SPECIAL SECTION: SPECIAL COMMUNICATION

The Uniform Postacute Assessment Tool: Systematically Evaluating the Quality of Measurement Evidence

Mark V. Johnston, PhD, Daniel Graves, PhD, Maureen Greene, MS

ABSTRACT. Johnston MV, Graves D, Greene M. The uniform postacute assessment tool: systematically evaluating the quality of measurement evidence. Arch Phys Med Rehabil 2007;88:1505-12.

The U.S. Congress has mandated that the Centers for Medicare & Medicaid Services develop a uniform assessment instrument that characterizes patients' needs for postacute services. What scientific criteria should be used to evaluate the evidence for such a tool? The validity of a measure can be accurately graded only if the constructs measured and their applications are clearly defined. We argue that improving postacute placement is the main purpose of the uniform postacute assessment (recently renamed the Continuity Assessment Record and Evaluation). We argue that placement itself needs to be better defined and measured in terms of transitions in the level and type of treatment and care. Domains that should be measured to provide appropriate rehabilitative placement recommendations include level of skilled medical and nursing care, therapies, routine living support, family support, ability to participate in self-care, and patient preference. Almost no research has been performed to quantify and predict the needed intensity of rehabilitative therapy, a major lacuna in evidence. Criteria and examples are provided for research that will provide minimal, probably adequate, or strong evidence for the validity of systems that recommend care transitions. A long-term program of research and systematic evidence synthesis is needed to support guidelines that improve postacute placement.

Key Words: Evidence-based medicine; Medicare; Occupational therapy; Outcome assessment (health care); Physical therapy (specialty); Rehabilitation; Reliability and validity.

TO INFORM AND IMPROVE both clinical and policy decisions, postacute care (PAC) needs an evidence basis. This evidence basis should accumulate over time. Criteria and methodologies have been developed for grading the level or strength of evidence for treatments and are relatively well known.[1-3] These methodologies have provided the technical basis for modern systematic reviews and the evidence-based medicine movement, which has enhanced the identification and speeded the accumulation and application of best medical knowledge.[1,3] It is possible to define criteria and methodologies for grading and synthesizing the strength of evidence for measures and measurement systems as well. The application of such criteria may be expected to speed progress and improve the application of measurement tools, including those used in postacute placement and policymaking.

By defining scientific criteria a priori, before reviewing the actual evidence, we can establish a basis for evaluation that is independent from the findings themselves and thus less subject to bias by transient situational considerations. With well-synthesized evidence, policy debate can at least start from a factual basis. Standardized criteria for grading the level of evidence can enable us to distinguish what is definitely known from what is probably or only possibly known, thus assisting application of measures and guiding priorities for continued research.

Judgments of the scientific evidence for a measurement system should be based on well-established scientific standards of reliability and validity.[4-6] Although these principles and standards provide a necessary and useful basis, summary and interpretation are needed to apply them. The past literature on measurement reliability and validity is voluminous, highly detailed (at the level of statistical methods), and often vague (at the level of general principles and criteria). The phrase "reliable and valid" is frequently unqualified; a measure has it or not. When reliability and validity are treated as categories or dichotomies, the same evidence may be used to support or criticize because shades of gray must be labeled as either black or white. Discourse about measures easily becomes a debate of good versus bad or like versus dislike. To increase discrimination, more refined criteria are needed to evaluate the level of evidence for a measure.

A uniform postacute assessment (UPAA) tool is potentially so important that the highest-quality criteria should be applied to its development, validation, improvement, and uses over time. We will use the abbreviation UPAA to emphasize the generic, long-term view taken in this article but will occasionally use the more current name, the Continuity Assessment Record and Evaluation tool.

Objectives

This article is designed to identify research methods and questions that are critical to producing a valid and useful postacute assessment system. Our objectives were as follows:

1. To briefly describe a method of grading the level of evidence for a measure or measurement system based on established principles of reliability and validity of probabilistic measurement.[4,5]
2. To examine the conceptual basis for UPAA measurement, particularly the primary stated use of a UPAA—improved postacute placement[7,8]—and its conceptual bases. We sketch domains of patient care that must be distinguished to validate a UPAA, focusing on the needed intensity of continuing rehabilitative therapy.

From the Department of Occupational Therapy, College of Health Sciences (Johnston) and Nursing (Greene), University of Wisconsin-Milwaukee, Milwaukee, WI; and the Institute for Rehabilitation and Research, Baylor College of Medicine, Houston, TX (Graves).

Presented in part to the State-of-the-Science Symposium on Post-Acute Rehabilitation, February 12, 2007, Arlington, VA.

Supported in part by the National Institute on Disability and Rehabilitation Research (grant no. H133G060218).

No commercial party having a direct financial interest in the results of the research supporting this article has or will confer a benefit upon the author(s) or upon any organization with which the author(s) is/are associated.

Reprint requests to Mark V. Johnston, PhD, Occupational Therapy, University of Wisconsin-Milwaukee, PO Box 413, Milwaukee, WI 53201, e-mail: *johnsto@uwm.edu.*

0003-9993/07/8811-00097$32.00/0
doi:10.1016/j.apmr.2007.08.117

3. To briefly review the literature on needed intensity of rehabilitative therapies such as physical therapy (PT), occupational therapy (OT), and speech-language pathology (SLP). We cite illustrative articles to provide an overview of the literature but do not attempt a definitive evidence review because extant literature is not routinely indexed in terms of measurement as a concept or subject heading.
4. To identify a hierarchy of issues on the validity of the UPAA as a guide or tool for postacute placement, concentrating on the intensity of rehabilitative therapies needed by patients. We identify the criteria and study designs needed to validate a UPAA as a measure for type and intensity of therapies, grading these studies by strength of evidence they provide. Validation criteria that provide exploratory, minimal, probably adequate, and strong evidence are described.

