Issues in Selecting Outcome Measures to Assess Functional Recovery After Stroke

Sharon Barak* and Pamela W. Duncan[†‡]

*Department of Physical Therapy, College of Public Health and Health Professions, and †Department of Aging and Geriatric Research, College of Medicine, University of Florida, Gainesville, Florida; and ‡Rehabilitation Outcomes Research Center, Department of Veterans Affairs, Gainesville, Florida

Summary: Most patients who survive a stroke experience some degree of physical recovery. Selecting the appropriate outcome measure to assess physical recovery is a difficult task, given the heterogeneity of stroke etiology, symptoms, severity, and even recovery itself. Despite these complexities, a number of strategies can facilitate the selection of functional outcome measures in stroke clinical trial research and practice. Clinical relevance in stroke outcome measures can be optimized by incorporating a framework of health and disability, such as the International Classification of Functioning, Disability, and Health (ICF). The ICF provides the conceptual basis for measurement and policy formulations for disability and health assessment. All outcome measures selected should also have sound psychometric properties. The essential psychometric properties are reliability, validity, responsiveness, sensibility, and established minimal clinically important difference. It is also important to establish the purpose of the measurement (discriminative, predictive, or evaluative) and to determine whether the purpose of the study is to evaluate the efficacy or effectiveness of an intervention. In addition, when selecting outcome measures and time of assessment, the natural history of stroke and stroke severity must be regarded. Finally, methods for acquiring data must also be considered. We present a comprehensive overview of the issues in selecting stroke outcome measures and characterize existing measures relative to these issues. **Key Words:** Disability evaluation, outcome assessment, measurement, stroke, cerebrovascular accident, recovery.

INTRODUCTION

Stroke is a leading cause of disability in the United States.[1] Regardless of the initial severity of the disability and neurological deficit, most stroke survivors exhibit some degree of recovery over time.[2–5] Assessment of recovery in individuals after stroke is important for both clinical practice and research,[6] but selecting outcome measures is a difficult process. Outcome measurement in stroke is difficult due to the various etiologies of stroke, heterogeneity of symptoms, variability in severity, and the possibility of spontaneous recovery after stroke.[7] Despite such complexities, several strategies can facilitate the selection of functional outcome measures in stroke clinical trial research and practice.

Clinical relevance in stroke outcome measures can be optimized by incorporating a framework of health and disability. The International Classification of Functioning, Disability and Health (ICF) is the World Health Organization framework for health and disability. The ICF provides the conceptual basis for measurement and policy formulations for disability and health. According to the ICF model, outcomes may be measured at the following levels: body functions and structure (impairment), activities, and participation. Activities and participation are affected by environmental and personal factors.[8]

All outcome measures selected should also have sound psychometric properties. The essential psychometric properties are reliability, validity, responsiveness to change, sensibility,[7] and minimal clinically important difference (MCID). Reliability of an outcome measure refers to the extent to which a score is free of random error[9]; validity is the capacity of an instrument to measure what it is intended and presumed to measure[10]; responsiveness to change is the ability of an outcome measure to detect clinically important changes[7]; sensibility refers to the overall appropriateness, importance, and ease of use of the instrument[2–5]; and the MCID helps to define a threshold that is considered to be an important improvement.[11–13] Generic outcome measures are useful for comparisons across populations and with

Address correspondence and reprint requests to: Sharon Barak, MSESS, Department of Physical Therapy, College of Public Health and Health Professions, P.O. Box 100154, Gainesville, FL 32610. E-mail: sbarak@phhp.ufl.edu.

normal age and gendered values. Condition-specific measures are more suitable for assessments within a specific client group. Whether the measure has been used within the stroke population is an important characteristic of an outcome measure, and it should be regarded as a relative indicator of how well the instrument might function within a given sample of individuals who have experienced stroke.[7,14]

Another essential factor in selecting outcome measures is to establish the purpose of the measurement. The purposes of outcome measures could be discriminative, predictive, or evaluative. Discriminative studies are designed to separate patients into discrete classes that can be defined according to specific diagnostic criteria. In predictive studies, patients are classified into groups against a known criterion or gold standard. Evaluative studies are intended to reflect clinically important changes.[15]

When selecting outcome measures use and timing of use, the natural history of stroke and stroke severity must be considered. Recovery after stroke is strongly influenced by time since onset and by baseline stroke severity. Individuals with stroke usually experience some degree of recovery.[16] More than 80% of patients with mild stroke reach maximum improvement in activities of daily living (ADL) function within 3 weeks, and thus the assessment of only ADL in this subgroup of individuals with stroke is insufficient to capture the full extent of stroke impact according to the ICF model.[16] These individuals may continue to have limitations in physical function, instrumental activities of daily living (IADL), and participation. Thus, a more global emphasis is needed in poststroke assessment for those with mild stroke. Patients with more severe stroke may not achieve independence in ADL, and in that population ADL assessment, as well as assessment of the other domains, is appropriate.[17]

The outcome measures selected for clinical trials may differ depending on whether the study is efficacy-oriented or effectiveness-oriented. The goal of efficacy trials is to optimize the chance of detecting a biological effect with as few patients as possible.[18,19] Because impairment scales may be the most sensitive to change and have the greatest capacity to differentiate between treatment groups, they are particularly useful for efficacy studies.[20] The aim of effectiveness trials is to determine whether interventions have beneficial results when they are administrated in the context of ordinary clinical practice.[21] Studies that focus on effectiveness are broadly conceptualized and are assessed not only for primary outcomes but also for a wide range of outcomes relevant to public health, such as comorbidity, quality of life, and cost effectiveness.[22,23]

It is also important for researchers and clinicians to understand the different methods of acquiring data. The main data acquisition methods consist of self-administered questionnaires, interviewer-administered interviews, or observational assessments, along with varying models including telephone-administered, face-to-face, or computerized/web-based methods. The assessments can be performed with the patient or with a proxy, such as a family member or health care provider.

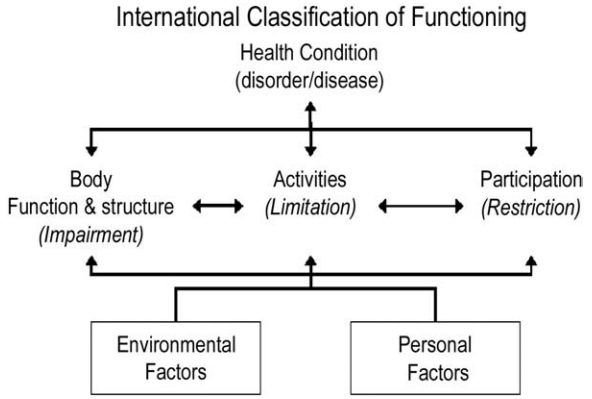

FIG. 1. The International Classification of Functioning (ICF) model. ICF domains are described from the perspective of the body, the individual, and society in two basic lists: (1) body function and structures and (2) activities and participation. *Functioning* is an umbrella term encompassing all body functions, activities, and participation. ICF also lists environmental and personal factors that interact with all these constructs.[27] Adapted from the WHO International Classification of Functioning, Disability and Health (2001).[27]

The purpose of this article is to provide a comprehensive overview of the issues in selecting stroke outcome measures and to characterize existing measures relative to these issues.

CONCEPTUAL FRAMEWORK

The ICF provides a conceptual framework for selection and classification of outcome measures.[24–26] Outcomes may be measured at any of these levels:

1. Body functions or structure (impairment): problems in body function or structure as a significant deviation or loss.
2. Activities: the execution of a task or action by an individual.
3. Participation: the involvement in a life situation.

Activities and Participation are affected by environmental and personal factors (referred to as contextual factors within the ICF).[27] The ICF model is presented in FIG. 1.

Impairments

Measures of impairment are the most closely related to the volume of brain loss and are probably the best markers of prognosis[28]; however, the extent to which measures of impairments will relate to the volume of brain loss will vary according to the region of the brain affected and stroke type. Nevertheless, according to the European Stroke Initiative,[29] poststroke disability assessment should comprise the impairment domains of motor weakness, sensory and propri-

oceptive deficits, and cognition impairments.[29] As impairment scales may be the most sensitive to change and have the greatest capacity to differentiate between treatment groups, they are particularly useful for efficacy studies.[20] However, for clinical significance and health policy it is important to relate changes in impairments to changes in activity and participation.

Activity

Activities measures are the most frequently used primary outcome measures in stroke. The most common domain of activity measurement is basic ADLs. However, in an unselected stroke population, approximately 60% of the patients will make a "complete recovery" in basic ADL.[8] Thus, measures of ADLs may have a ceiling effect and may not show a difference between groups in outcome, significantly reducing the power of any study.[28] For example, most patients with mild stroke spontaneously achieve independence in ADLs early, and therefore, make it difficult to detect an intervention effect on ADL. Thus, ADLs measures, such as the Barthel Index, will have a ceiling effect in stroke patients with mild deficits and other significant limitations may not be captured (e.g., important improvements in higher level functions such as household maintenance, shopping and quality-of-life status). Therefore, researchers need to stratify patients into different degrees of initial severity.[7] In the minor and moderate stroke strata the benchmarks of recovery must include measures of higher level of activity (i.e., IADLs) or mobility since they may be more sensitive to differences between groups[28,30] and they do not suffer from ceiling effect.[28] In the severe stroke patients' strata, assessment of recovery of basic ADLs and mobility may be an appropriate primary outcome measure.[17]

A challenge in all activity and mobility measures is that the link between the extent of loss at the level of pathology and impairment is not perfectly correlated and other factors may influence the outcome.[7,31] For example, an individual may improve in motor function, but without good social support to encourage independence, he or she may not become more independent in ADL, IADL, or participation.[32]

Participation

Although legislation, reports, and classification schemes promote the concept of participation as an important component of disability, the development of measures capturing the essence of participation has just begun.[33] One possible reason for the delay in development is that tasks subsumed within the participation level are relatively complex, more dependent on environmental influences and on social support, and usually assessed in the community by self or proxy report.[25]

The concept of quality of life is reflected in both participation and activity. However, quality of life is defined differently in quality of life models than in ICF activity and participation. In quality of life models, quality of life involves several core dimensions, including physical functioning, emotional well-being, social functioning, and role activities, as well as health perceptions and global assessment of life satisfaction.[34] The ICF defines activity as the execution of a task or action by an individual, and participation as the involvement in a life situation.[27]

Depression is another factor influencing stroke recovery. Depression may be considered an impairment that strongly influences activity and participation. It should always be considered for measurement in clinical practice and research, because symptoms occur in about one-third of poststroke patients.[35] Depending on the purpose of the study, depression may be considered as a primary outcome or a modifier of the relationship between impairment, activity and participation. There are established measures of depression that have been validated for stroke patients (e.g., Geriatric Depression Scale,[36] Beck Depression Inventory,[37] and Center for Epidemiologic Studies Depression).[38] Depression assessments should be taken from the patient rather than a proxy.

