J Rehabil Med 2011; 43: 661–672

EDUCATIONAL REVIEW

STRATEGIES FOR ASSESSMENT AND OUTCOME MEASUREMENT IN PHYSICAL AND REHABILITATION MEDICINE: AN EDUCATIONAL REVIEW*

Ayşe A. Küçükdeveci, MD[1], Alan Tennant, PhD[2], Gunnar Grimby, MD, PhD, FRCP[3] and Franco Franchignoni, MD[4]

From the [1]University of Ankara, Faculty of Medicine, Department of Physical Medicine and Rehabilitation, Turkey, [2]University of Leeds, Department of Rehabilitation Medicine, Faculty of Medicine and Health, UK, [3]Section of Clinical Neuroscience and Rehabilitation, Sahlgrenska Academy at University of Gothenburg, Sweden, and [4]Unit of Occupational Rehabilitation and Ergonomics, Salvatore Maugeri Foundation, Clinica del Lavoro e della Riabilitazione, IRCCS, Veruno (NO), Italy

The aim of this educational review, which is based upon expert opinion, is to describe to clinicians training in Physical and Rehabilitation Medicine and research students training to work in the field, the appropriate attributes and standards required for assessment and outcome measurement. "What to assess" is discussed in the context of the conceptual framework provided by the International Classification of Functioning, Disability and Health, supplemented with quality of life as an additional construct. The reasons for making the assessment, and the context in which the assessment will be used, are then considered. Examples of recommendations of some international organizations regarding what and how to assess are presented. Suggestions are made about the selection of assessment tools, including examples from two diagnostic groups: stroke and rheumatoid arthritis. Finally, the basic psychometric standards required for any assessment tool, and additional requirements for outcome assessment, are explained.

Key words: outcome assessment; psychometrics; rehabilitation; ICF.

J Rehabil Med 2011; 43: 661–672

Correspondence address: Ayşe A. Küçükdeveci, Ankara 85 Sitesi, 176. Sok, No:12, Beysukent, TR-06800 Ankara, Turkey. E-mail: ayse@tepa.com.tr

Submitted July 20, 2010; accepted May 23, 2011

INTRODUCTION

Physical and Rehabilitation Medicine (PRM) aims to enable persons with health conditions who experience, or are likely to experience disability, to achieve and maintain optimal functioning in interaction with the environment (1). Rehabilitation is a problem-solving and educational process that requires the use of assessments in order to identify the relevant problems. The term "assessment" includes techniques and procedures for classification and measurement of a variable pertaining to a person (2). Measurement is quantification of an observation by a standard unit. A wide variety of assessments is used in PRM across a variety of clinical or community settings, and these can be undertaken by a wide range of professionals and, in some circumstances, may be self-completed by patients. An assessment becomes a potential "outcome measure" when it is associated with the result of an intervention of some kind. In other words, "outcome" is defined as change in a state or situation that arises as a result of some process or intervention (3).

The aim of this educational review, which is based upon expert opinion, is to describe to clinicians training in PRM and research students training to work in the field, the appropriate attributes and standards required for assessment and outcome measurement. As such, every specialty will have its own emphasis on the type of assessment to be undertaken. These may vary by setting, such that those applied in an acute phase may differ from those applied in the community. In most cases the type of information being assessed can be catalogued by the International Classification of Functioning, Disability and Health (ICF) framework of the World Health Organization (WHO) (4). The ICF systematically classifies health and health-related states into two components: (*i*) body functions and structures; and (*ii*) activities and participation. Body functions are the physiological functions of body systems, whereas body structures are anatomical parts of the body. Impairments are problems in body function or structure, such as a significant deviation or loss. Impairments of body function are essentially signs and symptoms, and impairments of body structure represent aspects such as cranial nerve injury, or musculoskeletal damage. Activity is the execution of a task or action by an individual, whereas participation is involvement in a life situation. Activity limitations are difficulties an individual may have in executing activities such as dressing or feeding, transfer and mobility. Participation restrictions are problems an individual may experience in involvement in life situations, such as work or family life. According to the ICF framework, functioning is an umbrella term including body functions and structures and activities and participation. Disability is an umbrella term including impairments, activity limitations and participation restrictions. The ICF also incorporates environ-

*This is an educational review which is produced in collaboration with UEMS European Board of Physical and Rehabilitation Medicine.

This article has been fully handled by one of the Associate Editors, who has made the decision for acceptance, as the Editor-in-Chief being a coauthor.

 doi: 10.2340/16501977-0844

 ISSN 1650-1977

mental factors that interact with all of these constructs. "Personal factors" are also indicated, but as yet are not defined. In addition to the components of ICF, assessment of quality of life (QoL) can also be performed through concepts such as well-being, life satisfaction, or other models widely used, including the needs-based QoL scales (5–7). The conceptual model for functioning is shown in Fig. 1, modified from the original ICF schema to include potential aspects of environmental and personal factors, and to show QoL as an additional construct.

In considering an assessment of one or more of these constructs, either for a one-off need for clinical management, or on repeated occasions, and possible use as an outcome, a number of questions should be asked. These include what, why, and how should the assessment be made.

WHAT SHOULD BE ASSESSED?

The assessment can be at various levels, as classified below.

Body functions

Body functions, including psychological functions, are classified systematically into 8 sections in the ICF (4). Body functions that require assessment in most musculoskeletal conditions are: pain, mobility of joints, stability of joints, muscle power, muscle tone, muscle endurance, energy, sleep, emotional functions, exercise tolerance, gait pattern and sexual functions. Assessments of body functions in neurologically disabled people should also include: cognitive functions (consciousness, orientation, attention, memory, language, perception), touch and other sensory functions, voice and speech functions, defecation, urination and control of voluntary movement. The assessment of pain intensity by the visual analogue scale (VAS) or the Multidimensional Pain Inventory (MPI) (S1, S2), the Mini-Mental State Examination (MMSE) for some cognitive functions (S3), and the Modified Ashworth Scale (MAS) for muscle tone (S4) are examples of widely used assessments of body functions. Diagnosis-specific assessments can also be performed, such as haematological assessment (acute phase reactants) in rheumatoid arthritis (S5), or blood pressure measurement, or the evaluation of motor and sensory function by the Fugl-Meyer Assessment in stroke (S6).

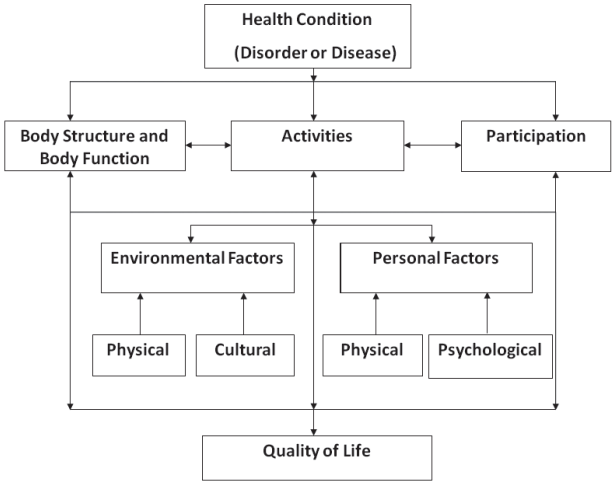

Fig. 1. The bio-psychosocial model of the International Classification of Functioning, Disability and Health (ICF) expanded to include quality of life.

Body structures

These can be assessed either by physical examination or by various imaging techniques. Joint deformities, muscle atrophy, structural impairments of various musculoskeletal regions determined by X-rays or other imaging methods, structural impairments of brain or spinal cord demonstrated by various imaging technique and pressure ulcers of the skin are examples of impairments of body structures usually assessed in the field of PRM. Radiological assessment scales, such as the Larsen Index (S7) or Bath Ankylosing Spondylitis Radiology Index (BASRI) (S8), and pressure ulcer grading scales (S9) are examples of scales used for the assessment of body structures.