In sum, we asked what is the consequence of a focused application of scientific measurement principles to validation of the UPAA?

ADVANCING MEASUREMENT: TOWARD GRADED QUALITY CRITERIA

An understanding of principles and elements of measurement reliability and validity is a necessary basis for evaluating the quality of measures. Standards for evaluating the reliability and validity of measures have been published for educational and psychologic tests[5] and for interdisciplinary medical rehabilitation.[4] As an essential basis for grading the quality of a measure, key information on reliability and validity needs to be systematically summarized in an evidence table. Table 1 presents a framework for such an evidence table. The framework is based on years of experience applying measurement principles and incorporates recent experience in grading the quality of outcome measures in spinal cord injury.[9] It uses terminology found in our older publication on measurement standards in rehabilitation[4] as well as concepts from item response theory (IRT) and Rasch analysis,[10] which were developed to overcome limitations of classical test theory. Other researchers have also proposed improved methods of summarizing and rating measurement evidence that go beyond classical methods (eg, Andresen[6]).

As shown in the table, essential types of information for evaluating measurement quality include the following.

1. Content validity: although essential, content validity is only a starting point. Validation of a measure by a committee of experts who agree on its content is a common procedure, but it is not enough.
2. Administrative characteristics: feasibility and cost, for instance, are critical in practice.
3. Internal consistency, reliability, reproducibility, and biases: measurement error affects virtually all uses of a measure.
4. Indicators of the internal validity of the scale and sensitivity to change: IRT and Rasch analysis provide strong indicators of the quality of a measure that can and should be used in grading measurement quality.[6]
5. Criterion-oriented validity: direct evidence is needed that a scale predicts its most important practical criterion or improves the main decision it affects or is supposed to inform.
6. Information on applicability to particular patient groups.

Understanding and grading the validity of a measure should be based on a summary of all available factual evidence.

The Need for a Conceptual Foundation

A conceptual basis is needed so that reasonable inferences and limitations to the measurement system can be understood. In past efforts to summarize measurement evidence,[9] we found that the level of evidence for a measure could be reliably graded only if the construct and its main application were clearly defined. When constructs or applications were vague, judgments of level of evidence became ambiguous; a measure could be valid for one application but of unknown validity for another application, and a scale would be judged as highly valid by a rater with one construct in mind but invalid by another rater who defined function or quality of life in a different way. In sum, the construct being measured and the application need to be defined to grade quality of measurement evidence.

Our overall grading of measurement quality is based on views of the unity of the concept of validity, emphasizing construct validity, which subsumes and integrates various aspects of validity, including content validity, reliability and internal structural characteristics, generalizability, external validity, and consequential validity.[11] However, in rating proposed UPAA measures, we particularly emphasize consequential validity (ie, usefulness in practice) because the UPAA encompasses the measurement of multiple constructs. An essential question should be answered: How well does the scale perform in its primary application? Although it is not realistic to expect that a measurement system will be completely validated for all of the uses to which it may be put, a measure should at least be validated for its main application. We will argue that judgment of level of validity is limited by the criteria and methods used in the validation study.

At the same time, indicators of the internal validity of the measure itself—content validity, reliability, and internal structure—are also a needed step not only to understand results found in application but also to understand and facilitate wider uses. Internal characteristics of a measure affect or constrain validity across multiple applications. So both reliability or internal characteristics and criterion-oriented validity need to be considered to judge the overall validity of a measure.

It is possible to synthesize these considerations into an overall grading of measurement validity. We sketch such a scheme designed to provide a simple summary of level of evidence to help readers distinguish measurement systems with stronger versus weaker supportive evidence. Although it is more refined than the traditional "reliable and valid" rubrics, the level grading (number of diamonds) should always be supplemented by explanation of the substantive strengths and limitations of the measure. Conclusions that the measure is (or is not) at 1 level of validity or another should always consider the substantive results as well as the study methodology. We apply these principles of evidence grading to the proposed UPAA later, focusing on issues of methodology and needed criterion domains.

KEY CONCEPTUAL ISSUES

Grading of the evidence for a measure begins by specifying a conceptual basis. We first address the issue of primary uses of the UPAA and then consider the major domains that need to be measured.

What is the Primary Criterion for Validation of a UPAA System?

The primary purposes of the UPAA are to serve as the basis for a payment system for postacute services, to monitor general quality across settings, and to inform "decisions for placement