Environmental factors will also have an impact on activity and participation and are organized in sequence from the individual's most immediate environment to the general environment.[27] The family is one example of an important environmental factor. Indeed, it has been demonstrated that early involvement of the family unit is strongly correlated with patient adherence to therapy, better understanding between patient and caregiver of achievable outcomes, and improved communication between patient and caregivers.[39] Thus, the family or social support may be a modifier that needs to be considered for clinical research.

Measurement of recovery at just one level gives only a partial picture of the recovery process. For example, many ADLs can be performed despite the presence of significant impairments. If only the level of activity is monitored, the patterns of neurological recovery may be disguised. Measuring stroke recovery at the impairment, activity, and participation levels allows the determination of the impact of changes in impairments on changes in activity and perceived quality of life.[7]

International experts identified the categories that account for the fundamental and most striking aspects of stroke related functioning. They created the Brief ICF Core Set.[40] The categories included in the brief ICF core set are given in TABLE 1. In TABLE 2 we have drawn on the work of Salter et al.,[24–26] Duncan et al.,[41] and Gresham et al.[42] to present an annotated list of the most commonly used instruments of stroke trials and clinical practice, classified by ICF categories of impairment, activities, and environmental factors.

TABLE 1. *The Brief ICF Core Set for Stroke—Adapted**

ICF Component and Category Title
Body functions
Consciousness functions
Orientation functions
Muscle power functions
Mental functions of language
Body structures
Structure of brain
Activities
Walking
Speaking
Toileting
Eating
Environmental factors
Immediate family

* Adapted from Geyh et al.[40]

OUTCOME MEASURES' PSYCHOMETRIC PROPERTIES

Psychometric properties are critical to the selection of any outcome measure. The essential psychometric properties are reliability, validity, responsiveness, and sensibility.[7] Whether the measure has been used within the stroke population and whether it has an established MCID are also important characteristics of outcome measures.

Reliability

Reliability of an outcome measure refers to the extent to which a score is free of random error. A score on an outcome measure is composed of two parts: true variance and measurement variance. True variance captures the variability in the attribute of interest. Measurement variance is random error and represents variability due to other factors. Measurement variance may be due to a variety of factors, including fatigue, cognitive factors, and mode of test administration.[9] Reliability is defined as the proportion of the score that contains information about the attribute of interest as opposed to measurement error. It is expressed as a coefficient from 0 to 1, with 1 representing perfect reliability.[7] There are three basic ways to evaluate the reliability of a given instrument: internal consistency, interrater reliability, and test–retest reliability

Internal consistency reliability. Internal consistency reliability is the most frequently used estimate of the reliability of a measure. A measure of internal consistency is the average degree of association among the items on a test.[43] To compute internal consistency, a single version of an instrument is administered to a single group of test subjects at a single time point. The data are then analyzed for consistency.[44] According to Andresen,[45] excellent internal consistency is reported at ≥ 0.80, adequate is 0.70–0.79, and poor is <0.70.[45]

Interrater reliability. Interrater reliability concerns variation between two or more raters who measure the same group of subjects.[46] Many potential threats to interrater reliability exist in any test situation. For instance, Blackburn et al.[47] evaluated the reliability of measurements obtained with the Modified Ashworth Scale in the lower extremities of people with stroke. They reported poor levels of interrater reliability, despite use of written guidelines. In their study, the assessors had not been trained specifically in the use of the scale, suggesting that

TABLE 2. *Most Commonly Used Stroke Outcome Measures*

Assessment Type and Name
Body Structure (Impairments)
Neurological scales
National Institutes of Health Stroke Scale[41,42]
Motor function
Fugl–Meyer Assessment[24–26,41,42]
Modified Ashworth[24–26]
Cognitive scales
Neurobehavioral Cognition Status Exam[41,42]
Mini Mental State Examination[24–26,41,42]
Speech and language functions
Boston Diagnostic Aphasia Examination[41,42]
Western Aphasia Battery[41,42]
Visual perception
Motor-free Visual Perception Test[24–26]
Depression scales
Beck Depression Inventory[24–26,41,42]
Center for Epidemiologic Studies Depression[41,42]
Geriatric Depression Scale[41,42]
Activities
Activities of Daily Living
Barthel Index[24–26,41,42]
Functional Independence Measure[24–26,41,42]
Balance
Berg Balance Scale[24–26,41,42]
Mobility and motor function
Timed Up-and-Go[24–26]
10 Meter walk[175]
6 Minutes walk[149]
Wolf Motor Function Test[176]
Motor Assessment Scale[41,42]
Rivermead Motor Assessment[24–26]
Motricity Index[41,42]
Chedoke McMaster Stroke Assessment Scale[24–26]
Modified Rankin Handicap Scale[24–26]
Instrumental Activities of Daily Living
Frenchay Activities Index[24–26,41,42]
Older Americans Resources and Services[41]
Participation
Health status and quality of life
Medical Outcomes Study Short Form 36[24–26,41,42]
Stroke Specific Quality of life[24–26]
EuroQoL-5D[24–26]
Stroke Impact Scale[24–26,41]
Sickness Impact Profile (stroke-adapted version)[24–26,41,42]
Family
Family assessment device[41,42]

guidelines need to be accompanied by training of test administrators to achieve improved reliability.[47] Generally, 80% agreement between raters is the minimum required.[44]

Test–retest reliability. Test–retest reliability is the correlation between scores obtained by the same person on two separate occasions. The interpretation is complicated by the fact that actual changes may have occurred in behavior or functional status during the time interval itself. Thus, low test–retest reliability does not necessarily reflect the psychometric properties of the test.[43] Excellent test–retest reliability is ≥0.75, adequate is 0.4–0.74, and poor is ≤0.40.[24–26] Fitzpatrick et al.[14] recommend a minimum test–retest reliability of 0.90 if the measure is to be used to evaluate the ongoing progress of an individual in a treatment situation. Test–retest reliability of measures is often established in chronic stroke subjects who are not continuing to experience recovery.

Validity

Demonstrating reliability in measurement is essentially providing the existence of a stable or generalizable concept; however, reliability says nothing about the nature of the concept. Thus, a set of items may yield a repeatable score, but one that may be an invalid indicator of the construct under study.[48] Validity is the capacity of an instrument to measure what it is intended to and presumed to measure. Many types of validity are referred to in the literature, such as face, content, discriminative, convergent, predictive, and criterion.[10] Of these, the most important are criterion and predictive validity.[7] Criterion validity refers to the performance of the instrument against an external gold standard or the actual outcome that the test was developed to assess.[43] Predictive validity is a form of criterion validity[24–26] and is the degree to which a test can predict how well an individual will do in a future situation.[43]

Responsiveness

Responsiveness is sensitivity to changes within patients over time, which may be indicative of therapeutic effects.[24–26] Responsiveness is most commonly evaluated through correlation with other scores, effect sizes, standardized response means, relative efficiency and sensitivity and specificity of change scores. For example, when examining sensitivity to change in an expected direction, the standardized effect method categorizes <0.5 as small, 0.5–0.8 as moderate, and ≥0.8 as large.[24–26] Assessment of possible floor and ceiling effects is included, because they indicate limits to the range of detectable change beyond which no further improvement or deterioration can be noted.[24–26] Such effects can seriously damage the capacity of a trial to detect change. If patients achieve the top score on a major outcome scale at baseline, no improvement can be detected. Conversely, if patients start out at the bottom of a scale, no deterioration can be measured.[7] There are adequate floor and ceiling effects when ≤20% of patients attain either the minimum (floor) or maximum (ceiling) score.[24–26]

Several investigators have examined the sensitivity of common outcome measures used in stroke rehabilitation. For example, English et al.[49] investigated the sensitivity of gait speed, the Berg Balance Scale, and the Motor Assessment Scale. Gait speed and the Berg Balance Scale were both sensitive to change and demonstrated large effect sizes. The Motor Assessment Scale item five (walking) also showed a large effect size and was able to detect change among lower functioning subjects. The effect sizes of the other items of the Motor Assessment Scale were small, and the majority of subjects showed no change over time on these measures. Houlden et al.[50] compared the responsiveness of the Barthel Index and the Functional Independence Measure (FIM). They concluded that the Barthel Index and the total and physical FIM scores showed similar responsiveness, and that the cognitive FIM score was least responsive. These findings suggest that none of the FIM scores have any advantages over the Barthel Index.[50] For additional examples of stroke outcome measure sensitivity studies published in the past few years, please refer to Wallace et al.[51] and Hsueh et al.[52]

Sensibility

Sensibility refers to the overall appropriateness, importance, and ease of use of an instrument; it is a major factor determining the success or failure of a clinical measure. The primary consideration in choosing an outcome measure is the correspondence between the dimensions of the measure (impairment, activity, or participation) and the goals of the intervention and the study.[7] For example, if the goal of the intervention is to improve upper extremity motor recovery, select measures that reflect upper extremity motor function. In addition, the measures that are selected must not be burdensome for the patient, yet should capture the range of their abilities.[7]

Has the measure been used within the stroke population?

An important factor to consider when evaluating outcome measures' psychometric properties is whether or not the measure has previously been used within the stroke population. Reliability and validity are not fixed qualities of measures. They should be regarded as relative indicators of how well the instrument might function within a given sample or for a given purpose.[14,53] Sensitivity to change may likewise be condition- or purpose-specific.[24–26] For example, as previously mentioned, the Barthel Index has a ceiling effect in stroke patients with mild deficits, yet it may be one of the most sensitive measures in patients with more severe impairments.[7] It is important for a measure to have been tested for use in the population within which it will be used.[24–26]

MCID and the concept of sliding dichotomy

In the presence of a plethora of available instruments and evidence of their psychometric properties, outcomes research is currently faced with the challenge of interpretability[54] of the scores. When health status is measured, it is worth knowing whether an observed difference indicates a clinically significant or trivial effect on the patient's health status or quality of life. A statistically significant difference in health status or quality of life measures might be of little clinical or practical importance; it is more important to know the MCID.[55] Pursuit of the MCID is one important area of current work in interpretability.[11] Jaeschke[55] first defined an MCID as being "the smallest difference in score in the domain of interest which patients perceive as beneficial." Since then, the definition has varied. Looking only at articles published in the past few years, we see definitions such as "the smallest difference in a score that is considered to be worthwhile or important."[12,56] Several stakeholders would share an interest in determining the MCID. Researchers would use this for sample size determination, drug companies need this for interpreting the results of trials, and clinicians could use this to guide clinical care.[11]

The use of continuous scales *versus* ordinal scales is an important consideration in the calculation of clinically significant results. When the outcome measure is continuous, such as gait velocity, it is important to determine whether the measure has a meaningful change or an absolute change, by establishing a MCID. When outcome measures are ordinal, however, they must generally be converted according to severity as a dichotomous outcome of "favorable" or "unfavorable," in order to determine clinical relevance; that is, a cutoff score must be established to demarcate a positive or negative test.[57] For example, the Berg Balance Scale can be used to predict if a stroke patient is at risk for falling. A cutoff score of <45 is typically used to indicate that an individual may be at grater risk for falling.[58] Thus, a score of ≥ 45 is considered to be a "favorable" outcome, and a score of <45 is an "unfavorable" outcome.