Activities

Activities are basic tasks or actions that represent the individual perspective of functioning. Assessments can be made of performance; that is, what an individual does in his or her current environment; or of capacity, which describes an individual's ability to execute a task or an action and ought to be done in a "standardized" environment. Although moderate to high correlations have been observed between capacity and performance, environmental and personal factors (such as motivation) have a great impact on the performance of activities (8, 9). Therefore, differentiating between capacity and performance of patients can contribute to making decisions in the rehabilitation process. In this respect, for example, the timed walking test in a standardized environment may ascertain capacity, whereas the measurement of everyday physical activity with an accelerometry-based activity monitor can be helpful to identify the performance levels of individuals in their natural environment (10).

Although in the ICF, the domains in "Activities and Participation" are given as single list and the components of "Activities" and "Participation" are not distinguished, it is also possible to designate some domains as activities and others as participation. For example, in PRM, it would be reasonable to operationalize "Activities" as a separate level of assessment. In this case, the domains, learning and applying knowledge, general tasks and demands, communication, mobility, self-care and, to some extent, domestic life, could be considered as "Activities". Most of the assessment tools used in the PRM field assess such activities (11, 12). The assessment may focus upon a special activity, such as mobility or dexterity, or a combination of such activities. For example, the Rivermead Mobility Index assesses mobility (S10), whereas the Nine-Hole Peg Test evaluates dexterity (S11). The Barthel Index (S12) and the Health Assessment Questionnaire (S13) are examples of generic assessment tools for physical activities of daily living, whereas the Functional Independence Measure (FIM™) evaluates both physical and cognitive aspects of daily life (S14).

Participation

Participation represents the societal perspective of functioning, and includes interpersonal interactions and relationships, life activities, such as domestic life, education, work and employment, and community, social and civic life. In the past, there was infrequent assessment of participation, although there were some available assessment tools based on the International Classification of Impairments, Disabilities and Handicaps (ICIDH), such as the London Handicap Scale (S15) or the Craig Handicap Assessment and Reporting Technique (S16). However, as participation is a key outcome in rehabilitation, its assessment is important. In this respect, there has been an increased development and use of participation measures based on the ICF (13, 14), such as the Impact on Participation and Autonomy Questionnaire (IPAQ) (S17) or Participation Profile (S18).

Quality of life

QoL is defined by the WHO as "individuals' perceptions of their position in life in the context of the culture and value systems in which they live and in relation to their own goals, expectations, standards and concerns" (S19). In recent years, there has been increasing interest in using QoL as an outcome measure in the field of PRM (6). Although health professionals and QoL experts tend to agree upon the subjective nature of the QoL, there has been no consensus on its definition. Factors making up QoL are broadly reflective of physical and material well-being, personal development and fulfilment, relations with other people, recreation, and, social, community, and civic activities (S20). There are two competing paradigms for the assessment of QoL and related concepts. The first argues that it is a multi-domain concept and is influenced by numerous factors, some of which may be unrelated to the individual's health or disease. The second paradigm for QoL is based upon more specific constructs such as subjective well-being or life satisfaction (5, 15). Thus, the differences between the paradigms can be viewed as the former being concerned with health status (or health-related QoL (HRQoL)) and measured often by a profile of several dimensions, whereas the latter is concerned with the subjective impact of the condition, usually measured by a single construct.

Health-related QoL (HRQoL) relates to the first paradigm and refers to the extent to which one's usual or expected physical, emotional and social well-being is affected by a medical condition or its treatment (16). There are 4 main different types of measures used to assess HRQoL (17). Generic health profiles, such as the Medical Outcomes Study Short Form 36 (SF-36) (S21) or Nottingham Health Profile (S22), provide a broad set of components related to HRQoL. Generic utility measures that have been developed for economic evaluation are derived from individuals' preferences for different health states and are expressed as a single number along a continuum from 0 "representing death" to 1 "representing perfect health" (e.g. EuroQoL) (S23). Disease- or condition-specific measures focus on aspects of health that are relevant to a particular health condition, for example the Fibromyalgia Impact Questionnaire (S24). Finally, there are individualized measures that allow the respondent to select and weight the most important areas of his own life, such as the Patient Generated Index (S25) or the Schedule for the Evaluation of Individual Quality of Life (SEIQoL) (S26).

The second paradigm for QoL is based upon largely single domains, such as "subjective well being" or, for example, the "needs-based" model as developed by McKenna & Doward (7). Another example is the assessment of "life satisfaction" (15). These approaches concentrate upon the persons' subjective experience of the impact of the condition, including, for example, the impact that the condition has upon meeting the individual's needs. The approach allows for an interpretation of the impact of health status (e.g. impairment and activity limitation) upon the individual's quality of life, as depicted in Fig. 1.

In summary, in choosing what to assess, it is therefore crucial that the conceptual framework is fully understood and, for outcome measurement, thought is given to where the likely impact of treatment will be found. If the focus is upon an impairment of body function, then the primary outcome should be associated with that function. If, on the other hand, the focus is upon the individual's role in society, then a measure of participation restriction would be appropriate as the primary outcome. However, in all cases, it is also possible to consider the wider bio-psychosocial model of the ICF in more detail and allow, for example, an examination of the mediating role of psychological factors upon the impact of pain upon activity limitation in low back pain. Thus the conceptual framework provided by the ICF, particularly where it is modified to include QoL as an additional construct, as in Fig. 1, provides a fertile environment for hypothesis generation and testing.

WHY SHOULD THE ASSESSMENT BE MADE?

Another step is to determine why the assessment is to be made. There might be various reasons for this.

Clinical decision-making in individual patients

The rehabilitation process can be considered to include 4 stages: assessment, goal-setting, intervention and re-assessment (18). In the assessment stage, the presence and the severity of the patient's problems (impairments, activity limitations, participation restrictions), prognostic factors and the patient's wishes and expectations are identified. Considering all these identified, short-term and long-term goals for the patient are established in the goal-setting stage. Then, at the intervention stage, all supportive and therapeutic interventions are undertaken according to the goals set. At the re-assessment stage, the effects of interventions against the goals set are evaluated. The process is iterative and if there are still problems requiring intervention, the cycle continues until the goals are achieved, and/or new goals are set. At most stages of this rehabilitation process, the rehabilitation team uses various assessment tools to establish the presence and the severity of problems, inform intervention planning and monitor progress, and predict recovery and discharge planning (19). Using standard assessment tools enhances communication among the team members.

A wide range of outcome measures are used in PRM concerning different aspects of disability (11). For example, a range

of impairment assessment tools can be used by one or more members of the rehabilitation team in routine patient care (12, 19, 20). Examples of impairment assessment tools used in clinical practice include the Modified Ashworth Scale for spasticity (S4), the Waterlow score for pressure ulcer risk (S27), the Multidimensional Pain Inventory (S2) for pain intensity, the Mini-Mental State Examination for cognitive screening (S3), the Boston Diagnostic Aphasia Examination for speech and language functions (S28), and the American Spinal Injury Association (ASIA) Impairment Scale for motor and sensory functions in spinal cord injury (S29). In addition, other instruments may be used to assess for conditions such as depression, for example, the Beck Depression Inventory (S30).

The most commonly used assessment tools for routine clinical practice in rehabilitation centres are concerned with activity limitation, and include the Barthel Index and the FIM™ (12, 21). These are generic activity limitation scales and used for decision-making by the whole rehabilitation team. However, they may not be sufficient for use in outpatient rehabilitation setting (because they often approach the ceiling within the in-patient setting). There are other scales used to measure activity limitations, which are more appropriate in an outpatient setting, including the Frenchay Activities Index, and the Nottingham Extended Activities of Daily Living Index, both of which were originally developed for patients with stroke, although which have been used in a variety of conditions (S31, S32, S33). Examples of specific activity assessment scales are the 10-m walk test for walking (S34), the Berg Balance Scale for balance (S35) or the nine-hole peg test for dexterity (S11).