Table 1: Areas for Synthesizing Information About Scales

Area	Key Items
Construct, content, and main application	Describe the construct being measured, nature of items used, and main application or use of the scale or measure.
	Subscales. Note each subscale (or dimension) using labels of author, expert, and statistical support for each subscale.
Administration	Type/mode. The basic type/nature of the measure (eg, observation, self-report, mechanism).
	Burden/cost indicators. Number of items, expense, time to administer, special setting, training, etc.
Internal consistency, reliability, and bias	Internal consistency and reliability indicators. Cronbach α is reported in CTT, but other statistics are used in modern IRT. Not applicable to single-item scales.
	Interrater, test-retest, and other reproducibility characteristics. The type of statistic needed depends on the use of the measure and the level of scaling (eg, categoric vs ordinal vs interval). For observations of performance or other judgments of complex observable phenomena, interrater reliability needs to be reported. If a measure is used to evaluate change, test-retest reproducibility, and stability should be tested.
	Bias. Bias is when a measurement procedure gives results that are systematically too high or low. Thus, it is conceivable that functional assessments are biased upward at acute hospital discharge but biased downward at subsequent postacute admission. Biases are entailed in multisetting assessments need to be studied to compare baselines and outcomes across settings.
Internal validity indicators	Core internal validity indicators. Model-fit statistics provide a central indicator of internal validity of multi-item, additive scales (eg, information function in IRT, item and person separation in Rasch analysis). Not applicable to single-item scales or categorizations.
	Sensitivity to change and ceiling floor information. Functional and health assessments in a UPAA should be sensitive to the change that would be expected in various treated groups. Tests of dynamic stability (eg, stacked vs racked Rasch analyses) can ensure that a measure has the same interval properties over time or across settings. Measurement ceiling and floor issues become critical with cross-setting measures designed to assess function and health from hospital to community. Data are needed to know whether lack of measured gain in a setting or population is because of a lack of actual improvement or a lack of scale sensitivity.
Criterion-oriented validity: prediction and discrimination	Report the predictive coefficient for the most important predictive use or criterion. Although the best possible criterion should be used, a true criterion standard may not exist; the relative quality of the criterion used should be evaluated or graded.
	Clinical utility (also called prescriptive validity). Evidence of whether and how the UPAA affects clinical decision making will be needed.
Applicability to the particular patient population	Data on the particular group. Evidence on the UPAA needs to be considered for major diagnostic and/or functional groups. Replicable problems or misfit in application should be noted. Information on frequency and type of person misfit should be reported.
	Language(s) and multicultural issues. Some items (eg, communicative function indicators, instructions to patients) are sensitive to language and culture.
	Norms. A UPAA can provide a wealth of normative information for various patient groups. Factors that most strongly affect a score, whether it be diagnosis, age, sex, race, or none of these, should be used to group scores.
	Extent of use. As a scale is used in a larger number of studies, the number of known inferences from it increases, effectively increasing the validity of the scale for various purposes. Extent of use can easily be graded.

Abbreviations: CTT, classified test theory; IRT, item response theory.

in PAC upon hospital discharge."[12(p3)] Among these, there is some consensus that improving "placement decision making" is the major purpose for the UPAA.[13] The act directing development of the UPAA emphasized "the needs of the patient and the clinical characteristics of the diagnosis to determine the appropriate placement of such patient in a post-acute care site."[7(p36)] The placement immediately after acute discharge is the first and main placement decision, although the UPAA should also inform subsequent care decisions. We focus on postacute placement recommendations and policies as the primary use of the UPAA.

Defining "placement." Does type of institutional placement (ie, inpatient rehabilitation facility (IRF), skilled nursing facility (SNF), long-term care hospital (LTCH), home health agency, and so on) map reliably into patient care needs? There is little evidence that these categories define a coherent conceptualization of level and type of patient care needs. Predicting discharge placement has proven to be remarkably difficult,[14] and nonclinical factors such as ownership and propinquity clearly affect the choice of rehabilitative setting.[15] The categories themselves are a mix of varying care capabilities. SNFs, for instance, vary greatly in the nature and quality