However, an instrument that defines function dichotomously as "favorable" *versus* "unfavorable" does not accord with every day clinical practice[59] and may be too coarse to detect smaller degrees of MCID.[7]

The concept of sliding dichotomy is a novel approach that answers both of the major objections to the conventional dichotomous analysis. The idea is that, instead of taking a single definition of "good" outcome for all patients, the definition is tailored to each individual patient's baseline prognosis on entry to the trial.[59] For a patient with a very severe injury, independence in basic ADLs alone might be regarded as a good outcome. For a patient with a mild injury, however, only a return to community participation would be regarded as a good outcome.[59] In practice, the approach would be implemented by grouping patients into a number of bands according to their baseline prognosis. Each band would have a customized dichotomy of the outcome scale to differentiate between "good" and "bad" outcome. The total number of good outcomes in the intervention group would be compared with the corresponding number of good outcomes in the control group.[59] (For further information about the concept of sliding dichotomy, see Murray et al.[59])

Characteristics of the most commonly used stroke outcome measures relative to the psychometric properties described above are presented in the Appendix. MCID is one of the psychometric properties evaluated. Unfortunately, this psychometric property has not been evaluated in the majority of the most common stroke outcome measures. Nonetheless, establishment of MCID is critical in designing effectiveness studies or in clinical trials that will influence clinical decision making and health policy.

PURPOSE OF MEASUREMENT

There are three purposes of measurement: discriminative, predictive, and evaluative.[15] Each of these three purposes of measurement scale has a useful role to play in rehabilitation, but mismatching the types can result in incorrect assessment information.[43]

Discriminative scales

Discriminative scales are used to distinguish between individuals or groups with respect to underlying dimension when no external criterion or gold standard is available for validating these measures.[60] These scales are used in between-subjects experimental designs that use separate samples for each treatment condition.[61] Thus, if one had two groups of patients with stroke and wanted to examine the differences between the two groups in ADLs, one would require a discriminative scale.

Predictive scales

Predictive scales are used to classify individuals into a set of predefined measurement categories when a gold standard is available. This gold standard is subsequently used to determine whether individuals have been classified correctly. Let us assume that investigators had developed a mobility instrument that took >1 hour to administer. Because an hour represents a rather long test, it would be desirable to have a shorter version. One might choose a subsample from the original test and examine the performance of the new, shorter instrument using the original as a gold standard.[60]

Evaluative scales

Evaluative scales are used to measure the magnitude of longitudinal change in an individual or group[60] (within

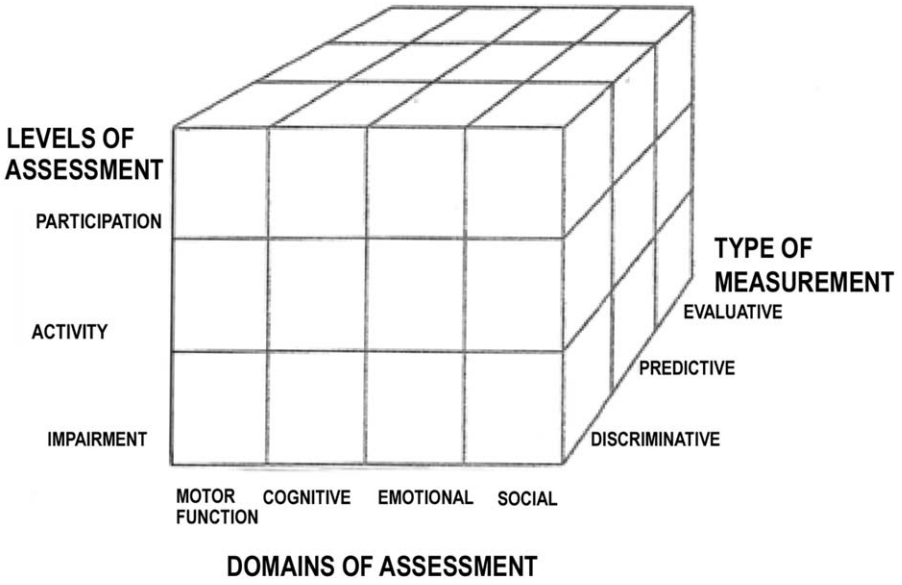

FIG. 2. A 3-dimensional model for functional assessment. Along one axis are the three areas of assessment (impairment, activity, and participation). Along the second axis are the domains of assessment that are generally accepted as relevant for rehabilitation outcomes, as well as in health status assessment. Along the third axis are the types of measurement. Adapted from Turner[43] (updated by the authors for terms and constructs).

subjects experimental design). Within-subjects experimental designs are experiments in which two sets of data are obtained from the same sample. They compare treatment effects by looking at changes in performance within each participant across treatments.[61] Thus, for an evaluative scale, we might ask whether a particular change in a patient's ADL score represents a trivial, small but important, moderate, or large improvement or deterioration.[62]

In summary, the distinction in type of measurement adds a third dimension to the conceptual framework. The 3-dimensional model of functional assessment was described by Turner[43] and its terms and constructs were updated by the authors. Along one axis are the three areas of assessment: impairment, activity, and participation. Along the second axis are the domains of assessment that are generally accepted as relevant for rehabilitation outcomes as well as in health status assessment. Along the third axis are the types of measurement. The modified 3-dimensional model of functional assessment is presented in FIG. 2. This model can be used to guide the questions the user needs to ask at the onset of the assessment task: What is the appropriate unit of analysis? How many, and which content domains are relevant? What is my assessment goal? Answers to these questions should help identify a preliminary set of outcome measures instruments, which can then be examined more closely for evidence of psychometric quality.[43]

NATURAL HISTORY OF STROKE AND STROKE SEVERITY

When considering the use of an outcome measure or the time of assessment, natural history of stroke and stroke severity should be considered.

Approximately 25% of patients worsen during the first 24 hours following stroke.[63] Beyond that first period, however, individuals with stroke usually experience some degree of recovery. Recovery is the most dramatic during the first 30 days after a stroke.[16,64,65] By the end of the first 3 months, patients who survive stroke almost always have less physical disability. Thus, measurements of activities (e.g., the Barthel Index and FIM) tend to show a plateau of gains by 3 months after stroke, partly owing to insensitivity of the scale to further improvements.[66] However, based on initial stroke severity there are different trajectories of recovery. For example, in more severe strokes recovery may be more protracted (FIG. 3).[67]

Efficacy and effectiveness trials

As we select outcome measures for interventions, it is particularly important to understand the distinction between efficacy-oriented and effectiveness-oriented clinical trials.

Efficacy trials. The focus of efficacy trials is usually a newly developed intervention or a promising modification of a well established one.[21] Whatever the investigative issue, the intention is to conduct a well-controlled experiment under ideal conditions, using relative homogeneous samples.[18,21] The goal of efficacy trials is to optimize the chance of detecting a biological effect with as few patients as possible.[18,19] Because impairment scales may be the most sensitive to change and have the greatest capacity to differentiate between treatment groups, they are particularly useful for efficacy studies.[20] Thus, the study endpoint will most likely reflect the impairment the treatment is attempting to minimize.[20] For example, if the intervention goal is to improve upper extremity motor function, then the measure selected will be Fugl–Meyer upper extremity, and researchers and clinicians should not expect major changes in mobility assessments. Finally, the duration of follow-up for

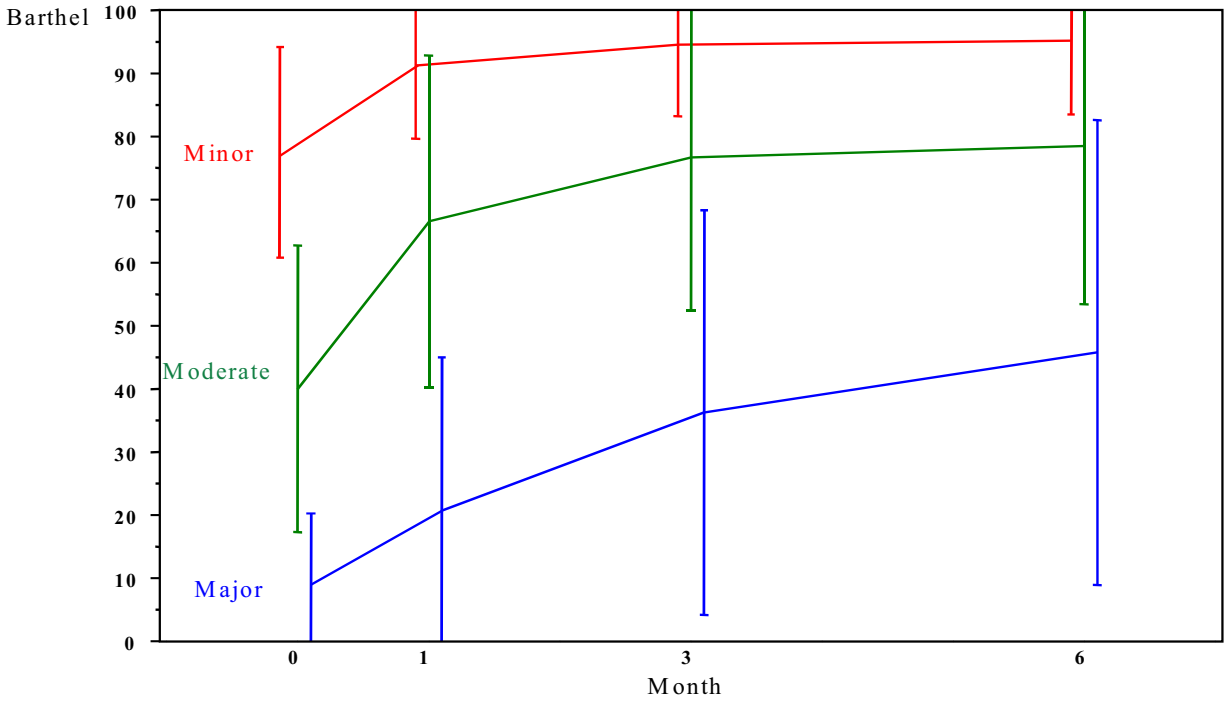

FIG. 3. The trajectory of Barthel ADL recovery for stroke patients with different levels of initial stroke severity. Subjects are stratified for severity using the Orpington Prognostic Scale. In the Barthel Index, a score of 0 represents complete inability and a score of 100 represents complete ability on all items. Adapted from the Kansas City Stroke Study, unpublished data (data collection started in October 1995 and was completed in 1999).