Clinical audit

Clinical audit is a quality improvement process that seeks to improve patient care and outcomes through systematic review of care against explicit criteria and the implementation of change. It has been defined as "the systematic, critical analysis of the quality of medical care, including the procedures used for diagnosis and treatment, the use of resources, and the resulting outcome and quality of life for the patient" (22). Where indicated by the audit, changes are implemented at an individual, team, or service level and further monitoring is used to confirm improvement in healthcare delivery (23). The key component of clinical audit is that performance is reviewed (or audited) to ensure that what *should* be done is *being* done, and if not, it provides a framework to enable improvements to be made (24). Aspects of the structure, process and outcome are addressed in this audit or evaluation process (23, 25) Structural evaluation refers to resources required, such as the information about the numbers and skills of the staff and the provision of equipment and physical space. Process evaluation refers to the actions of healthcare providers, including information about assessments, investigations, all therapeutic interventions and documentation. Finally, outcome evaluation refers to the results or the outcomes of the interventions, which reflect the effects of both structure and the process. Outcome evaluation requires the use of outcome criteria or outcome measures. In PRM, outcome evaluation in clinical audit is usually performed by comparing the functional performance of the patient before and after the intervention (25, 26).

The Uniform Data System for Medical Rehabilitation (UDSMR) and the Patient Evaluation and Conference System (PECS) are examples of measurement systems developed and used in the USA for outcome evaluation in medical rehabilitation (25). The UDSMR was founded in 1987 to serve as a repository and provide data management functions for information about inpatient medical rehabilitation throughout the USA. The FIM™, which is now widely used at the international level is a part of the UDSMR, and measures the level of functional independence of medical rehabilitation patients in terms of basic daily activities. Outcome measures used routinely for clinical practice and data collection are usually included as outcome criteria in clinical audit (19, 26). In addition to the level of functional independence, other parameters such as the length of stay, complications or discharge placement, are also used as outcome criteria (27–29). In the context of quality improvement or audit, the choice of outcome measure requires full discussion and agreement by the service provider and the service purchaser as they may have different goals (3).

Research study

One of the main fields where assessment is inevitable is for research. Various assessments or outcome measurements might be required in both clinical and epidemiological studies. The choice of the assessment tool depends on the type and the aim of research being made. If it is a clinical trial investigating the efficacy of an intervention, then an outcome measure that can capture the expected effect should be used. However, in the field of PRM, interventions are usually complex and might have diverse effects necessitating the use of several primary outcome measures (30). For example, if a specific therapy for unilateral neglect in stroke is being tested, then besides using a measure assessing neglect (such as the Behavioural Inattention Test), the effects of the treatment on daily functioning should also be evaluated by using activity and/or participation scales (31). In some instances, in order to capture the expected specific effect, the assessment should focus on the specific outcomes defined, and the use of more generic measures encompassing many domains should be avoided (30). For example, when investigating the efficacy of electrical stimulation for upper limb spasticity, measures evaluating upper limb function would be preferred to a global activity measure, such as the FIM™ (32). In population studies where causes, features or consequences of disability, or the needs of target populations are sought, the assessment methods or tools should be relevant and applicable for the situation. Generic outcome measures are usually suitable for case-mix groups, and simple assessment methods are preferred (33, 34).

Policy-making

Results of both clinical audits, and clinical and epidemiological research, guide policy and decision-makers in both policy-making and planning healthcare services. In addition, they may seek methods that guide them in optimal allocation of limited resources. This necessitates an economic evaluation, such as cost-effectiveness and cost-benefit analyses, wherein statistics are used to calculate the monetary cost or benefit per gained

unit of outcome (35, 36). Cost-effectiveness and cost-benefit analyses have been used primarily for specific clinical outcomes, such as change in global health and functional status (35, 37). For example FIM-efficiency (FIM Gain/Length of stay) is used as a surrogate marker for cost-efficiency and, similarly, the Northwick Park Dependency Scale and Care Needs Assessment was developed to provide a cost-related outcome measure for rehabilitation (38). Utility measures, such as the EuroQoL or the quality-adjusted life year (QALY), are commonly used in economic evaluation in healthcare (39); they are preference-based measures specifically designed to assess the value or desirability of a particular health status or outcome.

WHERE, AND IN WHAT CONTEXT WILL THE ASSESSMENT BE USED?

The setting

Rehabilitation occurs in many settings, which can be broadly categorized as acute, post-acute or long-term, including the community. In the acute setting, the measurement focus is usually on impairment, whereas in the post-acute rehabilitation setting, both impairment and activity measures are used (40, 41). In the long-term setting or community phase, participation and QoL measures may be more relevant (42). For example in traumatic spinal cord injury, the assessment in the acute setting is focused mainly on impairments of motor and sensory function, evaluated by ASIA standards, as well as impairments of cardiovascular, respiratory and skin functions that appear as the common complications of the acute stage (43). In the post-acute rehabilitation setting, besides impairment assessment, a variety of activity measures are used, such as the FIM™, Barthel Index, Quadriplegia Index of Function (QIF), Spinal Cord Independence Measure (SCIM) or the Walking Index for Spinal Cord Injury (WISCI) (44). Finally, in the long-term, participation measures, such as the Craig Handicap Assessment and Reporting Technique (CHART), and QoL instruments, such as the Satisfaction with Life Scale, may be preferred (S36, S37).

Single or multiple diagnoses

For a single diagnosis, disease-specific measures may be appropriate, whereas for multiple diagnoses, generic measures are more appropriate. The primary difference between a disease-specific and generic instrument is that, in the former case, it will have been validated for a given specific diagnosis, while, in the latter, it can be applied across different diagnoses (although there should be evidence to this effect; see below). The use of generic measures has several advantages, including the reduced need for developing and testing different instruments for all patient groups separately, and uniformity of measurement in rehabilitation facilities. Furthermore, they allow comparison of the burden of diseases or disabilities among patient groups and in some cases with healthy populations (45). Disease-specific measures are confined to the problems of the relevant patient groups and expected be more sensitive to change (S38).

Even in a single diagnostic group, the choice of the instrument can vary depending on the purpose of assessment. If a detailed assessment of a domain is required, a relevant focused measure targeted at the specific domain may be necessary. On the other hand, if the aim is to measure the outcome of a rehabilitation programme, more practical, generic measures may be preferred.

One or more countries

For assessments including more than one country, cross-cultural validity of the chosen measures should be established for the relevant countries (S39, S40) (see below). International clinical trials bring additional requirements including, where appropriate, the training and consequent inter-rater reliability of those making assessments (where they are not self-completed).

ARE THERE GUIDELINES OR RECOMMENDATIONS PUBLISHED FOR WHAT IS NEEDED?

When one has to make an assessment in a certain situation, the critical question is "what should be measured and how?" Although the ICF works as a universal framework to answer the question "what should be measured", it can be difficult to select the domains or categories for a specific situation. In order to improve and standardize the assessment and outcome measurement, some international organizations, special interest or working groups have developed recommendations or guidelines. Two examples are given below.

OMERACT recommendations

Outcome Measures for Rheumatoid Arthritis in Clinical Trials (OMERACT) is an international, informally organized network initiated in 1992 aimed at improving outcome measurement in rheumatology (S41). Data-driven recommendations are prepared and updated by expert working groups. Recommendations include core sets of measures for most rheumatological conditions, such as rheumatoid arthritis, osteoarthritis, osteoporosis, ankylosing spondylitis, systemic lupus erythematosus, psoriatic arthritis, gout and fibromyalgia (46, S42, S43, S44). For example, OMERACT, with the endorsement of WHO and International League Against Rheumatism, suggested a preliminary core set for use in rheumatoid arthritis clinical trials. This core set includes the following measures: pain, patient global assessment, physical disability, swollen joints, tender joints, acute phase reactants, and physician global assessment; in studies of 1 or more years' duration, radiographs of joints should be performed (47). Besides the recommendation of what to measure, how it should be done is also reported. For instance, the Bath Ankylosing Spondylitis Functional Index or the Dougados Functional Index are suggested for measuring function in ankylosing spondylitis (48). Details of these recommendations can be accessed at their website, www.omeract.org.