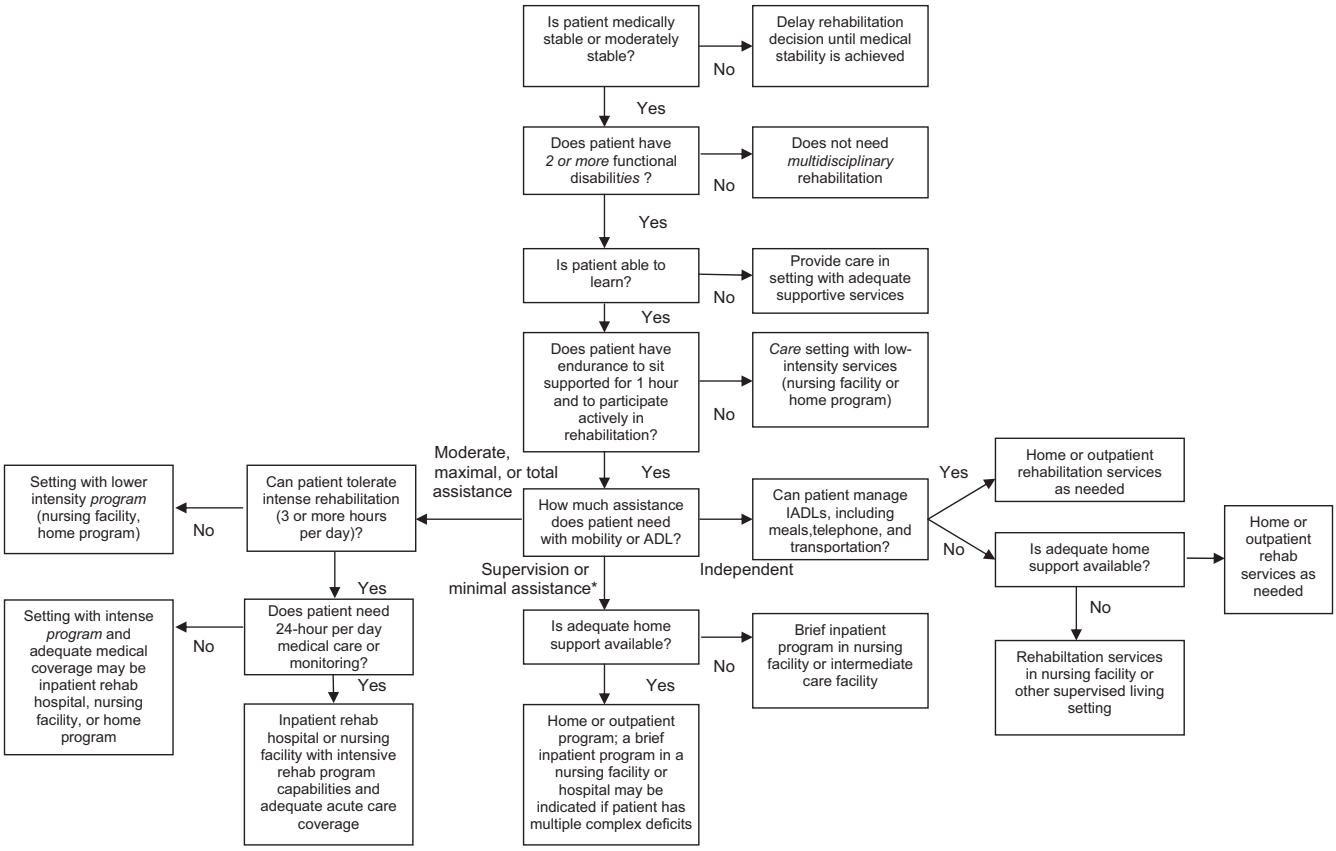

Fig 1. The core of the algorithm for rehabilitation placement of the poststroke patient. NOTE. Use of italics represents a modification from the algorithm as published. Source: Gresham et al.[19(p80-1)] Abbreviations: ADL, activities of daily living; IADL, instrumental activities of daily living. *Under special circumstances, some patients with multiple, complex, functional deficits may be appropriate for inpatient programs.

of rehabilitative services they offer because licensure requirements and services vary.[16] Postacute placement is determined not only by the clinical care needs of patients but also by a complex mix of institutional, personal, social, and political variables and values.[17] IRF patients receive both intense (24h/d) medical and nursing supervision and intense therapy; these needs are separate dimensions that correlate inversely.[18] These institutional categories involve disparate dimensions of care. The use of such internally mixed categories as validation criteria can confuse scientific efforts to improve measurement and prediction. Although it is impractical to ignore current placement categories, clinical application of the UPAA will surely require careful attention to patient care needs, not simply destinations. In sum, the term "placement" should be redefined in terms of transitions in patient care, considering skilled and unskilled medical and nursing care and rehabilitative therapies, wherever they are to be provided, as separate dimensions or domains.

What Are the Domains to Be Assessed for Rehabilitative Placement?

Many variables are relevant to rehabilitative placement.[12,13,17] Few evidence-based or even well-defined guidelines for rehabilitative placement have been published, but there is at least one: the guideline for placement of poststroke patients developed by the Agency for Health Care Policy and Research (AHCPR) (now the Agency for Health Care Research and Quality) in the 1990s.[19] The placement algorithm within it (fig 1) is generic, not mentioning stroke-specific items. Extensive work was done to develop this guideline, involving a large group of expert and experienced clinicians. Information on the reliability[18] of the placement algorithm is available, but validity studies are indirect and limited.[20,21] One would classify most of the guideline as having content validity alone, but because it is the best-developed attempt in the literature to provide an objective system for making postacute rehabilitative placement decisions, we will structure our discussions on it. The guideline implies that the following domains need to be considered to make postacute rehabilitative placement recommendations for patients with complex, disabling conditions.

Medical stability and needs for medical and nursing care and monitoring. The guideline begins with the question of whether the patient is "medically stable or moderately stable." Evaluating "medical stability" and consequent implications for medical care and monitoring is relevant to all care transitions. Medical and nursing complexity, including secondary and tertiary diagnoses as well as primary diagnosis and other indicators of severity, needs to be considered for sensible postacute placement. A patient with complex care needs will likely do better in an IRF or LTCH than an SNF or routine care nursing home. Further down the decision tree (see the left 2 columns) but linked with medical and nursing conditions, a judgment must be made of whether the person "needs" 24-hour-a-day medical care or monitoring. Nursing care is mentioned at multiple points throughout the guideline, and a judgment is required as to whether a site can adequately provide for these

needs. Although medical stability is a well-studied issue, research that quantifies the associated level and type of medical and nursing care is far more limited.

Extent of functional disabilities. Patients with severe or complex disabilities need a coordinated, multidisciplinary rehabilitation program or at least should be considered for it. Patients with a single or simpler disability (eg, lower-extremity alone) may do well with a single therapy (eg, PT). Indicators of function are found at other points in the decision tree (eg, center right); degree of independence in activities of daily living (ADLs) and instrumental ADLs largely determine whether routine self-care needs can be met at home, given the availability of paid or unpaid support. Practical methods of assessing the adequacy of family support need development.

Ability to learn, attend, and participate in self-care. These are critical skills. Simple methods of assessing this complex domain are needed because referral recommendations will be made by nurses, social workers, physical therapists, occupational therapists, and physicians in acute care settings.

Endurance and ability to participate actively in rehabilitation. This is a key domain for assessment of postacute rehabilitative needs. The stroke guideline defines "endurance to sit supported for 1 hour" as the key marker or threshold for consideration of active rehabilitative therapies.[19] Based on Medicare regulations, the ability to "tolerate" 3 hours or more of activity therapy (PT, OT) daily is treated as determining eligibility for intense rehabilitation (despite prior research raising questions about this cutpoint[22]). The ability "to participate actively in rehabilitation" can be judged if the person is given rehabilitative therapy in the acute hospital and patient participation is recorded, but we could find only 1 study[18] on the reliability of such judgments.

