization is all-but-universally believed to be an essential characteristic of quality rehabilitation.

The phase of development of knowledge regarding the intervention or question at issue is not considered in standard grading methods, so studies may be criticized as "low level" when in fact the research design was optimal. Many lower-level studies have developed treatment protocols, clarified outcome measures, and reported positive results. In such cases, the intervention is ready for testing in a larger clinical trial. If a phase 1 drug study—which may provide only level 3 evidence—had positive results, the conclusion would not be that further development should stop because methods were weak and the sample size small. The conclusion would be that the therapeutic agent warrants testing in larger, more rigorous trials. Phase 1 studies, if positive, are followed by small randomized clinical trials (phase 2 studies), which in turn are followed by large, definitive (phase 3) studies. Should not this logic of scientific development be applied to rehabilitation research as well?

The evidence reviews that follow this Overview do not identify what interventions should be the priority for future development and testing, but they do present many essential points that will assist in identifying the logical next steps in rehabilitation research. Although systematic evidence reviews delineate the limitations of rehabilitation research, they also show that rehabilitation is rich with research opportunities and findings that, with further development, can provide the results needed to improve clinical practice.

## ACKNOWLEDGMENTS

We thank Andrea Robinson Vincenti and Melissa Dolese, who assisted ably with technical preparation of this work, and the Clinical Practice Committee of the American Congress of Rehabilitation for their assistance in the intellectual development of evidence-based rehabilitation.

### REFERENCES

1. Sackett DL, Straus SE, Richardson WS, et al: *Evidence-Based Medicine: How to Practice and Teach EBM.* Edinburgh, Churchill Livingstone, 2000
2. Edlund W, Gronseth G, So Y, Franklin G: American Academy of Neurology Clinical Practice Guideline Process Manual. 2004 ed. American Academy of Neurology, 2004
3. Gray JAM: *Evidence-Based Healthcare,* ed 2. Edinburgh, Churchill Livingston, 2001
4. West S, King V, Carey TS, et al: Systems to rate the strength of scientific evidence. AHRQ Evidence report/technology assessment: Number 47. (Prepared by the Research Triangle Institute-University of North Carolina Evidence-based Prac-

tice Center under Contract No. 290-97-0011). Rockville, MD, Agency for Healthcare Research and Quality, 2002 Apr. Report No.: AHRQ Publication No. 02-E016