ICF Core Sets

The ICF classification, which serves as a framework and a common language to address the impact of a health condition on human functioning, comprises 1,545 categories divided over the 4

ICF components (body functions, body structures, activities and participation, environmental factors) (4). In order to make this comprehensive classification applicable in healthcare, ICF Core Sets have been developed for specific diseases or conditions (49). ICF Core Sets are selections of ICF categories relevant for specific diseases or conditions, which can be used in clinical studies or health statistics (brief ICF core sets) or to guide multidisciplinary assessments (comprehensive ICF core sets). For clinical practice and research, they list the ICF categories that should be measured, but they provide no information about how to measure them. ICF Core Sets of chronic conditions that may be relevant for the field of physical and rehabilitation medicine include chronic ischaemic heart disease, diabetes mellitus, obesity, obstructive pulmonary diseases, breast cancer, depression, osteoarthritis, osteoporosis, low back pain, rheumatoid arthritis, chronic widespread pain, ankylosing spondylitis, stroke, multiple sclerosis and spinal cord injury (49, 50). Validity testing of some, but not all, of these core sets has been reported (S45). In addition, ICF Core Sets for post-acute (early) rehabilitation setting have also been developed and validated (S46, S47). Nevertheless, these remain taxonomies that are thought to be relevant to a specific condition and, as yet, do not constitute measurement. It has been suggested that the assignment of existing standardized instruments to ICF categories and the operationalization of the ICF qualifiers can contribute to further improvements of ICF-based rehabilitation management in the future (S48). An ICF qualifier scale has been proposed to evaluate the extent of a patient's problem in each of the ICF categories (4). The qualifier scale of the components body functions and structures and activities and participation have 5 response levels, ranging from 0 to 4: no/mild/moderate/severe/complete problem. The qualifier scale of the component environmental factors has 9 response levels, ranging from −4 to +4. A specific environmental factor can be a barrier (−1 to −4), or a facilitator (1 to 4), or can have no influence (0) on a patient's life. If a factor has an influence, the extent of the influence (either positive or negative) can be coded as mild, moderate, severe, or complete. At present there is some concern over these qualifiers, as the reliability of the ICF codes when measured with the current ICF qualifiers has been reported to be relatively low (51). The scoring of ICF qualifiers has also been found to be difficult and time-consuming by the raters (52), and a reduction in the number of response levels has been proposed in order to improve the reliability of the assessments performed by the ICF qualifier scale (53). Recent work has tried to operationalize the qualifiers by application of the Rasch measurement model (54).

WHAT ASSESSMENT TOOLS ARE AVAILABLE?

Selection of assessment tools

Existing assessment tools and outcome measures can be identified through literature review in electronic bibliographic databases, web search, dedicated instrument databases, books and manual searching through specific journals (with iterative retrieval of articles listed in their references). There are a number of sources where appropriate assessment tools and outcome measures may be summarized and, in some cases, evaluated, as follows:

Systematic reviews. A systematic review is a summary of research that uses explicit methods to perform a thorough literature search and critical appraisal of individual studies to identify the evidence relevant to a specific question. It often, but not always, uses statistical techniques (meta-analysis) to combine these valid studies, or at least uses grading of the levels of evidence depending on the methodology used. Systematic reviews are crucial to evidence-based medicine. They are often based upon web searches of established databases, such as MEDLINE and PubMed. They may give detailed explanation of, for example, the range of balance measures available to the rehabilitation process (20) or functional outcome measures for the hemiparetic upper limb (32). A formalized library of such reviews can be found in the Cochrane Library, although most of these in the rehabilitation domain will concern practice, rather than available outcome measures (S49).

Other summary studies.There have been several published studies that have identified the range of assessments used across different countries (11, 12, 55, 56). These typically report on frequency of use rather than on explanation of what is being measured or any quality criteria of the scales themselves.

Single studies. Many assessment scales in use have been originally published as a paper reporting on the reliability and/or validity of the scale (S4, S8, S10, S17, S24, S35). Thus, a MEDLINE search for measuring a specific construct (perhaps also accompanied by an abstract and/or title word of "reliability") may elicit several relevant assessments.

Books. There is a wide range of books that review available assessments and outcome measures (57–61). These may have more information about the quality of those assessments, reporting various psychometric criteria (see below).

Assessment tools for stroke and rheumatoid arthritis

Examples of assessment tools for two different diagnostic groups, stroke and rheumatoid arthritis are presented in Tables I and II. Assessment domains and sub-domains that might be relevant for each diagnostic group and available assessment tools for the corresponding domain or sub-domain are listed. It has been shown that many of the listed scales for stroke can be mapped onto the ICF classification, often being associated with the activities and participation component, with mobility being the category most frequently covered in a wide range of the instruments (62). For rheumatoid arthritis, one study reported that a comparison of instruments showed that the different health status measures covered different components, and that they covered the different components with different levels of precision (63).

HOW SHOULD THE QUALITY OF THE ASSESSMENT TOOLS BE JUDGED?

Assuming that a choice is made of the instrument, or instruments that seem appropriate for the context and topic to be measured, the next task is to review the quality of the selected

Table I. *Examples of assessment tools for stroke*

Assessment domain	Assessment tool
Body functions	
Consciousness	Glasgow Coma Scale (S50)
Global cognitive functions	Mini-Mental State Examination (S3), Neurobehavioral Cognitive Status Examination (S51)
Memory functions	Rivermead Behavioural Memory Test (S52)
Attention functions	Behavioural Inattention Test (S53)
Visual perception functions	Motor-free Visual Perception Test (S54)
Speech and language functions	Boston Diagnostic Aphasia Examination (S28)
Emotional functions	Beck Depression Inventory (S30), Hospital Anxiety and Depression Scale (S55)
Motor functions	Fugl-Meyer Assessment (S6), Brunnstrom's stages of motor recovery (S56), Modified Ashworth Scale (S4)
Composite neurological functions	National Institutes of Health Stroke Scale (S57), Canadian Neurological Scale (S58)
Activities and participation	
Activities of daily living	Barthel Index (S12), Functional Independence Measure (S14)
Instrumental activities of daily living	Frenchay Activities Index (S31)
Mobility	Berg Balance Scale (S35), Rivermead Mobility Index (S10), Timed Up and Go Test (S59)
Dexterity	Nine-Hole Peg Test (S11)
Activities and participation	London Handicap Scale (S15), WHODAS II (S60), Impact on Participation and Autonomy Questionnaire (S17)
QoL/Health-related QoL	
	SF-36 (S21), NHP (S22), EuroQoL (S23), Stroke Impact Scale (S61), Stroke-Specific Quality of Life Scale (S62), Stroke-Adapted Sickness Impact Profile (S63)

QoL: Quality of life; WHODAS II: World Health Organization Disability Assessment Schedule II; SF-36: Medical Outcomes Study Short Form 36; NHP: Nottingham Health Profile.

instrument(s). There are a range of quality "standards" that must be met (64–66), as well as gaining a clear understanding of the type of data that will be derived from the instruments, as this will affect the type of analysis available.

What type of information will be obtained from this assessment?