Patient preferences or values are a final essential domain. Although not formally considered in the decision tree, there was strong agreement among guideline experts that patients (and family) are critical.[19] Much of discharge planning in practice involves communication and compromise about alternative discharge services, feasibility, and patient-family values. The success and appropriateness of therapy depend also on the person's motivation to participate, goals, and circumstances, and so these too must be considered or measured. We do not emphasize patient satisfaction and values as primary criteria for validation of need domains in the UPAA, but we do suggest that they be incorporated into an operational UPAA system and used in reports on "appropriateness" of placements.

MEASURING THE NEED FOR REHABILITATION

Next, we address issues related to general placement, focusing on a domain that is often neglected but that is critical to rehabilitative placement, the needed intensity of active rehabilitative therapies.

Literature Search

We conducted a series of searches of PubMed, CINAHL, and Health and Psychosocial Instruments (HAPI) to identify evidence on measures or scales for evaluating or predicting the needed intensity of active rehabilitation therapies after acute hospital care. These searches were challenging because, although a large number of studies are potentially relevant, studies are not routinely indexed by measurement criteria relevant to measurement quality. For instance, a search of PubMed on terms related to *activity therapies* and *rehabilitation nursing*, *measurement and assessment*, and *rehabilitation* or *post-acute care* found a large number of studies (N=25,460), but few (n=5201) mentioned terms related to *intensity* or *dosage*. After eliminating nonadult, non-English, and laboratory studies, only 87 were found that mentioned "placement" or "care transition" or "patient discharge." By using a similar search strategy, 113 possibly relevant articles were found in CINAHL and 107 in HAPI. After reading titles and abstracts, a few relevant articles were reviewed. Relevant articles were frequently found by reading reference lists or personal recommendations.

Most research has been within setting and has not studied transitions across settings or predictors of care needs after acute hospital discharge. The literature is too disparate to provide guidelines for care transitions without extensive literature searches for each impairment, pathology, or disability, and much expert opinion would be required to synthesize implications for PAC needs. Local coverage determinations developed by Medicare's fiscal intermediaries and based on informed expert opinion, tradition, and rationale are also relevant. They show the complexity of developing defensible care guidelines and the inconsistent decisions that result when using current processes.[23]

A number of studies[14,17,24,25] have identified general predictors of home and/or SNF placement for various groups. We also found studies that identified statistically significant predictors of rehabilitative discharge destination (eg, age, home support, number of comorbid conditions after hip and knee arthroplasty,[26] severity of brain injury[27]). In addition, descriptive studies[28] have identified problems and considerations in the discharge process. The extent of family support is a critical item noted if not well-assessed in all postacute needs assessment systems.[12,13,29] Such studies provide valuable information on factors involved in rehabilitative placement decisions, but more evidence is needed to provide a sound basis for clinically precise and prescriptive guidelines.

Only in the case of routine, long-term care needs is there substantial validity literature; attempts have long been made to develop scales of needed level of long-term care,[30] and ADL disability scales are designed to measure the unskilled care needs or dependencies in this domain.[31] Other articles in this issue address considerations involved in choosing a scale of basic ADL function that builds on the FIM instrument, the so-called Minimum Data Set for Medicare nursing homes, and the Outcome and Assessment Information Set for home health agencies.

Factors related to the needed intensity of rehabilitative therapies. Multiple factors are relevant to the ability to participate in and profit from active rehabilitation therapies. Functional performances themselves involve multiple dimensions (eg, the correlated dimensions of speed/coordination and endurance/strength[32]), and participation in rehabilitative therapies requires not only physical activity and endurance but also attention or mental effort. Pain and medical stability also constrain participation in therapy. Approaches to measurement have ranged from single ratings of global participation[33] to multiple formal performance tests.[32] Instrumented methods have been used (eg, to measure heart rate and oxygen consumption to assess exercise tolerance for cardiac rehabilitation).[34]

Our literature searches found little direct evidence on the issues of needed "intensity," "amount," "level," or "dose" of active rehabilitative therapies (PT, OT, SLP, rehabilitation nursing, or nursing education) or on the validity of methods of quantifying major constructs that should affect this domain of concern. Extensive searches of Medline and CINAHL yielded 43 to 87 articles depending on search details. Most of these were only tangentially or marginally relevant, but a few studies deserve mention. Keith[35] has published a theoretical article on

treatment strength in rehabilitation. In an observational study, Duncan[20] and Reker[21] and colleagues have shown that postacute treatment of stroke patients according to AHCPR rehabilitative guidelines is associated with better functional outcomes[20] and greater patient satisfaction[21] than placements that depart from these guidelines. Johnston et al[18] evaluated the reliability of the AHCPR guideline for postacute placement of stroke patients. The placement guideline was found to be reasonably reliable for home and nursing facility placement with rehabilitation therapy services, but it was not reliable for placements in low-intensity outpatient rehabilitation and high-intensity inpatient rehabilitation programs.[18(p539)] Kramer et al[36] have shown that rehabilitation in a more intense setting, inpatient rehabilitation hospitals, leads to somewhat improved functional outcomes compared with less intense care in nursing homes for stroke patients but not for hip fracture patients, but characteristics of patients who do best in the alternative settings remain unclear. No research was found that provided what we would label as strong evidence for needed intensity of rehabilitative therapies; we lack data on the accuracy with which various measurement procedures predict participation in and benefit from rehabilitative activity therapies.