clinical endpoints (functional outcome) does not need to exceed 3 months in typical efficacy studies; shorter periods may be possible. A shorter time period will likely reduce variation in clinical outcome due to subsequent events unrelated to the study.[20]

Effectiveness trials. The aim of effectiveness trials is to determine whether interventions have beneficial results when they are administrated in the context of ordinary clinical practice. As such, effectiveness trials are principally concerned with the external validity of treatment outcomes.[21] Studies that focus on effectiveness are broadly conceptualized, use heterogeneous samples that are recruited in a variety of practice settings, and are assessed not only for primary outcomes but also for a wide range of outcomes relevant to public health, such as comorbidity, quality of life, and cost effectiveness.[22,23] In addition, participants tend to be followed for a longer duration, and data analysis can place greater emphasis on differences among subgroups. Features such as these just listed may indeed enhance the generalizability of a study, but they may also introduce possible confounds that allow the results to be attributed to factors other than the intervention itself.[21]

In effectiveness studies, the most clinically relevant outcome must be assessed. For the most part, these will include activities and participation measures. The most commonly used outcome measure in effectiveness studies have been the Barthel Index and the Rankin or modified Rankin scale[17]; however, the Barthel Index is known to be insensitive to small changes in functional status and to have significant ceiling effects. The Rankin scale has been criticized as inherently insensitive and for mixing objective and subjective items, which span impairment, activity, and participation aspects of recovery.[28] Given the limitations of the Barthel Index and Rankin Scale, the new Stroke Impact Scale (SIS) has been increasingly endorsed in effectiveness trials. The SIS has been developed to be a more comprehensive measure of health outcomes for stroke populations. The SIS incorporates meaningful dimensions of function and health-related quality of life into one self-report questionnaire. The SIS version 3 includes 59 items and assesses eight domains (strength, hand function, ADL and IADL, mobility, communication, emotion, memory and thinking, and participation or role function).[66]

Methods of acquiring data

Another important issue in selecting outcome measures is methods of acquiring data.

When assessment requires a form of self-report, several modes of assessment exist: trained interviewers vs. self-administered, administration by a healthcare professional or other proxy, and computerized adaptive test (CAT).

Trained interviewers *versus* self-administered. Questionnaires are either administered by trained interviewers or self-administered. Although having trained interviewers is resource intensive, it both ensures compliance and minimizes errors and missing items. The self-administered

approach is much less expensive, but increases the number of missing patients and missing responses. A compromise between the two approaches is to have the instrument completed under supervision. Another compromise is the telephone interview, which minimizes errors and missing data but dictates a relatively simple questionnaire structure.[60]

Proxy and healthcare professional's report. Impairment and activity measures can be performance-based, but participation and quality of life are most often self-reported.[24–26] Self-report measures are limited, however, by the cognition and communication problems of stroke survivors.[68] For example, in a large study that used mail-administered quality of life questionnaires, 50% of the stroke subjects were unable to complete the questionnaires by themselves.[69] Moreover, study results can be seriously compromised and misleading if subjects who are suffering from severe deficits are excluded. The inclusion of proxy data will increase sample size, improve generalizability, and reduce sample bias.[70] Nonetheless, the use of proxy respondents should be approached with caution.[24–26]

Proxy assessors tend to assess patients as more disabled than they appear on other measures of functional disability, including self-reported methods. This discrepancy becomes more pronounced for patients with more impaired levels of functioning.[71–73] This discrepancy could be explained by a difference in interpretation. Proxy respondents may be rating actual, observable performance, whereas patients may rate their perceived capability—what they think they are capable of doing, rather than what they actually do.[72] Unfortunately, a similar discrepancy has been noted in ratings when using healthcare professionals as proxy respondents, although in the opposite direction. Healthcare professionals may tend to rate patients higher than the patients themselves would.[73,74] Again, the discrepancy may be due to a difference in frame of reference. A healthcare professional may use a more disabled group as reference norm, whereas patients would simply compare themselves to prestroke conditions.[74] Clinicians and researchers also need to pay attention to measurement consistency. If researchers and clinicians use proxy respondents at the beginning of the intervention ("pre-test") they should be consistent and use the proxy respondents throughout the study or intervention ("post-test").

Data acquisition has typically relied on traditional, fixed-length tests, which tend to be long and require administering items that are high (or even too high) for those with low trait values and items that are low (or too low) for those with high trait values.[75] However, do all items need to be administered to every person? Can we get an accurate estimate of function if we administer fewer items, and do so without sacrificing precision? Can individual assessment be personalized by drawing from a large item-pool, based on that person's responses? The use of CAT methodology with a large item-pool is the new assessment frontier, and it may provide an effective solution to these measurement challenges.[76]

CAT. Computerized adaptive testing has been applied in educational and psychological testing for decades,[76] and it is currently being used to administer the Graduate Record Examination. Only recently has CAT technology been applied to rehabilitation and health service research.[77] Unlike fixed-length paper-and-pencil tests, CAT tests provide different test-item sets for each examinee based on that person's estimated trait (or ability) level.[78] An adaptive test first asks questions in the middle of the ability range, and then, based on the responses, asks subsequent questions that focus on relevant functional levels. Thus, precise information regarding an individual's functional ability level is obtained, with fewer items administered,[76] and the information about each individual can be assessed most efficiently.[78]

CAT is ideally suited to item response theory (IRT) methods.[79] IRT makes it possible to estimate an individual's trait levels with any subset of items in an item pool. Methods based on IRT overcome the limitations of ordinal data, provide detailed examination of item performance and respondent validity, and control for rater severity.[80] IRT methods have been widely used in the field of education[81]; in rehabilitation, they have been used to psychometrically assess the FIM.[82]

The simplest of the IRT models, the Rasch model, represents the essential elements for developing measures that are both efficient and precise.[83] The Rasch model breaks down assessing an individual into its most basic elements, person ability minus item difficulty. In using this formula to determine a person's ability level, the most information about an individual is obtained when person ability matches item difficulty or when the individual has a 50% probability of passing or being successful on an item.[84] Thus, it is unnecessary to administer all test items to every person. For example, if a person has a 50% probability of being successful at standing without any assistance device, it would be imprudent to ask that individual a very easy task (e.g., to sit down on a chair) or a very complex task (e.g., to run upstairs).

SUMMARY

Clinical investigators and clinicians are increasingly concerned with the selection of appropriate outcome measures, because these measures will have an impact on detecting treatment effects. There is no general consensus, however, on the battery of measures that should be used in clinical stroke trials and clinical practices. Thus, to improve the selection of stroke outcome measures, we offer for consideration the following recommendations:

1. Clinical relevance in stroke outcome measures can be optimized by incorporating the framework of

Health and Disability, the ICF. This model will help establish the domains of outcome measures.

2. All outcome measures should have established psychometric properties (e.g., reliability, validity, and sensitivity to change) and should have been tested in individuals with stroke.
3. The purpose of measurement should guide researchers and clinicians in identifiable areas of function that should be assessed (e.g., impairment, ADL, IADL).
4. The natural history of stroke and stroke severity must be considered when outcome measures are selected.
5. The type of study (efficacy *versus* effectiveness studies) should also dictate the type of outcome measures selected.
6. The mode of administration has to be taken into consideration (e.g., phone, interview, or self-report).

Acknowledgments: This article is the result of work supported by resources and facility usage at the Rehabilitation Outcomes Research Center (RORC), North Florida/South Georgia Veterans Health System, Gainesville, FL. The RORC is funded by ROCO1-124.