The type of information obtained from assessments generally falls into two groups; those that categorize patients into one

Table II. *Examples of assessment tools for rheumatoid arthritis*

Assessment domain	Assessment tool
Body functions	
Sensation of pain	Visual analogue scale (S1), verbal rating scale (S1), Multidimensional Pain Inventory (S2), AIMS2-Pain section (S64), NHP-Pain section (S22), SF36-Pain section (S21), Rheumatoid Arthritis Pain Scale (S65)
Sensation of muscle stiffness	Duration of morning stiffness
Sensitivity to pressure	Tender joint count
Mobility of joint functions	Joint range of motion
Muscle power functions	Grip strength
Haematological system functions	Erythrocyte sedimentation rate, C-reactive protein
Energy and drive functions	Multidimensional Assessment of Fatigue Scale (S66), visual analogue scale
Sleep functions	Medical Outcomes Study, Sleep measure (S67)
Emotional functions	Hospital Anxiety Depression Scale (S55), Beck Depression Inventory (S30)
Body Structures	
Structures related to movement	Swollen joint count Joint damage: Larsen Index (S7), Sharp Index (S68)
Body functions/Body structures	
Composite	DAS28 (S69)
Activities & Participation	
Activities	Health Assessment Questionnaire (S13), AIMS2 – mobility (S64), AIMS2 – walking & bending (S64), AIMS2 – hand & finger function (S64), AIMS2 – arm function (S64), AIMS2 – self care (S64), AIMS2 – household tasks (S64)
Participation	AIMS2 – social activity (S64), AIMS2 – support (S64), AIMS2 – work (S64), Rheumatoid Arthritis Work Instability Scale (S70)
Activities and participation	London Handicap Scale (S15), WHODAS II (S60)
QoL/Health-related QoL	
	SF-36 (S21), NHP (S22), EuroQoL (S23), Rheumatoid Arthritis Quality of Life Scale (S71)

AIMS2: Arthritis Impact Measurement Scales 2; SF-36: Medical Outcomes Study Short Form 36; NHP: Nottingham Health Profile; DAS28: disease activity score 28; WHODAS II: World Health Organization Disability Assessment Schedule II.

or more groups, and those that provide an estimate of the level of the construct under consideration. Sometimes an assessment instrument may do both, that is it can be used both to categorize patients, and to give an estimate of the level of the construct. The majority of assessments used in rehabilitation fall into the latter category, and the type of measurement of the estimate may differ. The data produced by the assessment is reported as one of the 4 levels of Stevens' classification, as explained below (S72):

Categorical. Categorical or nominal data are more often used to classify people into groups, for example by gender or ethnicity. The key attribute of a categorical variable is that there is no implied ordering between categories. However, sometimes ordinal scales (see below) are used to classify patients into clinical groups, for example those with or without clinical depression.

Ordinal. Many assessment tools used in everyday clinical practice are ordinal scales, which are based upon a score derived from a set of tasks or questions. In this scale type, the numbers assigned to tasks or questions represent the rank order. Ordinal scales order people by the magnitude of the construct under consideration, for example activities of daily living as in the Barthel Index (S12) and Health Assessment Questionnaire (S13), or pain as in the Multidimensional Pain Inventory (S2). A key characteristic of these scales is that the distances between the raw score points are unequal and mathematical calculations, such as change scores, are invalid (67). In general, different sets of statistical procedures (non-parametric) are available for use with ordinal scales (68). However, recently methods to transform these ordinal scales into interval level measures have gained widespread use, opening up a wider range of available statistical procedures (69, 70).

Interval. An interval scale has sequential units with numerically equal distances between them. An example of interval scale measurement is the Celsius temperature scale, where the unit of measurement is 1/100 of the difference between the melting temperature and the boiling temperature of water at atmospheric pressure. The scale has an arbitrary "zero point", negative values can be used, and 40ºC is not twice as warm as 20ºC. Most of the ordinal scales that have been Rasch transformed will have interval scaling with an arbitrary zero (see below).

Ratio. A ratio scale is an interval scale that has a zero point representing the total absence of the quantity being measured. Measurement of range of motion with a goniometer provides data at the ratio scale level, where there is a meaningful zero and as a consequence it is possible to say that 20º is twice the range of motion as 10º. Weight and height measures are similar in the type of measurement. It is probably fair to say that most of the measurement in physical and rehabilitation medicine at this level are concerned with body functions and structures.

What are the basic standards?

Given an awareness of the different type of measurement available from assessments, irrespective of this, certain standards are required of any instrument. These have historically been catalogued under the rubric of psychometrics (71). They are primarily concerned with reliability and validity. Reliability is concerned with whether the instrument delivers an estimate of the person's level of the trait being measured in a consistent manner. It assesses the extent to which the instrument is free from random error. Validity addresses whether the instrument measures what it is intended to measure (72). Validity has traditionally been separated into 3 distinct types; namely, content, criterion and construct validity. However, contemporary thinking suggests that these distinctions are arbitrary (73). According to the modern psychometric approach, there are 4 main stages to the construction and testing of an assessment instrument: content validity, internal construct validity, reliability and external validity.

Content validity. Content evidence remains a critical first step in establishing the validity of an assessment scale in both traditional and contemporary approaches (74). Content validity is the extent to which an instrument contains items critical or appropriate to the construct being measured. It is concerned with whether the items adequately cover the expected substantive content of the instrument. Content validity is established by a systematic, qualitative approach including focus groups, expert panels, etc. (64). Face validity, which might seem to have a superficial resemblance to content validity, is a judgement of the appearance of the instrument indicating that the item set looks appropriate (73).

Internal construct validity. Irrespective of how the items for a scale have been determined, there must be evidence that they can be summated to give a score, that is they represent a unidimensional construct (75). In classical test theory, this evidence is most likely to come from a test of "factorial validity". Factorial-, structural- or internal construct validity is the degree to which the measure of a construct conforms to the theoretical definition of the construct (76). Most recently factorial validity has been established by confirmatory factor analysis (77), to identify one or more unidimensional constructs.

Another approach to investigate the dimensional structure and scalability of the assessment scales is Rasch analysis. The Rasch measurement model is based on item response theory and transforms the ordinal scales into interval measure. This requires that the data from the questionnaire satisfies the expectation of the Rasch measurement model. The Rasch model (69) is the current standard for the development of unidimensional scales delivering metric quality outcomes in healthcare (70, 78) and has in recent years been used frequently in rehabilitation research (79, 80). Briefly, data collected from questionnaires (or for assessments completed by clinical staff) which include items for a new (or existing) scale, which are intended to be summated into an overall score (which may be at the subscale, or overall level) are tested against the expectations of this measurement model. The model defines how responses to items should be if measurement (at the metric level) is to be achieved. For the Rasch model, dichotomous (69) and polytomous versions are available (81, 82). The response patterns

achieved are tested against what is expected (a probabilistic form of Guttman scaling) (83), and a variety of fit statistics determine if this is the case (84).

In addition to determining if the data conform to these expectations, the process of Rasch analysis involves a number of other actions, which are important to the quality of the measurement obtained. As well as testing some of the assumptions of the model, such as unidimensionality and local independence of items, aspects such as measurement invariance are routinely tested. Within the framework of Rasch measurement, the scale should work in the same way, irrespective of which group (based on age, gender or for example disease subtype) is being assessed (85). For example, in the case of measuring pain, younger and older age groups should have the same probability of affirming an item if they have the same level of pain. If for some reason one group did not display the same probability of affirming the item (in the dichotomous case), then this item would be deemed to display differential item functioning (DIF), and would also violate the requirement of unidimensionality (86). Consequently, every item should be checked for DIF by age and gender and for any clinical or other subgroup relevant for the analysis.

There are several Rasch software packages available for this purpose, but the most widely used in health outcomes are WINSTEPS (S73) and RUMM (S74).

Reliability. Once it is shown that a set of items can be summated to give a score, the reliability of that score can be assessed. There are different types of reliability and, in rehabilitation the most commonly reported reliabilities are internal consistency reliability, test-retest and inter-rater reliability. Internal consistency reliability is reported as a Cronbach's alpha statistic, and is concerned with how a set of items or tasks, whose scores are intended to be summated together, have acceptable inter-item correlations (87). Different levels of reliability are required for different use. Where the object is to compare groups, as in many research situations, the reliability can be much lower than when an instrument is used to aid individual clinical decisions. Generally, for the former, an alpha of 0.7 or greater is required, and for the latter, a value of 0.9 or greater is required, although often the value 0.85 is used (71, 88).