Levels of continuing medical and nursing care needs. Projecting continuing medical and nursing care needs remains a critical related issue. Although the literature on medical and nursing care needs within institutional settings is huge, the literature on the best methods of forecasting care needs in subsequent settings is small. Nursing care needs can be estimated or approximately measured, and although nursing needs correlate with ADL dependency, the 2 are not the same.[37] Dependency at discharge can be scaled reliably and predicts subsequent resource use in home care.[38] Variables predicting use of specialized discharge planning services have been identified (eg, age, disability, living alone, walking limitation).[39] Patients can be stratified into levels of risk for postdischarge medical services on the basis of patient age and the Medical Outcomes Study 36-Item Short-Form Health Survey physical function and social function scores.[40]

The development and validation of methods of evaluating transitions in needed level of medical and nursing care and supervision is a great issue. We did not review this complex literature further, but the issues involved are central to PAC transitions in both rehabilitative and alternative settings. Westra et al's study[24] of the Uniform Needs Assessment Instrument shows that it is possible to provide a standard set of information and improve the identification of subsequent patient needs.

Results of our search support Coleman's summary that "Currently, there are no evidence-based criteria to inform hospital personnel in determining what the appropriate PAC setting is for a beneficiary with a known set of conditions and skilled care needs."[13(p88)] Indeed, defining and validating assessments of "skilled care needs" is a challenge in itself. Work is needed to develop the evidence basis for rehabilitative and other care transitions.

Types and Levels of Needed Validation Studies

The level of evidence for a measure depends on the validation methodologies used, including not only the evidence of content relevance, reliability, and internal structure but also predictive or criterion-referenced validity. It is commonly said that measures should be validated against a criterion standard, but, in fact, a well-validated criterion may not exist. The quality of the criterion used needs to be evaluated and is 1 (but not the only) critical indicator of level of evidence for a measure.

Table 2: Overall Summary Rating of Measurement Validity for an Application

Graphic	Labels and Brief Descriptions
○	Preliminary or questionable: opinion-based content validity alone.
◆	Minimal: in addition to content validity, data indicate acceptable internal structure and reliability, but there is little or no direct evidence of validity for the application. Alternatively, evidence shows limited or unclear practical validity without determining reliability, error rates, or biases.
◆◆	Moderate or fairly adequate: in addition to the above, data indicate that the scale performs reasonably well for its main-defined purpose (eg, predicting a criterion indicating appropriate patient placement), and evidence on reliability, error rates, and biases are provided. Although the scale is reasonably "valid" for this use, questions remain (eg, about common generalization or extensions or limitations to use).
◆◆◆	Strong: Strong associations (with small confidence intervals) are found to be an important criterion standard in diverse settings and populations of clinical or community practice; reliability, errors, bias, and limitations are also documented.

In the following sections, we will sketch several types of criterion validation studies and suggest how they might be graded into levels by using graphical symbols to provide a simple summary (table 2).

Preliminary Level: Content Validity

Content validity implies a process in which domains and items are generated and carefully reviewed by a group of experts. In this case, one would choose clinicians experienced in providing recommendations for postacute placement and therapy intensity and experts who are familiar with the relevant literature. This procedure is used by Medicare's fiscal intermediaries in promulgating local coverage decisions, but expert opinion alone does not provide a satisfactory level of evidence from a scientific perspective. The suggested label for this level of evidence is *preliminary* (○).

Minimal Validity Level

Reliability and internal structural validity. Knowledge of the degree of error or unreliability is needed to use a measure and interpret results. If the proposed measure includes multiple items that are to be summed, analyses of internal structure are needed to identify the actual dimensions or subscales in the scale and to test whether summing makes sense.[4,5,41] Single-item scales or categorizations have no internal structure, but reliability still needs to be evaluated. Knowledge of internal structure and reliability is an important addition to content validity, increasing knowledge of likely generalizability and validity, but is far from enough to provide a strong recommendation for a measure.

Initial concurrent validation. Direct but weak evidence of criterion validity of a scale is available, but needed reliability or internal validity analyses have not been done. For instance, a study might predict clinical judgments or actual placements but not very accurately or without evidence that the criterion used is sensitive to patient needs. A UPAA must be validated against actual treatment provided and clinical judgment, but, if the circumstances are

undefined, we cannot distinguish whether placements were suboptimal or the UPAA screen was flawed. Concurrent validation against such a pragmatic criterion provides useful information, but, without good evidence that the criterion is appropriate, validation of the new scale is still inadequate. A label for this level of evidence would be *minimal*, specifying reliability, and/or initial concurrent validity (◆).

Moderate Level: Fairly Adequate Evidence

Current good or best practices. Current practices can provide a useful criterion for validation of a measurement system, provided the quality of care and placement decisions can be justified. Treatment and placement decisions made for patients with good funding by experienced clinical teams in a region with a variety of placement alternatives, excluding ownership and inappropriate financial biases, can provide a criterion for UPAA validation. Inclusion and exclusion and/or statistical control procedures might be used to reduce factors not reflecting patient care needs.

The judgment of experienced clinical professionals, typically nurses, therapists, and/or clinical case managers, who have evaluated the patient and readiness for therapy or care transitions is arguably the best available criterion for appropriate placement or care transitions. The difficulty with this criterion is that "clinical judgment" is ill defined and varies across clinicians and by local practice area. Nonetheless, the field will want to know the relationship between assessments based on the UPAA and current best practices. Studies comparing UPAA projections to well-informed clinical judgments may be expected to improve inferences from the UPAA and knowledge of its strengths and limitations.

Improvement in function. Measured outcomes also provide needed criteria for the validation of rehabilitative placement systems. If improved patient function is used as a criterion in an observational outcomes study, then it is highly important to control for established indicators of case severity in rehabilitation, such as functional severity at baseline assessment, age, demographic factors, diagnosis, and comorbid conditions.[42,43] Control for selection biases can greatly enhance the validity of studies of desired health outcomes.[43] Participation in therapy should also be considered in such a criterion validation study; a valid screen or measure should discriminate cases that fully engage in active therapy from those that do not. A suggested label for this intermediate level of evidence is *moderate* or *fairly adequate* (◆◆).