REFERENCES

1. American Heart Association. 2001 heart and stroke statistical update. Dallas, TX, 2000.
2. Duncan PW, Goldstein LB, Matchar D, Divine GW, Feussner J. Measurement of motor recovery after stroke: outcome assessment and sample size requirements. Stroke 1992;23:1084–1089.
3. Loewen SC, Anderson BA. Predictors of stroke outcome using objective measurement scales. Stroke 1990;21:78–81.
4. Wade DT, Wood VA, Hewer RL. Recovery after stroke: the first 3 months. J Neurol Neurosurg Psychiatry 1985;48:7–13.
5. Kinsella G, Ford B. Acute recovery from patterns in stroke patients: neuropsychological factors. Med J Aust 1980;2:663–666.
6. Roberts L, Counsell C. Assessment of clinical outcomes in acute stroke trials. Stroke 1998;29:986–991.
7. Duncan PW. Measuring recovery of function after stroke: clinical and measurement issues in selecting stroke outcome measures in clinical trials. In: Goldstein LB, editor. Restorative neurology: advances in pharmacotherapy for recovery after stroke. New York: Futura Publishing; 1998. p. 225–240.
8. Jorgensen HS, Pedersen PM, Kammersgaard L, Raaschou HO, Olsen TS. Epidemiology of stroke related disability. In: Duncan PW, editor. Clinics in geriatric medicine: stroke. Philadelphia: WB Saunders; 1999. p. 785–800.
9. Bergner M, Rothman ML. Health status measures: an overview and guide for selection. Annu Rev Public Health 1987;8:191–210.
10. Stewart AL. Psychometric consideration in functional status instruments. In: WONCA Classification Committee, editors. Functional status measurement in primary care. New York: Springer-Verlag; 1990.
11. Kirwan JR. Minimum clinically important difference: the crock of gold at the end of the rainbow? J Rheumatol 2001;28:439–444.
12. Hays RD, Woolley JM. The concept of clinically meaningful difference in health-related quality-of-life research. How meaningful is it? Pharmacoeconomics 2000;18:419–423.
13. Bellamy N, Carr A, Dougados M, Shea B, Wells G. Towards a definition of "difference" in osteoarthritis. J Rheumatol 2001;28:427–430.
14. 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.
15. Kirshner B, Guyatt GH. A methodological framework for assessing health indices. J Chronic Dis 1985;38:27–36.
16. Jorgensen HS, Nakayama H, Raaschou HO, Vive-Larsen J, Stoier M, Olsen TS. Outcome and time course of recovery in stroke. Part ii: Time course of recovery. The Copenhagen stroke study. Arch Phys Med Rehabil 1995;76:406–412.
17. Duncan PW, Lai SM, Keighley J. Defining post-stroke recovery: implications for design and interpretation of drug trials. Neuropharmacology 2000;39:835–841.
18. March JS, Silva SG, Compton S, Shapiro M, Califf R, Krishnan R. The case for practical clinical trials in psychiatry. Am J Psychiatry 2005;162:836–846.
19. Devuyst G, Bogousslavsky J. Recent progress in drug treatment for acute stroke. J Neurol Neurosurg Psychiatry 1999;67:420–425.
20. Stroke Therapy Academic Industry Roundtable II. Recommendations for clinical trial evaluation of acute stroke therapies. Stroke 2001;32:1598–1606.
21. Fuhrer MJ. Overview of clinical trials in medical rehabilitation: impetuses, challenges, and needed future directions. Am J Phys Med Rehabil 2003;82:S8–S15.
22. Schoenwald SK, Hoagwood K. Effectiveness, transportability, and dissemination of interventions: what matters when? Psychiatr Serv 2001;52:1190–1197.
23. Burns BJ. Children and evidence-based practice. Psychiatr Clin North Am 2003;26:955–970.
24. Salter K, Jutai JW, Teasell R, Foley NC, Bitensky J. Issues for selection of outcome measures in stroke rehabilitation: ICF body functions. Disabil Rehabil 2005;27:191–207.
25. Salter K, Jutai JW, Teasell R, Foley NC, Bitensky J, Bayley M. Issues for selection of outcome measures in stroke rehabilitation: ICF participation. Disabil Rehabil 2005;27:507–528.
26. Salter K, Jutai JW, Teasell R, Foley NC, Bitensky J, Bayley M. Issues for selection of outcome measures in stroke rehabilitation: ICF activity. Disabil Rehabil 2005;27:315–340.
27. World Health Organization. Introduction. In: International classification of functioning, disability and health (ICF). Geneva: WHO; 2001:3–25.
28. Duncan PW, Jorgensen HS, Wade DT. Outcome measures in acute stroke trials: a systematic review and some recommendations to improve practice. Stroke 2000;31:1429–1438.
29. Hack W, Kaste M, Bogousslavsky J, et al. European stroke initiative recommendations for stroke management-update 2003. Cerebrovasc Dis 2003;16:311–337.
30. Duncan PW, Wallace D, Lai SM, Johnson D, Embretson S, Laster LJ. The stroke impact scale version 2.0. Evaluation of reliability, validity, and sensitivity to change. Stroke 1999;30:2131–2140.
31. de Haan R, Aaronson N, Limburg M, Hewer RL, van Crevel H. Measuring quality of life in stroke. Stroke 1993;24:320–327.
32. Glass TA, Matchar DB, Belyea M, Feussner JR. Impact of social support on outcome in first stroke. Stroke 1993;24:64–70.
33. Gray DB, Hollingsworth HH, Stark SL, Morgan KA. Participation survey/mobility: psychometric properties of a measure of participation for people with mobility impairments and limitations. Arch Phys Med Rehabil 2006;87:189–197.
34. Shumaker SA, Anderson RT, Czajkowski SM. Psychological tests and scales. In: Spilker B, editor. Quality of life assessments in clinical trials. New York: Raven Press; 1990. p. 95–113.
35. Hsieh LP, Kao HJ. Depressive symptoms following ischemic stroke: a study of 207 patients. Acta Neurol Taiwan 2005;14:187–190.
36. Yesavage JA, Brink TL, Rose TL, et al. Development and validation of a geriatric depression screening scale: a preliminary report. J Psychiatr Res 1982;17:37–49.
37. Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J. An inventory for measuring depression. Arch Gen Psychiatry 1961;4:561–571.
38. Radlof LS. The CES-D scale: a self report depression scale for research in the general population. J Appl Psychol Meas 1977;1:385–401.

39. Evans RL, Bishop DS, Matlock AL, Stranahan S, Smith GG, Halar EM. Family interaction and treatment adherence after stroke. Arch Phys Med Rehabil 1987;68:513–517.
40. Geyh S, Cieza A, Schouten J, et al. ICF core sets for stroke. J Rehabil Med 2004; 135–141.
41. Duncan PW, Zorowitz R, Bates B, et al. Management of adult stroke rehabilitation care: a clinical practice guideline [online]. Stroke 2005;36:e100–e143. Available at: http://stroke.ahajournals.org/cgi/content/full/36/9/e100/DC1.
42. Gresham GE; Post-Stroke Rehabilitation Guideline Panel. Post-stroke rehabilitation: clinical practice guideline no. 16. DHHS Publication AHCPR 95-0662. Washington, DC: U.S Government Printing Office; 1995.
43. Turner RR. Rehabilitation: issues in functional assessment. In: Spilker B, editor. Quality of life and pharmacoeconomics in clinical trials. Philadelphia: Lippincott-Raven Publishers; 1996. p. 839–851.
44. Higgins PA, Straub AJ. Understanding the error of our ways: mapping the concepts of validity and reliability. Nurs Outlook 2006;54:23–29.
45. Andresen EM. Criteria for assessing the tools of disability outcomes research. Arch Phys Med Rehabil 2000;81:S15–S20.
46. Portney LG, Watkins MP. Reliability. In: Foundations of clinical research: applications to practice. 2nd ed. Upper Saddle River, NJ: Prentice Hall; 2000:79–110.
47. Blackburn M, van Vliet P, Mockett SP. Reliability of measurements obtained with the Modified Ashworth scale in the lower extremities of people with stroke. Phys Ther 2002;82:25–34.
48. Heitzmann CA, Kaplan RM. Assessment of methods for measuring social support. Health Psychol 1998;7:75–109.
49. English CK, Hillier SL, Stiller K, Warden-Flood A. The sensitivity of three commonly used outcome measures to detect change amongst patients receiving inpatient rehabilitation following stroke. Clin Rehabil 2006;20:52–55.
50. Houlden H, Edwards M, McNeil J, Greenwood R. Use of the Barthel Index and the Functional Independence Measure during early inpatient rehabilitation after single incident brain injury. Clin Rehabil 2006;20:153–159.
51. Wallace D, Duncan PW, Lai SM. Comparison of the responsiveness of the Barthel Index and the motor component of the Functional Independence Measure in stroke: the impact of using different methods for measuring responsiveness. J Clin Epidemiol 2002;55:922–928.
52. Hsueh IP, Lin JH, Jeng JS, Hsieh CL. Comparison of the psychometric characteristics of the Functional Independence Measure, 5 item Barthel Index, and 10 item Barthel Index in patients with stroke. J Neurol Neurosurg Psychiatry 2002;73:188–190.
53. Lorentz WJ, Scanlan JM, Borson S. Brief screening tests for dementia. Can J Psychiatry 2002;47:723–733.
54. Guyatt GH, Cook DJ. Health status, quality of life, and the individual. JAMA 1994;272:630–631.
55. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials 1989;10:407–415.
56. Beaton DE, Boers M, Wells GA. Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. Curr Opin Rheumatol 2002;14:109–114.
57. Portney LG, Watkins MP. Validity of measurements. In: Foundations of clinical research: applications to practice. Mehalik C, editor. 2nd ed. Upper Saddle River, NJ: Prentice Hall; 2000. p. 79–110.
58. Berg KO, Wood-Dauphinee SL, Williams JI, Maki B. Measuring balance in the elderly: validation of an instrument. Can J Public Health 1992;83:S7–S11.
59. Murray GD, Barer D, Choi S, et al. Design and analysis of phase III trials with ordered outcome scales: the concept of the sliding dichotomy. J Neurotrauma 2005;22:511–517.
60. Guyatt GH, Jaeschke R, Feeny DH, Patrick DL. Measurements in clinical trials: choosing the right approach. In: Spilker B, editor. Quality of life and pharmacoeconomics in clinical trials. Philadelphia: Lippincott-Raven Publishers; 1996. p. 41–49.
61. Gravetter FJ, Wallnau LB. Hypothesis tests with two independent samples. In: Statistics for the behavioral sciences. Knight V, Stoddard F, Bruckman R, editors. 5th ed. Belmont, CA: Wadsworth/Thomson Learning; 2000.
62. Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. Ann Intern Med 1993;118:622–629.
63. Castillo J. Deteriorating stroke: diagnostic criteria, predictors, mechanisms and treatment. Cerebrovasc Dis 1999;9(Suppl 3):1–8.
64. Binkofski F, Seitz RJ. Modulation of the bold-response in early recovery from sensorimotor stroke. Neurology 2004;63:1223–1229.
65. Carmichael ST, Tatsukawa K, Katsman D, Tsuyuguchi N, Kornblum HI. Evolution of diaschisis in a focal stroke model. Stroke 2004;35:758–763.
66. Lai SM, Studenski S, Duncan PW, Perera S. Persisting consequences of stroke measured by the Stroke Impact Scale. Stroke 2002;33:1840–1844.
67. Studenski SA, Wallace D, Duncan PW, Rymer M, Lai SM. Predicting stroke recovery: three- and six-month rates of patient-centered functional outcomes based on the Orpington Prognostic Scale. J Am Geriatr Soc 2001;49:308–312.
68. Duncan PW, Lai SM, Tyler D, Perera S, Reker DM, Studenski S. Evaluation of proxy responses to the Stroke Impact Scale. Stroke 2002;33:2593–2599.
69. Dorman PJ, Slattery J, Farrell B, Dennis MS, Sandercock PA. A randomized comparison of the EuroQoL and short form-36 after stroke. United Kingdom collaborators in the international stroke trial. BMJ 1997;315:461.
70. Magaziner J, Simonsick EM, Kashner TM, Hebel JR. Patient-proxy response comparability on measures of patient health and functional status. J Clin Epidemiol 1988;41:1065–1074.
71. Segal ME, Gillard M, Schall R. Telephone and in-person proxy agreement between stroke patients and caregivers for the Functional Independence Measure. Am J Phys Med Rehabil 1996;75:208–212.
72. Hachisuka K, Ogata H, Ohkuma H, Tanaka S, Dozono K. Test–retest and inter-method reliability of the self-rating Barthel Index. Clin Rehabil 1997;11:28–35.
73. Sneeuw KC, Aaronson NK, de Haan RJ, Limburg M. Assessing quality of life after stroke. The value and limitations of proxy ratings. Stroke 1997;28:1541–1549.
74. McGinnis GE, Seward ML, DeJong G, Osberg JS. Program evaluation of physical medicine and rehabilitation departments using self-report Barthel. Arch Phys Med Rehabil 1986;67:123–125.
75. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Med Care 2000;38:II28-II42.
76. Andres PL, Black-Schaffer RM, Ni P, Haley SM. Computer adaptive testing: a strategy for monitoring stroke rehabilitation across settings. Top Stroke Rehabil 2004;11:33–39.
77. Dijkers MP. A computer adaptive testing simulation applied to the FIM instrument motor component. Arch Phys Med Rehabil 2002;84:384–393.
78. Butcher JN, Perry J, Hahn J. Computers in clinical assessment: historical developments, present status, and future challenges. J Clin Psychol 2004;60:331–345.
79. Weiss DJ. Adaptive testing by computer. J Consult Clin Psychol 1985;53:774–789.
80. Velozo CA, Kielhofner G, Lai JS. The use of Rasch analysis to produce scale-free measurement of functional ability. Am J Occup Ther 1999;53:83–90.
81. Segall DO. General ability measurement: an application of multidimensional item response theory. Psychometrika 2001;66:79–97.
82. Linacre JM, Heinemann AW, Wright BD, Granger CV, Hamilton BB. The structure and stability of the Functional Independence Measure. Arch Phys Med Rehabil 1994;75:127–132.
83. Wright BD, Stone MH. Best test design. Chicago: Mesa Press; 1979.
84. Smith RM. Rasch measurement models: interpreting WINSTEPS/BIGSTEPS and FACETS output. Chicago: Mesa Press; 1999.