Test-retest reliability is an indication that the instrument remains stable over repeated administrations. It is undertaken on people who are not expected to change, and administered twice, 2–4 weeks apart. Adequate test-retest reliability should be considered a requirement for any instrument used to measure outcome, when estimates will be derived from two or more occasions. Inter-rater reliability is concerned with whether or not raters agree in their assessment of an individual.

External validity. Given evidence of content and internal construct validity, and reliability, it is then possible to compare the instrument with other "external" scales. There are several types of external validity that address this issue, all of which, in some way or other, will see if the scale score has an expected association with other "external" comparators. Consequently these types are collectively referred to as external validity (89). External validity includes "criterion" and "external construct" validity. Criterion validity is where the scale score is compared to a criterion variable, or "gold standard". Concurrent validity and predictive validity are forms of criterion validity where, in the former, the scale score is compared with other instruments intended to measure the same or similar construct and, in the latter, the scale score is predictive of some expected event or outcome. Convergent validity and discriminative validity are forms of external construct validity. Convergent validity addresses that the scale has expected correlation with another measure that is theoretically predicted to correlate with, whereas discriminant validity shows that the scale does not correlate with dissimilar or unrelated measures (90).

Additional requirements for when the assessment is going to be used as an outcome measure

Floor and ceiling effect. It would be inappropriate to give a set of questions intended to measure higher functional activities to patients admitted to the rehabilitation ward. This would result in what is called a "floor effect"; everyone would score zero. Likewise, there would be little point in giving a scale focussing on high dependency needs to ambulant outpatients, as they are likely to score the maximum on such a scale; a "ceiling effect". Floor or ceiling effects are considered to be present if more than 15% of respondents achieved the lowest or highest possible score, respectively. The consequence of a floor effect is that deterioration may be missed and, for a ceiling effect, improvement may be missed (90).

Responsiveness. Responsiveness or sensitivity to change is the ability of an instrument to detect change over time. There are two approaches to the assessment of responsiveness (91). The first approach is distribution-based, the most common being the "effect size" (92). This allows a comparison between different instruments, where the assessment is taken at the same time (thus removing any confounding brought about by the different efficacy of a particular intervention). It is based on the change score divided by the standard deviation at either time-point (but usually baseline), and is thus presenting the magnitude of change in standard deviation units. Variations on this theme can be found, for example, with the standardized response mean (SRM), which divides the mean change by the standard deviation of the change scores, or with the smallest real difference (93). One fundamental problem with this approach is that such calculations assume interval-scaled normally distributed variables. In practice these requirements are largely ignored, even though recent evidence has suggested that effect sizes can be inflated by inappropriately applying ordinal data (94). Non-parametric effect sizes are available (95).

The second approach is anchor-based; that is, there is some external value that determines that a significant change has occurred, the most common being the minimal clinically important difference (MCID) (96). This attempts to identify a meaningful magnitude of change by comparing the change score on an assessment with some summary perceived measure of change. MCID reflects the interpretability of the scale scores, which can be defined as the degree to which one can assign meaning to quantitative scores (45). MCID shares the methodological weakness of the effect size in ignoring the

ordinal nature of both the scale and the global measure used for comparison. The presence of a floor or ceiling effect is likely to understate the level of responsiveness using both approaches.

More recently, it has been proposed that change based upon Rasch analysis of individual patient's scores may give a much more sensitive measurement of change than the traditional effect size approach (97).

Additional factors that should be taken into consideration

In addition to satisfying the basic psychometric requirements of reliability, validity and responsiveness, there are a number of other considerations that must be taken into account when it comes to the choice of assessment.

Feasibility. Feasibility addresses the respondent and administrative burden of the assessment tool. Some assessments take a long time to complete and may be quite inappropriate, for example, in a busy outpatient clinic. Furthermore, some assessments can only be undertaken by staff with certain skills, or belonging to a particular profession. Thus it is important to ascertain the operational requirements for any assessment, and make a judgment based upon the feasibility of the assessment in the context of the intended use. However, new approaches based upon item response theory, including Rasch analysis, and computer adaptive testing, which tailors tests to the patients' level of the attribute being measured, have the capacity to reduce the respondent burden and the time needed to complete the test, and these approaches may transform the operational context of measurement in the future (98–100).

Cost. Not all assessments are free for use within public health or not-for-profit settings. A wide variety of license agreements are used and some can be very expensive, for example €2 or more per assessment. Consequently it is essential to determine the licensing conditions for any assessment. However, there are many instruments that can be used without licensing. Sometimes there are other costs; for example, in the requirement for training to administer an assessment or, in some circumstances, the requirement to deposit data elsewhere as a condition of use. These requirements may be ongoing, and thus there will be a commitment to long-term expenditure if the use of the assessment becomes permanent.

Language adaptation. Sometimes an assessment is not available in the language required. Adapting such an assessment, be it for use by professionals, or a self-completed questionnaire to be used by patients, requires both scientific rigour, and can be expensive. There are a series of guidelines to help with such a task, especially for patient-reported outcomes (S39, S40). Producing a good translation and cultural adaptation of an outcome measure (particularly if patient-reported) requires checking the semantic, idiomatic, experiential and conceptual equivalence between source and final versions. This means analysing many times in different ways to ascertain if the instrument functions as required with "real" target people. A correct translation process (including cognitive debriefing interviews) is just the first step. Full adaptation requires that scaling and psychometric properties of the new language version are assessed and compared with those of the source version, also applying item response theory methods (e.g. DIF techniques) (S40). If multi-language versions of an assessment are required for a particular study, it is essential that they have been properly adapted and that there is published evidence that they are reliable, valid and free of cross-cultural DIF (85).

CONCLUSION

The choice of an appropriate assessment or outcome instrument is an important aspect of clinical practice, audit and research. Considerable care must be taken to ensure that the best possible assessments are chosen for the task in hand, and that, wherever possible, they conform to all modern quality standards for measurement.

REFERENCES[1]

1. Stucki G, Cieza A, Melvin J. The International Classification of Functioning, Disability and Health (ICF): a unifying model for the conceptual description of the rehabilitation strategy. J Rehabil Med 2007; 39: 279–285.
2. Tesio L. Functional assessment in rehabilitative medicine: principles and methods. Eura Medicophys 2007; 43: 515–523.
3. Wade DT. Outcome measurement and rehabilitation. Clin Rehabil 1999; 13: 93–95.
4. World Health Organization. International Classification of Functioning, Disability and Health (ICF). Geneva: World Health Organization; 2001.
5. Deiner E. Subjective well-being. Psychol Bull 1984; 95: 542–575.
6. Fuhrer MJ. Subjectifying quality of life as a medical rehabilitation outcome. Disabil Rehabil 2000; 22: 481–489.
7. McKenna SP, Doward LC. The needs-based approach to quality of life assessment. Value Health 2004; 7 Suppl 1: S1–S3.
8. Parker K, Kirby RL, Adderson J, Thompson K. Ambulation of people with lower-limb amputations: relationship between capacity and performance measures. Arch Phys Med Rehabil 2010; 91: 543–549.
9. Holsbeeke L, Ketelaar M, Schoemaker MM, Gorter JW. Capacity, capability, and performance: different constructs or three of a kind? Arch Phys Med Rehabil 2009; 90: 849–855.
10. van den Berg-Emons RJ, Bussmann JB, Stam HJ. Accelerometry-based activity spectrum in persons with chronic physical conditions. Arch Phys Med Rehabil 2010; 91: 1856–1861.
11. Haigh R, Tennant A, Biering-Sørensen F, Grimby G, Marinček C, Phillips S, et al. The use of outcome measures in Physical Medicine and Rehabilitation within Europe. J Rehabil Med 2001: 33: 273–278.
12. Skinner A, Turner-Stokes L. The use of standardized outcome measures in rehabilitation centres in the U.K. Clin Rehabil 2006; 20: 609–615.
13. Noonan VK, Kopec JA, Noreau L, Singer J, Chan A, Masse LC, et al. Comparing the content of participation instruments using the International Classification of Functioning, Disability and Health. Health Qual Life Outcomes 2009; 7: 93.
14. Magasi S, Post MW. A comparative review of contemporary participation measures' psychometric properties and content cover-