5. Alderson P, Green S, Higgins JPT: *Cochrane Reviewers' Handbook 4.2.2.* Chichester, UK, John Wiley and Sons, 2004
6. Ottenbacher KJ, Maas F: How to detect effects: Statistical power and evidence-based practice in occupational therapy research. *Am J Occup Ther* 1999;53:181–8
7. Cook TD, Campbell DT: *Quasi-Experimentation: Design and Analysis Issues for Field Settings.* Chicago, Rand McNally, 1979
8. Shadish WR, Cook TD, Campbell DT: *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston, Houghton Mifflin, 2002
9. Rosner B: *Fundamentals of Biostatistics*, ed 6. Belmlont, CA, Duxbury Thomson, 2005
10. West S, King V, Carey TS, et al: Systems to rate the strength of scientific evidence: Summary. AHRQ Evidence report/technology assessment: Number 47. Agency for Healthcare Research and Quality, 2002 Mar. Report No.: AHRQ Publication No. 02-E015
11. Harris RP, Helfand M, Woolf SH, et al: Current methods of the US Preventive Services Task Force: A review of the process. *Am J Prev Med* 2001;20(3 suppl):21–35
12. Sample Power 2.0 [computer program]. Version 2.0. Chicago, SPSS, 2001
13. Oxman AD, Guyatt GH, Singer J, et al: Agreement among reviewers of review articles. *J Clin Epidemiol* 1991;44:91–8
14. Benjamini Y, Hochberg Y: Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J Roy Stat Soc* 1995;57:289–300
15. Pocock SJ, Elbourne DR: Randomized trials or observational tribulations? *N Engl J Med* 2000;342:1907–9
16. Concato J, Shah N, Horwitz RI: Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med* 2000;342:1887–92
17. Norris SL, Atkins D: Challenges in using nonrandomized studies in systematic reviews of treatment interventions. *Ann Intern Med* 2005;142(12 pt 2):1112–9
18. Normand SL, Sykora K, Li P, et al: Readers guide to critical appraisal of cohort studies: 3. Analytical strategies to reduce confounding. *BMJ* 2005;330:1021–3
19. Mamdani M, Sykora K, Li P, et al: Reader's guide to critical appraisal of cohort studies: 2. Assessing potential for confounding. *BMJ* 2005;330:960–2
20. Kline RB: *Principles and Practice of Structural Equation Modeling, Second Edition*. New York, Gilford, 2004
21. Thompson WD: Statistical analysis of case-control studies. *Epidemiol Rev* 1994;16:33–50
22. Little RJ, Rubin DB: Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annu Rev Public Health* 2000;21:121–45
23. Finkel SE: *Causal Analysis with Panel Data*. Thousand Oaks, CA, Sage, 1995
24. Vray M, Girault D, Hoog-Labouret N, et al: Methodology for small clinical trials. *Therapie* 2004;59:273–6
25. Chatfield C: *The Analysis of Time Series: An Introduction*, ed 6. Boca Raton, FL, CRC Press, 2004
26. Jenicek M: *Foundations of Evidence Based Medicine*. New York, Parthenon Publishing Group, 2003
27. Jenicek M: *Clinical Case Reporting in Evidence-Based Medicine*, ed 2. Oxford, Oxford University Press, 2001
28. Ottenbacher KJ, Hinderer SR: Evidence-based practice: Methods to evaluate individual patient improvement. *Am J Phys Med Rehabil* 2001;80:786–96
29. US Department of Education WWC. WWC Study Review Standards, 2005. Available at: http://www.whatworks.ed.gov/reviewprocess/standards.html
30. Oxford Center for Evidence Based Medicine. Levels of Evidence and Grades of Recommendation, 2005. Available at: http://www.cebm.net/levels_of_evidence.asp. Accessed December 7, 2005
31. Gold MR, Siegel JE, Russel LB, et al: *Cost-Effectiveness in Health and Medicine.* New York, Oxford University Press, 1996
32. Johnston MV, Keith RA, Hinderer SR: Measurement standards for interdisciplinary medical rehabilitation. *Arch Phys Med Rehabil* 1992;73:S3–23
33. *International Classification of Impairments, Disability, and Handicap*. Geneva, World Health Organization, 1980
34. *International classification of Functioning, Disability, and Health: ICF.* Geneva, World Health Organization, 2001
35. Kravitz RL, Duan N, Braslow J: Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *Milbank Q* 2004;82:661–87
36. Haley SM, Coster WJ, Binda-Sundberg K: Measuring physical disablement: The contextual challenge. *Phys Ther* 1994;74:443–51
37. Norcross JC, Beutler LE, Levant RE (eds): *Evidence-Based Practices in Mental Health*. Washington, DC, American Psychological Association, 2006
38. Jensen PS, Weersing R, Hoagwood KE, et al: What is the evidence for evidence-based treatments? A hard look at our soft underbelly. *Ment Health Serv Res* 2005;7:53–74
39. Borrelli B, Sepinwall D, Ernst D, et al: A new tool to assess treatment fidelity and evaluation of treatment fidelity across 10 years of health behavior research. *J Consult Clin Psychol* 2005;73:852–60
40. Smith GC, Pell JP: Parachute use to prevent death and major trauma related to gravitational challenge: Systematic review of randomised controlled trials. *BMJ* 2003;327:1459–61
41. Mann WC, Ottenbacher KJ, Fraas L, et al: Effectiveness of assistive technology and environmental interventions in maintaining independence and reducing home care costs for the frail elderly: A randomized controlled trial. *Arch Fam Med* 1999;8:210–7
42. Fuhrer MJ, Jutai JW, Scherer MJ, et al: A framework for the conceptual modelling of assistive technology device outcomes. *Disabil Rehabil* 2003;25:1243–51
43. Philips Z, Ginnelly L, Sculpher M, et al: Review of guidelines for good practice in decision-analytic modelling in health technology assessment. *Health Technol Assess* 2004;8:iii–xi, 1
44. Scherer MJ: *Assistive Technology: Matching Device and Consumer for Successful Rehabilitation*. Washington DC, American Psychological Association, 2001
45. Rubenstein LZ, Josephson KR, Wieland GD, et al: Effectiveness of a geriatric evaluation unit: A randomized clinical trial. *N Engl J Med* 1984;311:1664–70
46. Organised inpatient (stroke unit) care for stroke. *Cochrane Database Syst Rev* 2002;1:CD000197
47. Powell J, Heslin J, Greenwood R: Community based rehabilitation after severe traumatic brain injury: A randomised controlled trial. *J Neurol Neurosurg Psychiatry* 2002;72:193–202
48. Luther SL, Nelson A, Powell-Cope G: Provider attitudes and beliefs about clinical practice guidelines. *SCI Nurs* 2004;21:206–12
49. Campbell DT, Stanley JC: *Experimental and Quasi-Experimental Designs for Research.* Boston, Houghton Mifflin, 1963
50. Montori VM, Smieja M, Guyatt GH: Publication bias: A brief review for clinicians. *Mayo Clin Proc* 2000;75:1284–8

51. Olson CM, Rennie D, Cook D, et al: Publication bias in editorial decision making. *JAMA* 2002;287:2825–8

52. Muche R, Rohlmann F, Buchele G, Gaus W: [The use of randomisation in clinical studies in rehabilitation medicine: Basics and practical aspects]. *Rehabilitation (Stuttg)* 2002;41:311–9

53. Salmon WC: *Causality and Explanation*. Oxford, Oxford University Press, 1997

54. Law M: *Evidence-based Rehabilitation: A Guide to Practice*. Thorofare, NJ, Slack, 2002

55. Pagliarulo MA: *Introduction to Physical Therapy*, ed 2. New York, CV Mosby, 2001

56. Zola IK: Toward the necessary universalizing of a disability policy. *Milbank Q* 1989;67(suppl 2, pt 2):401–28

57. Johnston MV, Stineman M, Velozo CA: Outcomes research in medical rehabilitation: Foundations from the past and directions for the future, in Fuhrer M (ed): *Assessing Medical Rehabilitation Practices: The Promise of Outcomes Research*. Baltimore, Paul H. Brookes, 1997, pp 1–41

58. Whyte J, Hart T: It's more than a black box: It's a Russian doll. Defining rehabilitation treatments. *Am J Phys Med Rehabil* 2003;82:639–52

59. Keith R, Lipsey MW: The role of theory in rehabilitation assessment, treatment, and outcomes, in Glueckauf RI, Sechrest LB, Bond GR, McDonel E (eds): *Improving Assessment in Rehabilitation and Health*. Newbury Park, CA, Sage, 2005, pp 33–58