85. D'Olhaberriague L, Litvan I, Mitsias P, Mansbach HH. A reappraisal of reliability and validity studies in stroke. Stroke 1996; 27:2331–2336.
86. Lyden PD, Lau GT. A critical appraisal of stroke evaluation and rating scales. Stroke 1991;22:1345–1352.
87. Brott T, Adams HP, Olinger CP, et al. Measurements of acute cerebral infarction: a clinical examination scale. Stroke 1989;20: 864–870.
88. Lyden P, Brott T, Tilley B, et al. Improved reliability of the NIH stroke scale using video training. NINDS TPA stroke study group. Stroke 1994;25:2220–2226.
89. Muir KW, Weir CJ, Murray GD, Povey C, Lees KR. Comparison of neurological scales and scoring systems for acute stroke prognosis. Stroke 1996;27:1817–1820.
90. Gladstone DJ, Danells CJ, Black SE. The Fugl–Meyer assessment of motor recovery after stroke: a critical review of its measurement properties. Neurorehabil Neural Repair 2002;16:232–240.
91. Wolf SL, Catlin PA, Ellis M, Archer AL, Morgan B, Piacentino A. Assessing the Wolf Motor Function test as outcome measure for research in patients after stroke. Stroke 2001;32:1635–1639.
92. Morris DM, Uswatte G, Crago JE, Cook EW, Taub E. The reliability of the Wolf Motor Function test for assessing upper extremity function after stroke. Arch Phys Med Rehabil 2001;82: 750–755.
93. Sloan RL, Sinclair E, Thompson J, Taylor S, Pentland B. Interrater reliability of the modified Ashworth scale for spasticity in hemiplegic patients. Int J Rehabil Res 1992;15:158–161.
94. Gregson JM, Leathley MJ, Moore AP, Smith TL, Sharma AK, Watkins CL. Reliability of measurements of muscle tone and muscle power in stroke patients. Age Ageing 2000;29:223–228.
95. van Wijck FM, Pandyan AD, Johnson GR, Barnes MP. Assessing motor deficits in neurological rehabilitation: patterns of instrument usage. Neurorehabil Neural Repair 2001;15:23–30.
96. Pandyan AD, Price CI, Rodgers H, Barnes MP, Johnson GR. Biomechanical examination of a commonly used measure of spasticity. Clin Biomech 2001;16:859–865.
97. Kiernan RJ, Mueller J, Langston JW, Van Dyke C. The Neurobehavioral Cognitive Status Examination: a brief but quantitative approach to cognitive assessment. Ann Intern Med 1987;107: 481–485.
98. Lamarre CJ, Patten SB. A clinical evaluation of the Neurobehavioral Cognitive Status Examination in a general psychiatric inpatient population. J Psychiatry Neurosci 1994;19:103–108.
99. Schwamm LH, Van Dyke C, Kiernan RJ, Merrin EL, Mueller J. The Neurobehavioral Cognitive Status Examination: comparison with the Cognitive Capacity Screening Examination and the Mini-Mental State Examination in a neurosurgical population. Ann Intern Med 1987;107:486–491.
100. Osmon DC, Smet IC, Winegarden B, Gandhavadi B. Neurobehavioral Cognitive Status Examination: its use with unilateral stroke patients in a rehabilitation setting. Arch Phys Med Rehabil 1992;73:414–418.
101. Toedter LJ, Schall RR, Reese CA, Hyland DT, Berk SN, Dunn DS. Psychological measures: reliability in the assessment of stroke patients. Arch Phys Med Rehabil 1995;76:719–725.
102. Dick JP, Guiloff RJ, Stewart A, et al. Mini-Mental State Examination in neurological patients. J Neurol Neurosurg Psychiatry 1984;47:496–499.
103. Tombaugh TN, McIntyre NJ. The Mini-Mental State Examination: a comprehensive review. J Am Geriatr Soc 1992;40:922–935.
104. Agrell B, Dehlin O. Mini-Mental State Examination in geriatric stroke patients. Validity, differences between subgroups of patients, and relationships to somatic and mental variables. Aging 2000;12:439–444.
105. Grace J, Nadler JD, White DA, et al. Folstein vs. Modified Mini-Mental State Examination in geriatric stroke. Stability, validity, and screening utility. Arch Neurol 1995;52:477–484.
106. Gresham GE; Post-Stroke Rehabilitation Guideline Panel. Attachments. In: Post-stroke rehabilitation: clinical practice guideline no. 16. DHHS Publication AHCPR 95-0662. Washington, DC: U.S. Government Printing Office; 1995.
107. Goodglass H, Kaplan E. The assessment of aphasia and related disorders. 2nd ed. Media, PA: Williams & Wilkins; 1983.
108. Goodglass H, Kaplan E. Test procedures and rationale. In: Manual for the Boston Diagnostic Aphasia Examination (BDAE). Philadelphia: Lea and Febiger; 1983.
109. Kertesz A. The Western Aphasia Battery. New York: Grune and Stratton; 1982.
110. Su CY, Chang JJ, Chen HM, Su CJ, Chien TH, Huang MH. Perceptual differences between stroke patients with cerebral infarction and intracerebral hemorrhage. Arch Phys Med Rehabil 2000;81:706–714.
111. Mazer BL, Korner-Bitensky NA, Sofer S. Predicting ability to drive after stroke. Arch Phys Med Rehabil 1988;79:743–750.
112. Aben I, Verhey F, Lousberg R, Lodder J, Honig A. Validity of the Beck Depression Inventory, Hospital Anxiety and Depression Scale, SCL-90, and Hamilton Depression Rating Scale as screening instruments for depression in stroke patients. Psychosomatics 2002;43:386–393.
113. Roberts RE, Vernon SW, Rhoades HM. Effects of language and ethnic status on reliability and validity of the Center for Epidemiologic Studies-Depression Scale with psychiatric patients. J Nerv Ment Dis 1989:177:581–592.
114. Roberts RE, Vernon SW. The Center for Epidemiologic Studies Depression Scale: its use in a community sample. Am J Psychiatry 1983;140:41–46.
115. Shinar D, Gross CR, Price TR, Banko M, Bolduc PL, Robinson RG. Screening for depression in stroke patients: the reliability and validity of the Center for Epidemiologic Studies Depression Scale. Stroke 1986;17:241–245.
116. Parikh RM, Eden DT, Price TR, Robinson RG. The sensitivity and specificity of the Center for Epidemiologic Studies Depression Scale in screening for post-stroke depression. Int J Psychiatry Med 1988;18:169–181.
117. Comstock GW, Helsing KJ. Symptoms of depression in two communities. Psychol Med 1976;6:551–563.
118. Burns A, Lawlor B, Craig S. Rating scales in old age psychiatry. Br J Psychiatry 2002;180:161–167.
119. Brink TL, Yesavage JA, Lum B, et al. Depressive symptoms and depressive diagnoses in a community population. Arch Gen Psychiatry 1982;45:1078–1084.
120. Robinson RG, Price TR. Post-stroke depressive disorders: a follow-up study of 103 patients. Stroke 1982;13:635–641.
121. Agrell B, Dehlin O. Comparison of six depression rating scales in geriatric stroke patients. Stroke 1989;20:1190–1194.
122. Hsueh IP, Lee MM, Hsieh CL. Psychometric characteristics of the Barthel activities of daily living index in stroke patients. J Formos Med Assoc 2001;100:526–532.
123. Sulter G, Steen C, De Keyser J. Use of the Barthel Index and Modified Rankin Scale in acute stroke trials. Stroke 1999;30: 1538–1541.
124. Uyttenboogaart M, Stewart RE, Vroomen PC, De Keyser J, Luijckx GJ. Optimizing cutoff scores for the Barthel Index and the Modified Rankin Scale for defining outcome in acute stroke trials. Stroke 2005;36:1984–1987.
125. van der Putten JJ, Hobart JC, Freeman JA, Thompson AJ. Measuring change in disability after inpatient rehabilitation: comparison of the responsiveness of the Barthel Index and the Functional Independence Measure. J Neurol Neurosurg Psychiatry 1999;66: 480–484.
126. Duncan PW, Samsa GP, Weinberger M, Goldstein LB, Bonito A, Witter DM, et al. Health status of individuals with mild stroke. Stroke 1997;28:740–745.
127. Beninato M, Gill-Body KM, Salles S, Stark PC, Black-Schaffer RM, Stein J. Determination of the minimal clinically important difference in the FIM instrument in patients with stroke. Arch Phys Med Rehabil 2006;87:32–39.
128. Cavanagh SJ, Hogan K, Gordon V, Fairfax J. Stroke-specific FIM models in an urban population. J Neurosci Nurs 2000;32:17–21.
129. Adunsky A, Fleissig Y, Levenkrohn S, Arad M, Noy S. Clock drawing task, Mini-Mental State Examination and Cognitive-Functional Independence Measure: relation to functional outcome of stroke patients. Arch Gerontol Geriatr 2002;35:153–160.