High Level: Strong Evidence

A strong criterion for a screen or measure of needed treatment would be provided by the selection criteria used in a randomized controlled trial (RCT) that showed that the treatment was effective at improving valued health or functional outcomes in a typical practice setting. Evidence should also be provided that the treatment was necessary in the sense that it cannot be provided in a less intense setting. Information on the cost-effectiveness of alternative interventions and the value of outcomes attained is needed too, but perfect information should not be demanded of any measurement or decision-support system.

The availability of strong evidence can simplify the validity study; one examines the assessment system to see if it has the selection criteria used in the RCT. In practice, a general-purpose assessment system is unlikely to provide the precise items found in a validated clinical assessment but will provide correlated or partial information related to some of the selection criteria in the RCT (ie, it will function as a screen). There are established methodologies for the validation of a screening test.[2] One can test how well the UPAA predicts whether the patient is a candidate for the established treatment (eg, computing positive and negative predictive values). A suggested label for this level of evidence is *strong* (◆◆◆).

Study Limitations

We concentrated on major conceptual matters that are prerequisite to evaluating the quality of measures and to planning research. The quality of continuing PAC is a major application of a UPAA, but we addressed only the first, critical step, the initial postacute placement. We also focused on major methodologic issues in grading the level of evidence for the UPAA, but we did not attempt to present specific numeric criteria for required levels of reliability, validity, or cost-effectiveness.

CONCLUSIONS AND RECOMMENDATIONS

The UPAA tool has the potential to provide invaluable information to guide the development of a more organized and cost-effective system of postacute patient care. However, the UPAA tool needs an evidence basis. Principles of reliability and validity can be refined, creating a framework for grading the level of evidence for measures and assessment systems. Applying such a framework to evaluation of the UPAA yields certain insights. First, the conceptual basis for constructs involved and chief application of a measure need to be clearly defined to evaluate the evidence for a measure and its application.

Second, the UPAA and guidelines based on it need to be validated against patient outcomes and appropriateness of placements. We have suggested that "placement" is more meaningfully defined in terms of the level and type of care than administratively convenient labels based on type of institution. The needed intensity of rehabilitative therapies should be distinguished from the needs for skilled medical and nursing care, routine ADL care, and other factors as criteria to validate placement appropriateness.

Third, the quality of the outcome criterion used in validation studies is a critical consideration in grading the evidence for an assessment system.

Fourth, there is almost no evidence on the validity of methods to evaluate the needed intensity of rehabilitative therapies for use at discharge from the acute hospital. The absence of direct evidence on needed levels of rehabilitative therapies is a major lacuna that surely must be remedied.

An a priori framework for evaluating level of measurement evidence will provide a more rational and scientifically defensible basis for judging the adequacy of the UPAA and policies based on it than expert opinion or ad hoc judgments. We recommend that a process of systematic review, based on graded criteria for level of evidence, be initiated to evaluate accumulating evidence on the validity of the UPAA and other such assessment systems. Continued over the long-term, such a process will identify both progress and deficiencies in the assessment system and provide the factual basis needed for policy debates and guideline development.

Acknowledgments: We thank Allen Heinemann, PhD, and Margaret Crater, MA, for their editorial assistance and insightful comments.

References

1. Higgins JP, Green S. Cochrane handbook for systematic reviews of interventions 4.2.6. Chichester: John Wiley & Sons; 2006.
2. Edlund W, Gronseth G, So Y, Franklin GM, American Academy of Neurology clinical practice guideline process manual. St Paul: American Academy of Neurology; 2004.
3. Sackett DL, Straus SE, Richardson WS, Rosenberg WM, Haynes RB. Evidence-based medicine: how to practice and teach EBM. Edinburgh: Churchill Livingstone; 2000.