[1]Additional references (S1-74) will be found at https://doi.org/10.2340/16501977-0844.

age. Arch Phys Med Rehabil 2010; 91 Suppl 1: S17–S28.
15. Fugl-Meyer AR, Melin R, Fugl-Meyer KS. J Life satisfaction in 18- to 64-year-old Swedes: in relation to gender, age, partner and immigrant status. J Rehabil Med 2002; 34: 239–246.
16. Johnston MV, Miklos CS. Activity-related quality of life in rehabilitation and traumatic brain injury. Arch Phys Med Rehabil 2002; 83 Suppl 2: S26–S38.
17. Garratt A, Schmidt L, Mackintosh A, Fitzpatrick R. Quality of life measurement: bibliographic study of patient assessed health outcome measures. BMJ 2002; 324: 1417–1421.
18. Wade DT. Describing rehabilitation interventions. Clin Rehabil 2005; 19: 811–818.
19. Tyson S, Greenhalgh J, Long AF, Flynn R. The use of measurement tools in clinical practice: an observational study of neurorehabilitation. Clin Rehabil 2010; 24: 74–81.
20. Tyson S, Connell LA. How to measure balance in clinical practice. A systematic review of the psychometrics and clinical utility of measures of balance activity for neurological conditions. Clin Rehabil 2009; 23: 824–840.
21. Barak S, Duncan PW. Issues in selecting outcome measures to assess functional recovery after stroke. NeuroRx 2006; 3: 505–524.
22. National Health Service Review. Working Paper No 6 Medical Audit. London: HMSO; 1989.
23. National Institute of Clinical Excellence (NICE). Principles of best practice in clinical audit. Abingdon, UK: Radcliffe Medical Press; 2002.
24. Wikipedia.org [homepage on internet] San Francisco: Wikimedia Foundation Inc.; 2001 [updated 2010 June 8]. Clinical audit. Available from: http://en.wikipedia.org/wiki/Clinical_audit.
25. Granger CV, Black T, Braun SL. Quality and outcome measures for medical rehabilitation. In: Braddom RL, editor. Physical medicine and rehabilitation. China: Elsevier Inc.; 2007, p. 151–164.
26. Freeman JA, Hobart JC, Playford ED, Undy B, Thompson AJ. Evaluating neurorehabilitation: lessons from routine data collection. J Neurol Neurosurg Psychiatry 2005; 76: 723–728.
27. Cadilhac DA, Pearce DC, Levi CR, Donnan GA; Greater Metropolitan Clinical Taskforce and New South Wales Stroke Services Coordinating Committee. Improvements in the quality of care and health outcomes with new stroke care units following implementation of a clinician-led, health system redesign programme in New South Wales, Australia. Qual Saf Health Care 2008; 17: 329–333.
28. Abilleira S, Gallofré M, Ribera A, Sánchez E, Tresserras R. Quality of in-hospital stroke care according to evidence-based performance measures: results from the first audit of stroke, Catalonia, Spain. Stroke 2009; 40: 1433–1438.
29. Schouten LM, Hulscher ME, Akkermans R, van Everdingen JJ, Grol RP, Huijsman R. Factors that influence the stroke care team's effectiveness in reducing the length of hospital stay. Stroke 2008; 39: 2515–2521.
30. Wade DT, Smeets RJ, Verbunt JA. Research in rehabilitation medicine: methodological challenges. J Clin Epidemiol 2010; 63: 699–704.
31. Bowen A, Lincoln NB. Cognitive rehabilitation for spatial neglect following stroke. Cochrane Database Syst Rev 2007; 18: CD003586.
32. Ashford S, Slade M, Malaprade F, Turner-Stokes L. Evaluation of functional outcome measures for the hemiparetic upper limb: a systematic review. J Rehabil Med 2008; 40: 787–795.
33. Cigolle CT, Langa KM, Kabeto MU, Tian Z, Blaum CS. Geriatric conditions and disability: the Health and Retirement Study. Ann Intern Med 2007; 147: 156–164.
34. Holbrook TL, Anderson JP, Sieber WJ, Browner D, Hoyt DB. Outcome after major trauma: 12-month and 18-month follow-up results from the Trauma Recovery Project. J Trauma 1999; 46: 765–771.
35. Uhlig T, Finset A, Kvien TK. Effectiveness and cost-effectiveness of comprehensive rehabilitation programs. Curr Opin Rheumatol 2003, 15: 134–140.
36. Cimera RE, Rumrill PD. Economic analyses of rehabilitation services as research methodologies. Work 2008; 31: 483–487.
37. Brady BK, McGahan L, Skidmore B. Systematic review of economic evidence on stroke rehabilitation services. Int J Technol Assess Health Care 2005; 21: 15–21.
38. Turner-Stokes L. Politics, policy and payment – facilitators or barriers to person-centred rehabilitation? Disabil Rehabil 2007; 29: 1575–1582.
39. Räsänen P, Roine E, Sintonen H, Semberg-Konttinen V, Ryynänen OP, Roine R. Use of quality-adjusted life years for the estimation of effectiveness of health care: A systematic literature review. Int J Technol Assess Health Care 2006; 22: 235–241.
40. Scheuringer M, Stucki G, Huber EO, Brach M, Schwarzkopf S, Kostanjsek N, et al. ICF Core Set for patients with musculoskeletal conditions in early post-acute rehabilitation facilities. Disabil Rehabil 2005; 27: 405–410.
41. Stier-Jarmer M, Grill E, Ewert T, Bartholomeyczik S, Finger M, Mokrusch T, et al. ICF Core Set for patients with neurological conditions in early post-acute rehabilitation facilities. Disabil Rehabil 2005; 27: 389–395.
42. Bullinger M, Azouvi P, Brooks N, Basso A, Christensen AL, Gobiet W et al; TBI Consensus Group. Quality of life in patients with traumatic brain injury-basic issues, assessment and recommendations. Restor Neurol Neurosci 2002; 20: 111–124.
43. Furlan JC, Noonan V, Singh A, Fehlings MG. Assessment of impairment in patients with acute traumatic spinal cord injury: a systematic review of the literature. J Neurotrauma 2010 Apr 6. [Epub ahead of print].
44. Dawson J, Shamley D, Jamous MA. A structured review of outcome measures used for the assessment of rehabilitation interventions for spinal cord injury. Spinal Cord 2008; 46: 768–780.
45. Dekker J, Dallmeijer AJ, Lankhorst GJ. Clinimetrics in rehabilitation medicine: current issues in developing and applying measurement instruments. J Rehabil Med 2005; 37: 193–201.
46. Brooks P, Hochberg M. Outcome measures and classification criteria for the rheumatic diseases. A compilation of data from OMERACT (Outcome Measures for Arthritis Clinical Trials), ILAR (International League of Associations for Rheumatology), regional leagues and other groups. Rheumatology (Oxford) 2001; 40: 896–906.
47. Boers M, Tugwell P, Felson DT, van Riel PL, Kirwan JR, Edmonds JP, et al. World Health Organization and International League of Associations for Rheumatology core endpoints for symptom modifying antirheumatic drugs in rheumatoid arthritis clinical trials. J Rheumatol 1994; 21 Suppl 41: S86–S89.
48. van der Heijde D, Bellamy N, Calin A, Dougados M, Khan MA, van der Linden S. Preliminary core sets for endpoints in ankylosing spondylitis. J Rheumatol 1997; 24: 2225–2229.
49. Stucki G, Grimby G. Applying the ICF in medicine. J Rehabil Med 2004; Suppl 44: 5–6.
50. Stucki G, Kostanjsek N, Ustun TB, Cieza A. ICF-based classification and measurement of functioning. Eur J Phys Rehabil Med 2008; 44: 315–328.
51. Okochi J, Utsunomiya S, Takahashi T. Health measurement using the ICF: test-retest reliability study of ICF codes and qualifiers in geriatric care. Health Qual Life Outcomes 2005; 3: 46.
52. Bautz-Holter E, Sveen U, Cieza A, Geyh S, Røe C. Does the International Classification of Functioning, Disability and Health (ICF) core set for low back pain cover the patients' problems? A cross-sectional content-validity study with a Norwegian population. Eur J Phys Rehabil Med 2008; 44: 387–397.
53. Uhlig T, Lillemo S, Moe RH, Stamm T, Cieza A, Boonen A, et al. Reliability of the ICF Core Set for rheumatoid arthritis. Ann Rheum Dis 2007; 66: 1078–1084.
54. Cieza A, Hilfiker R, Boonen A, Chatterji S, Kostanjsek N, Ustün BT, Stucki G. Items from patient-oriented instruments can be integrated into interval scales to operationalize categories of the International Classification of Functioning, Disability and Health. J Clin Epidemiol 2009; 62: 912–921.
55. Torenbeek M, Caulfield B, Garrett M, Van Harten W. Current use of outcome measures for stroke and low back pain rehabilitation in five European countries: first results of the ACROSS project. Int J Rehabil Res 2001; 24: 95–101.