130. Berg K, Wood-Dauphinee S, Williams JI. The balance scale: reliability assessment with elderly residents and patients with an acute stroke. Scand J Rehabil Med 1995;27:27–36.
131. Mao HF, Hsueh IP, Tang PF, Sheu CF, Hsieh CL. Analysis and comparison of the psychometric properties of three balance measures for stroke patients. Stroke 2002;33:1022–1027.
132. Shumway-Cook A, Brauer S, Woollacott M. Predicting the probability for falls in community-dwelling older adults using the Timed Up & Go test. Phys Ther 2000;80:896–903.
133. Shumway-Cook A, Woollacott MH. Motor control: theory and practical applications. Baltimore, MD: Williams & Wilkins 1995.
134. Podsiadlo D, Richardson S. The timed "up & go": a test of basic functional mobility for frail elderly persons. J Am Geriatr Soc 1991;39:142–148.
135. Whitney SL, Poole JL, Cass SP. A review of balance instruments for older adults. Am J Occup Ther 1998;52:666–671.
136. Rockwood K, Awalt E, Carver D, MacKnight C. Feasibility and measurement properties of the Functional Reach and the Timed Up and Go tests in the Canadian study of health and aging. J Gerontol A Biol Sci Med Sci 2000;55:M70–M73.
137. Siggeirsdottir K, Jonsson BY, Jonsson H Jr, Iwarsson S. The timed 'up & go' is dependent on chair type. Clin Rehabil 2002; 16:609–616.
138. Collen FM, Wade DT, Bradshaw CM. Mobility after stroke: reliability of measures of impairment and disability. Int Disabil Stud 1990:12:6–9.
139. Perry J, Garrett M, Gronley JK, Mulroy SJ. Classification of walking handicap in the stroke population. Stroke 1995;26:982–989.
140. Goldie PA, Matyas TA, Evans OM. Deficit and change in gait velocity during rehabilitation after stroke. Arch Phys Med Rehabil 1996;77:1074–1082.
141. Salbach NM, Mayo NE, Higgins J, Ahmed S, Finch LE, Richards CL. Responsiveness and predictability of gait speed and other disability measures in acute stroke. Arch Phys Med Rehabil 2001; 82:1204–1212.
142. Collin C, Wade D. Assessing motor impairment after stroke: a pilot reliability study. J Neurol Neurosurg Psychiatry 1990;53: 576–579.
143. Kosak M, Smith T. Comparison of the 2-, 6-, and 12-minute walk tests in patients with stroke. J Rehabil Res Dev 2005;42:103–107.
144. Peeters P, Mets T. The 6-minute walk as an appropriate exercise test in elderly patients with chronic heart failure. J Gerontol A Biol Sci Med Sci 1996;51:M147–151.
145. Redelmeier DA, Bayoumi AM, Goldstein RS, Guyatt GH. Interpreting small differences in functional status: the six-minute walk test in chronic lung disease patients. Am J Respir Crit Care Med 1997;155:1278–1282.
146. Solway S, Brooks D, Lacasse Y, Thomas S. A qualitative systematic overview of the measurement properties of functional walk tests used in the cardiorespiratory domain. Chest 2001;119: 256–270.
147. Berry MJ, Rejeski WJ, Adair NE, Zaccaro D. Exercise rehabilitation and chronic obstructive pulmonary disease stage. Am J Respir Crit Care Med 1999;160:1248–1253.
148. Dobkin BH. Short-distance walking speed and timed walking distance: redundant measures for clinical trials? Neurology 2006; 66:584–586.
149. Eng JJ, Chu KS, Dawson AS, Kim CM, Hepburn KE. Functional walk tests in individuals with stroke: relation to perceived exertion and myocardial exertion. Stroke 2002;33:756–761.
150. Kunkel A, Kopp B, Muller G, Villringer K, Villringer A, Taub E, Flor H. Constraint-Induced Movement Therapy for motor recovery in chronic stroke patients. Arch Phys Med Rehabil 1999;80: 624–628.
151. Wolf SL, Lecraw DE, Barton LA, Jann BB. Forced use of hemiplegic upper extremities to reverse the effect of learned nonuse among chronic stroke and head-injured patients. Exp Neurol 1989;104:125–132.
152. Poole JL, Whitney SL. Motor assessment scale for stroke patients: concurrent validity and interrater reliability. Arch Phys Med Rehabil 1988;69:195–197.
153. Malouin F, Pichard L, Bonneau C, Durand A, Corriveau D. Evaluating motor recovery early after stroke: comparison of the Fugl–Meyer assessment and the Motor Assessment Scale. Arch Phys Med Rehabil 1994;75:1206–1212.
154. Lincoln N, Leadbitter D. Assessment of motor function in stroke patients. Physiotherapy 1979;65:48–51.
155. Kwakkel G, Kollen BJ, van der Grond J, Prevo AJ. Probability of regaining dexterity in the flaccid upper limb: impact of severity of paresis and time since onset in acute stroke. Stroke 2003;34: 2181–2186.
156. Cole B, Finch E, Gowland C, Mayo NE. Heart of the matter: template for outcome measures. Adult motor and functional activity measures. In: Basmajian J, editor. Physical rehabilitation outcome measures Toronto, Ontario: Canada Communication Group-Publishing; 1994. p. 38–78.
157. Gowland C, Stratford P, Ward M, Moreland J, Torresin W, Van Hullenaar S, et al. Measuring physical impairment and disability with the Chedoke–McMaster Stroke Assessment. Stroke 1993; 24:58–63.
158. Wolfe CD, Taub NA, Woodrow EJ, Burney PG. Assessment of scales of disability and handicap for stroke patients. Stroke 1991; 22:1242–1244.
159. Segal ME, Schall RR. Determining functional/health status and its relation to disability in stroke survivors. Stroke 1994;25:2391–2397.
160. Gurland BJ, Wilder DE. The care interview revisited: development of an efficient, systematic clinical assessment. J Gerontol 1984;39:129–137.
161. Kane RA, Kane RL. Multidimensional measures. In: Assessing the elderly: a practical guide to measurement. Lexington, Massachusetts: Lexington Books 209–247,1981.
162. Doble SE, Fisher AG. The dimensionality and validity of the Older Americans Resources and Services (OARS) activities of daily living (ADL) scale. J Outcome Meas 1998:2:4–24.
163. Kane RA, Kane RL. Measures of physical functioning in long-term care. In: Assessing the elderly: a practical guide to measurement. Lexington, Massachusetts: Lexington Books 25–66,1981.
164. Andresen EM, Meyers AR. Health-related quality of life outcomes measures. Arch Phys Med Rehabil 2000;81:S30–S45.
165. Ferguson RJ, Robinson AB, Splaine M. Use of the reliable change index to evaluate clinical significance in SF-36 outcomes. Qual Life Res 2002;11:509–516.
166. Dorman P, Slattery J, Farrell B, Dennis M, Sandercock P. Qualitative comparison of the reliability of health status assessments with the EuroQoL and SF-36 questionnaires after stroke. United kingdom collaborators in the international stroke trial. Stroke 1998;29:63–68.
167. Walters SJ, Munro JF, Brazier JE. Using the SF-36 with older adults: a cross-sectional community-based survey. Age Ageing 2001;30:337–343.
168. Coons SJ, Rao S, Keininger DL, Hays RD. A comparative review of generic quality-of-life instruments. Pharmacoeconomics 2000: 17:13–35.
169. Agency for Healthcare Research and Quality. Valuation of the EuroQoL-5d health states. Available at: http://www.Ahrq.Gov/rice/eq5dproj.Htm.2006, Accessed Date: December 2005.
170. Coast J, Peters TJ, Richards SH, Gunnell DJ. Use of the EuroQoL among elderly acute care patients. Qual Life Res 1998;7:1–10.
171. Duncan PW, Bode RK, Min Lai S, Perera S. Rasch analysis of a new stroke-specific outcome scale: the Stroke Impact Scale. Arch Phys Med Rehabil 2003;84:950–963.
172. van Straten A, de Haan RJ, Limburg M, van den Bos GA. Clinical meaning of the stroke-adapted sickness impact profile-30 and the sickness impact profile-136. Stroke 2000;31:2610–2615.
173. Golomb BA, Vickrey BG, Hays RD. A review of health-related quality-of-life measures in stroke. Pharmacoeconomics 2001;19: 155–185.
174. Miller IW, Bishop DS, Epstein NB, Keitner GI. The McMaster Family Assessment Device: reliability and validity. J Marital Fam Ther 1985;11:345–356.
175. Frytak J. Measurement. J Rehabil Outcomes Meas 2000;4:15–31.
176. Taub E, Miller NE, Novack TA, et al. Technique to improve chronic motor deficit after stroke. Arch Phys Med Rehabil 1993; 74:347–354.

APPENDIX

TABLE A1. *Psychometric Properties of the Most Commonly Used Stroke Outcome Measures*

Assessment Name	Time to Administer	Reliability	Validity	Responsiveness	Minimal Clinically Important Difference (MCID) or Cutoff Scores	Tested for Stroke Patients?	Strengths	Weaknesses
National Institutes of Health Stroke Scale	5–10 min.[41,42]	Excellent.[85,86]	Excellent.[85,86]	Low sensitivity.[41,42]	Scores ≥ 25 indicate very severe neurologic impairment, 15–24 severe impairment, 5–14 mild to moderately severe impairment, <5 mild impairment.[87]	Yes.[85,88,89]	Brief, reliable, can be administered by non-neurologists.[41,42]	Low sensitivity[41,42]; ceiling effect.[89]
Fugl–Meyer Assessment (FM)	30–40 min.[41,42]	Excellent.[24]	Excellent (caution with the balance subscale).[24]	Adequate.[24]	The MCID on the FM scale is not yet known; >10 points (10%) change in FM motor scores may represent clinically meaningful improvement based on clinical experience with this scale and consultation with physical therapists and stroke neurologists.[90]	Yes.[24,91,92]	Extensively evaluated measure, good validity and reliability for assessing sensorimotor function and balance.[41,42]	Considered too complex and time-consuming by many[41,42]; examines synergy patterns that no longer form the basis for many functionally oriented treatments.[91]
Modified Ashworth	Testing should be relatively brief.[24]	Adequate.[24]	Poor.[24]	Insufficient data.[24]	Not established.	Yes.[93,94]	Has widespread clinical acceptance, is routinely used to assess spasticity, is the current clinical standard.[95]	Some questions remain whether the scale is a valid measure of spasticity[96]; no standardized testing procedures or guidelines for the use of the scale exist, reliability of the test is dependent upon the muscle being assessed.[94]
Neurobehavioral Cognition Status Exam (NCSE)	10–20 min.[97]	The NCSE had good test–retest reliability (κ = 0.69), but the inter-rater reliability was not as good (κ = 0.57).[98]	Has well demonstrated validity.[97,99]	Sensitive to cognitive effects of stroke, although there was little discrimination between left-sided and right sided strokes.[100]	Patients who have scores that are lower than those in the average range on any test are impaired in that specific skill.[99] For geriatric population (77.6 years ± 5.2 years) the normal ranges for the different tests are: Orientation = 11.7 ± 0.7; Attention test = 7.7 ± 0.9; Comprehension = 5.9 ± 0.4; Repetition = 12.4 ± 0.8; Naming = 8.2 ± 1.1; Constructions = 4.4 ± 1.5; Memory = 10.1 ± 2.2; Calculations = 3.9 ± 0.3; Similarities = 5.6 ± 1.3; Judgment = 5.0 ± 0.8.[97]	Yes.[97,99–101]	Predicts gain in Barthel Index scores, unrelated to age.[41,42]	Does not distinguish right from left hemisphere, no reliability studies in stroke, correlates with education[41,42]; visual and motor problems make completion of block design difficult.[101]
Mini Mental State Examination	10 min.[41,42]	Excellent.[24]	Adequate.[24]	Insufficient data.[24]	A score of ≤23 is the generally accepted cutoff point indicating presence of cognitive impairment.[102] Levels of impairment have also been classified as none (24–30); mild (18–24), and severe (0–17).[103]	Yes.[104,105]	Widely used for screening.[41,42] Brief.[106]	Several functions with summed score, heavily language dependent, likely to misclassify patients with aphasia.[41,42]