4. Johnston MV, Keith RA, Hinderer SR. Measurement standards for interdisciplinary medical rehabilitation. Arch Phys Med Rehabil 1992;73(12 Suppl):S3-23.
5. American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (U.S.). Standards for educational and psychological testing. Washington (DC): American Educational Research Association; 1999.
6. Andresen EM. Criteria for assessing the tools of disability outcomes research. Arch Phys Med Rehabil 2000;81(12 Suppl 2): S15-20.
7. Deficit Reduction Act of 2005, Pub L No. 109-171, 120 Stat. 37, §5008 (2005).
8. Policy Council, Centers for Medicare & Medicaid Services, U.S. Department of Health and Human Services. Post acute care reform plan. September 28, 2006. Available at: http://www.cms.hhs.gov/SNFPPS/Downloads/pac_reform_plan_2006.pdf. Accessed August 27, 2007.
9. Johnston M, Graves D. Towards guidelines for evaluation of measures: an introduction with application to spinal cord injury. Presented to: The American Spinal Injury Association; 2006 June 24; Boston (MA).
10. Andrich D. Controversy and the Rasch model: a characteristic of incompatible paradigms? Med Care 2004;42(1 Suppl):I7-16.
11. Messick S. Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. Am Psychol 1995;50:741-9.
12. Kramer A, Holthaus DE. Uniform patient assessment for post-acute care. Final report. Aurora: Div Health Care Policy and Research, Univ Colorado at Denver and Health Sciences Center; 2006.
13. Coleman E. Hospital discharge assessment and data items that facilitate appropriate placement and efficient care transitions. In: Kramer AH, Danielle, editors. Uniform patient assessment for post-acute care. Aurora: Div Health Care Policy and Research, Univ Colorado at Denver and Health Sciences Center; 2006. p 86-99.
14. Kane RL, Finch M, Blewett L, Chen Q, Burns R, Moskowitz M. Use of post-hospital care by Medicare patients. J Am Geriatr Soc 1996;44:242-50.
15. Buntin MB, Garten AD, Paddock S, Saliba D, Totten M, Escarce JJ. How much is postacute care use affected by its availability? Health Serv Res 2005;40:413-34.
16. Kaplan SJ. Growth and payment adequacy of postacute care rehabilitation. Arch Phys Med Rehabil 2007;88:1494-9.
17. Ottenbacher K, Graham JE. The state-of-the-science: access to postacute care rehabilitation services. Arch Phys Med Rehabil 2007;88:1513-21.
18. Johnston MV, Wood K, Stason WB, Beatty P. Rehabilitative placement of poststroke patients: reliability of the Clinical Practice Guideline of the Agency for Health Care Policy and Research. Arch Phys Med Rehabil 2000;81:539-48.
19. Gresham G, Duncan P, Stason W, et al. Post-stroke rehabilitation. Clinical practice guideline no. 16. Washington (DC): Agency for Health Care Policy and Research; 1995.
20. Duncan PW, Horner RD, Reker DM, et al. Adherence to postacute rehabilitation guidelines is associated with functional recovery in stroke. Stroke 2002;33:167-77.
21. Reker DM, Duncan PW, Horner RD, et al. Postacute stroke guideline compliance is associated with greater patient satisfaction. Arch Phys Med Rehabil 2002;83:750-6.
22. Johnston MV, Miller LS. Cost-effectiveness of the Medicare three-hour regulation. Arch Phys Med Rehabil 1986;67:581-5.
23. Fiedler I. Inpatient rehabilitation facilities: standards for criteria for patient admissions. Paper presented to: State of the Science Symposium on Post-Acute Rehabilitation: Setting a Research Agenda and Developing an Evidence Base for Practice and Public Policy; 2007 Feb 12; Arlington (VA).
24. Westra BL, Holland DE, Aufenthie J, et al. Testing the Uniform Needs Assessment Instrument for hospital discharge planning with older adults. J Gerontol Nurs 1998;24:42-6.
25. Melchiorre PJ. Acute hospitalization and discharge outcome of neurologically intact trauma patients sustaining thoracolumbar vertebral fractures managed conservatively with thoracolumbosacral orthoses and physical therapy. Arch Phys Med Rehabil 1999; 80:221-4.
26. Munin MC, Kwoh CK, Glynn N, Crossett L, Rubash HE. Predicting discharge outcome after elective hip and knee arthroplasty. Am J Phys Med Rehabil 1995;74:294-301.
27. Barnes EF, Frank EM, Montgomery A, Nichols M. Factors predicting rehabilitative service provision in adults with traumatic brain injury. J Med Speech Lang Pathol 2005;13:69-84.
28. Tyson S, Turner G. Discharge and follow-up for people with stroke: what happens and why. Clin Rehabil 2000;14:381-92.
29. Gage B, Green J. State of the art: current CMS PAC Instruments. In: Kramer A, Holthaus DE, editors. Uniform patient assessment for post-acute care. Final report. Aurora: Div Health Care Policy and Research, Univ Colorado at Denver and Health Sciences Center; 2006. p 15-38.
30. McKenzie DA, Capuzzi CF, Will SJ. Alberta Assessment and Placement Instrument. Description and interrater reliability. Med Care 1989;27:937-41.
31. McDowell I. Physical disability and handicap. Measuring health: a guide to rating scales and questionnaires. 3rd ed. Oxford: Oxford Univ Pr; 2006. p 50-149.
32. Novy DM, Simmonds MJ, Lee CE. Physical performance tasks: what are the underlying constructs? Arch Phys Med Rehabil 2002;83:44-7.
33. Lenze EJ, Munin MC, Quear T, et al. The Pittsburgh Rehabilitation Participation Scale: reliability and validity of a clinician-rated measure of participation in acute rehabilitation. Arch Phys Med Rehabil 2004;85:380-4.
34. Bitzer EM, Aster-Schenck IU, Klosterhuis H, Dörning H, Rose S. [Developing evidence based guidelines on cardiac rehabilitation—phase 1: a qualitative review] [German]. Rehabilitation (Stuttg) 2002;41:226-36.
35. Keith RA. Treatment strength in rehabilitation. Arch Phys Med Rehabil 1997;78:1298-304.
36. Kramer AM, Steiner JF, Schlenker RE, et al. Outcomes and costs after hip fracture and stroke. A comparison of rehabilitation settings. JAMA 1997;277:396-404.
37. Turner-Stokes L, Tonge P, Nyein K, Hunter M, Nielson S, Robinson I. The Northwick Park Dependency Score (NPDS): a measure of nursing dependency in rehabilitation. Clin Rehabil 1998; 12:304-18.
38. Edwardson SR, Nardone P. The dependency at discharge instrument as a measure of resource use in home care. Public Health Nurs 1990;7:138-44.
39. Holland DE, Harris MR, Leibson CL, Pankratz VS, Krichbaum KE. Development and validation of a screen for specialized discharge planning services. Nurs Res 2006;55:62-71.
40. Fairchild DG, Hickey ML, Cook EF, et al. A prediction rule for the use of postdischarge medical services. J Gen Intern Med 1998;13:98-105.
41. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. Oxford: Oxford Univ Pr; 2003.
42. Iezzoni LI. Risk adjustment for measuring healthcare outcomes. Chicago: Health Administration Pr; 2003.
43. Kane RL. Understanding health outcomes research. 2nd ed. Boston: Jones & Bartlett; 2005.