56. Douglas H, Swanson C, Gee T, Bellamy N. Outcome measurement in Australian rehabilitation environments. J Rehabil Med 2005; 3: 325–329.
57. Barat M, Franchignoni F, editors. Assessment in physical medicine and rehabilitation. Pavia: PI-ME – Maugeri Foundation Books; 2004.
58. Bowling A. Measuring health. A review of quality of life measurement scales, 3rd edition. Buckingham: Open University Press; 2004.
59. Herndon RN. The Handbook of Neurologic Rating Scales, 2nd edition. NewYork, NY: Demos Medical Publishing; 2006.
60. McDowell I. Measuring health. A guide to rating scales and questionnaires, 3rd edition. New York: Oxford University Press; 2006.
61. Wade DT. Measurement in neurological rehabilitation. Oxford: Oxford University Press; 1992.
62. Schepers VP, Ketelaar M, van de Port IG, Visser-Meily JM, Lindeman E. Comparing contents of functional outcome measures in stroke rehabilitation using the International Classification of Functioning, Disability and Health. Disabil Rehabil 2007; 29: 221–230.
63. Stucki G, Cieza A. The International Classification of Functioning, Disability and Health (ICF) Core Sets for rheumatoid arthritis: a way to specify functioning. Ann Rheum Dis 2004; 63 Suppl 2: ii40–ii45.
64. Frost MH, Reeve BB, Liepa AM, Stauffer JW, Hays RD. What is sufficient evidence for the reliability and validity of patient-reported outcome measures? Value Health 2007; 10 Suppl 2: S94–S105.
65. Rothman M, Burke L, Erickson P, Leidy NK, Patrick DL, Petrie CD. Use of existing patient-reported outcome (PRO) instruments and their modification: the ISPOR good research practices for evaluating and documenting content validity for the use of existing instruments and their modification PRO Task Force report. Value Health 2009; 12: 1075–1083.
66. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol D, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res 2010; 19: 539–549.
67. Merbitz C, Morris J, Grip JC. Ordinal scales and foundations of misinference. Arch Phys Med Rehabil 1989: 70: 308–312.
68. Svensson E. Guidelines to statistical evaluation of data from rating scales and questionnaires. J Rehabil Med 2001; 33: 47–48.
69. Rasch G. Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press; 1960.
70. Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Rheum 2007; 57: 1358–1362.
71. Nunnally JC, Bernstein IH. Psychometric theory, 3rd edition. New York: McGraw-Hill; 1994.
72. Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating patient-based outcome measures for use in clinical trials. Health Technol Assess 1998; 2: i–iv, 1–74.
73. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: Theory and application. Am J Med 2006; 119: 166.e7–166.e16.
74. Johnston MV, Graves DE. Towards guidelines for evaluation of measures: an introduction with application to spinal cord injury. J Spinal Cord Med 2008; 31: 13–26.
75. Thurstone LL. Attitudes can be measured. Am J Sociol 1928; 33: 529–554.
76. Loevinger J. Objective tests as instruments of psychological theory. Psychol Rep 1957; 3, 635–694.
77. Bollen KA. Structural equations with latent variables. New York: John Wiley & Sons; 1989.
78. Tennant A, McKenna SP, Hagell P. Application of Rasch analysis in the development and application of quality of life instruments. Value Health 2004; 7 Suppl 1: S22–S26.
79. Tesio L. Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. J Rehabil Med 2003; 35: 105–115.
80. Gallagher P, Franchignoni F, Giordano A, MacLachlan M. Trinity amputation and prosthesis experience scales: a psychometric assessment using classical test theory and Rasch analysis. Am J Phys Med Rehabil 2010; 89: 487–496.
81. Andrich D. Rating formulation for ordered response categories. Psychometrika 1978; 43: 561–573.
82. Masters G. A Rasch model for partial credit scoring. Psychometrika 1982; 47: 149–174.
83. Guttman L. The basis for scalogram analysis. In: Stouffer SA, Guttman L, Suchman EA, Lazarsfield PF, Star SA, Clausen JA, editors. Studies in social psychology in World War II: Vol 4. Measurement and prediction. Princeton: Princeton University Press; 1950, p. 60–90.
84. Smith RM. Fit analysis in latent trait measurement models. J Applied Measurement 2000; 2: 199–218.
85. Tennant A, Penta M, Tesio L, Grimby G, Thonnard J-L, Slade A, et al. Assessing and adjusting for cross cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model: the Pro-ESOR project. Medical Care 2004; 42 Suppl 1: 37–48.
86. Holland PW, Wainer H. Differential item functioning. Hilldale NJ: Lawrence Erlbaum Associates; 1993.
87. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951; 16: 297–334.
88. Bland JM, Altman DG. Cronbach's alpha. BMJ 1997; 314: 572.
89. Smith GT. On construct validity: Issues of method and measurement. Psychol Assess 2005; 17: 396–408.
90. Streiner DL, Norman GR. Health measurement scales. A practical guide to their development and use. 3rd edition. Oxford: Oxford University Press; 2003.
91. Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol 2003; 56: 395–407.
92. Cohen J. Statistical power analysis for the behavioral sciences, 2nd edition. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
93. Beckerman H, Roebroeck ME, Lankhorst GJ, Becher JG, Bezemer PD, Verbeek AL. Smallest real difference, a link between reproducibility and responsiveness. Qual Life Res 2001; 10: 571–578.
94. Kersten P, White PJ, Tennant A. The visual analogue WOMAC 3.0 scale – internal validity and responsiveness of the VAS version. BMC Musculoskeletal Disorders 2010; 11: 80.
95. Wolf FM. Meta analysis. Qualitative applications in the social sciences 59. Newbury Park California: Sage; 1986.
96. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials 1989; 10: 407–415.
97. Hobart J, Stefan C, Thompson AJ. Effect sizes can be misleading: is it time to change the way we measure change? J Neurol Neurosurg Psychiatry 2010; 81: 1044–1048.
98. Haley S, Ni P, Hambleton R, Slavin M, Jette A. Computer adaptive testing improves accuracy and precision of scores over random item selection in a physical functioning item bank. J Clin Epidemiol 2006; 59: 1174–1182.
99. Bjorner JB Chang C-H, Thissen D, Reeve BB. Developing tailored instruments: item banking and computerized adaptive assessment. Qual Life Res 2007; 16: 95–108.
100. Elhan AH, Oztuna D, Kutlay S, Küçükdeveci AA, Tennant A. An initial application of computerized adaptive testing (CAT) for measuring disability in patients with low back pain. BMC Musculoskelet Disord 2008; 9: 166.