TABLE A1. *Continued*

Boston Diagnostic Aphasia Examination	1–4 h.[41,42]	Kuder–Richardson reliability coefficient for subtests: range 0.68–0.98 (about two-thirds range 0.90–0.98).[107]	Adequately evaluated.[106]	Not tested.	A score of 6 on the Aphasia Severity Rating Scale indicates no aphasia; scores of 5, 4, and 3 indicate mild to moderate aphasia.[107]	Yes.[108]	Widely used, comprehensive, sound theoretical rationale.[41,42]	Time to administer long, half of patients cannot be classified.[41,42]
Western Aphasia Battery	1–4 h.[41,42]	Adequately evaluated.[106]	Standardized in 365 aphasic and 162 normal individuals.[106]	Not tested.	A score of ≤93.8 represents presence of aphasia.[109]	Yes.[109]	Widely used, comprehensive.[41,42]	Time to administer long, aphasia quotients and taxonomy of aphasia not well validated.[41,42]
Motor-free Visual Perception Test	10–15 min.[24]	Excellent.[24]	Adequate.[24]	Insufficient data.[24]	Not reported.	Yes.[110]	Widely used[111]; simple, well tolerated by subjects.[110]	Provides a global score and, therefore, gives less information about specific visual dysfunction than a scale providing domain-specific scores.[110]
Beck Depression Inventory	10 min.[41,42]	Excellent.[24]	Excellent.[24]	Poor.[24]	A score of ≥10 is generally accepted cutoff score for the indication of possible depression.[112]	Yes.[112]	Widely used, easily administered, norms available. Good with somatic symptoms.[41,42]	Less useful in elderly and in patients with aphasia or neglect, high rate of false positives, somatic items may not be due to depression.[41,42]
Center for Epidemiologic Studies Depression	<15 min.[41,42]	High internal consistency (α 0.83–0.91), acceptable test–retest reliability,[113,114] high inter-rater reliability.[115]	Good construct validity in both clinical and community samples.[113,114]	Adequately evaluated.[115,116]	A cutoff score of 16 is generally used to distinguish depressed individuals from nondepressed,[117] with a score of ≥23 indicating significant depression.[118]	Yes.[101,115]	Brief, easily administered, useful in elderly, effective for screening in stroke population.[41,42]	Not appropriate for aphasic patients[41,42]; does not measure only depressive symptoms but a combination of symptoms common to both major depression and generalized anxiety disorders.[114]
Geriatric Depression Scale-long form (GDS)	10 min.[41,42]	Excellent (test–retest reliability = 0.85; internal consistency = 0.94.[119]	Concurrent vs. Zung and Beck scales and Hamilton scale.[120]	Adequately evaluated.[106]	Normal ≤10; mildly depressed; 11–20; and moderately to severely depressed ≥21.[119]	Yes.[101]	Brief, easy to use with elderly, cognitively impaired, and those with visual or physical problems or low motivation.[41,42] Agrell and Dehlin[121] compared the GDS to five other depression rating scales in a stroke population and found that the GDS and the Zung performed best of the six instruments.	High false-negative rates in minor depression.[41,42]
Barthel Index	≤20 min.[26]	Excellent.[24]	Excellent.[41,42,122]	Adequate.[24]	Poor outcome: <60[123]; >95 is a pivotal score for determining which patients do not require help from another person for everyday activities.[124]	Yes.[122,125]	Widely used for stroke; excellent validity and reliability.[41,42]	Large reported ceiling and floor effects.[126]

TABLE A1. *Continued*

Assessment Name	Time to Administer	Reliability	Validity	Responsiveness	Minimal Clinically Important Difference (MCID) or Cutoff Scores	Tested for Stroke Patients?	Strengths	Weaknesses
Functional Independence Measure (FIM)	~30 min.[26]	Excellent.[24]	Adequate.[26]	Adequate.[24]	According to Beninato et al.,[127] FIM change scores from admission to discharge associated with MCID were 22, 17, and 3 for the total FIM, motor FIM, and cognitive FIM, respectively. Wallace et al.[51] estimated the MCID for the motor subscore at 1–3 mo after stroke based on 1 level of change on the Modified Rankin Scale; they estimated the MCID on the motor subtest to be 11 points.	Yes.[52,71,125]	Widely used for stroke; measures mobility, activities of daily living, cognition, functional communication.[41,42] Use of 7-point scale increases sensitivity versus other disability scales.[106]	Ceiling and floor effects at the upper and lower ends of function[106]; reliability is dependent upon the individual conducting the assessment[128]; cognition subtest is not a preferred cognition assessment tool for stroke patients due to insufficient sensitivity.[129]
Berg Balance Scale	10–15 min.[130]	Excellent.[26]	Excellent.[26]	Excellent.[26]	A score of 56 indicates functional balance, scores < 45 indicate that an individual may be at greater risk of falling.[58]	Yes.[130]	Simple, well established with stroke patients, sensitive to change.[41,42]	May suffer from decreased sensitivity in early stages post stroke among severely affected patients.[131]
Timed Up-and-Go	A few minutes[26]	Excellent.[26]	Adequate.[26]	Insufficient data.[26]	Cutoff score for high risk for falls in community-dwelling older adults is $\geq$14 s to complete the test.[132] According to Shumway-Cook,[133] adults without neurological impairments who are independent with balance and mobility skills are able to perform the Timed Up-and-Go test in $<$10 s.	Yes.[134]	Quick, easy to administer, can be accomplished in community, timed scores are objective, requires no specialized equipment and training.[135]	May not be suitable for use among individuals exhibiting cognitive impairment[136]; addresses relatively few aspects of balance.[135,137]
10 Meter walk	A few minutes	High.[138]	Validity established in many studies.[138]	Sensitive measure of recovery of post-stroke mobility.[139]	When 10-m gait velocity measures are stratified into clinically meaningful functional ambulation classes such as household ambulation ($<$0.4 m/s), limited community ambulation (0.4–0.8 m/s), and community ambulation ($>$0.8 m/s), changes in 10-m gait velocity is clinically meaningful.[139]	Yes.[138,140]	Simple, related to the severity of impairment in the home and the community[139]; less likely to show a ceiling effect.[141]	A variation in gait speed of $\leq$25% limits the tests reliability [142]; the 10-m walk test does not provide a continuous-scale assessment of gait recovery following stroke. Early during rehabilitation phase, patients may not be able to walk 5–10 m, and are therefore not testable (floor effect)[143]; as patients improve, walking speed $>$5–10 m becomes a less credible measure of walking speed over more functionally relevant distances outside the home.[143]

TABLE A1. *Continued*

6 Minute walk	6 min.	Acceptable inter- and intrarater reliability (0.78 and 0.74, respectively).[143]	Considered a valid test for assessing exercise capacity of elderly patients with chronic heart failure and chronic obstructive pulmonary disease.[144]	Standardized response mean = 1.52.[143]	The MCID is estimated to be 54 m for chronic lung disease patients.[145] The MCID for the stroke population was not established.	Yes.[143]	Simple[143]; well-tolerated, and reflects activities of daily living[146]; a continuous variable without floor or ceiling effects[143]; quick and easy to implement and can be completed by many patients.[147]	Cannot assess other important aspects of gait such as quality of movement, balance, use of assist devices, and amount of physical assistance needed.[143] The 6-min walk is usually described as a measure of endurance, fatigability, and cardiovascular fitness,[148] but there are a number of stroke-specific impairments that could potentially alter the outcome of the test—for example, factors such as muscle weakness, balance impairment, and spasticity might influence the distance walked.[149]
Wolf Motor Function Test	30 min.[92]	High interrater reliability, internal consistency[91,92]; high test–retest reliability.[92]	Supported criterion validity.[91]	Sensitive to treatment effects in subjects undergoing constraint-induced therapy treatment.[150] Appears to be more sensitive than other upper extremity tools.[151]	Not reported.	Yes.[91]	Reliably measures functional ability in a variety of activities, tests a wide range of functional tasks and explores both performance time and quality of movement, detailed written protocol[92]; requires few tools and minimal training.[91]	Time consuming.[92]
Motor Assessment Scale	15 min.[41,42]	High.[152]	High concurrent validity.[152,153]	Item 5 (walking) showed a large effect size; the other items have small effect sizes (d = 0.36–0.5) and the majority of subjects showed no change over time.[49]	Not reported.	Yes.[49,152,153]	Brief assessment of movement and physical mobility[41,42]; good reliability and validity.[106]	Reliability assessed only in stable patients.[41,42]
Rivermead Motor Assessment	≤40 min.[26]	Adequate.[26]	Adequate.[26]	Poor.[26]	Collin and Wade[142] propose that a total score difference of ±3 may represent a clinically relevant change.	Yes (stroke specific).[26]	The time spent making the assessment is directly related to the patient's level of motor functioning[154]; can be self-reported.[138]	Time consuming[142]; the validity of the scale as a Guttman scale is questionable.[26]
Motricity Index (MI)	5 min.[41,42]	Good.[138]	Good.[138]	Sensitivity not tested.[41,42]	Based on clinical experience, Kwakkel et al.[155] considered MI arm score of ≥11, and MI leg score of ≥25 to be associated with good outcome of upper extremity dexterity at 6 mo after stroke.	Yes.[138]	Brief and simple assessment of motor function of arm, leg, and trunk.[41,42,138]	Sensitivity not tested.[41,42]
Chedoke McMaster Stroke Assessment Scale	1 h.[26]	Excellent.[26]	Excellent.[26]	Excellent.[26]	Change of 8 points on the disability inventory equates to clinically important changes as judged by client and caregiver.[156]	Yes (stroke specific).[26]	Improved interpretability and sensitivity to small physical changes[157]; requires little equipment.[26]	Complex to administer, long, not suited to proxy use.[26]

TABLE A1. *Continued*

Assessment Name	Time to Administer	Reliability	Validity	Responsiveness	Minimal Clinically Important Difference (MCID) or Cutoff Scores	Tested for Stroke Patients?	Strengths	Weaknesses
Modified Rankin Handicap Scale	15 min.[26]	Excellent.[26]	Adequate.[26]	Adequate.[26]	Score of ≤2 reflects a good outcome; 2, unfavorable outcome.[124]	Yes (stroke specific).[26]	Simple, well studied reliability, requires no special tools or training.[26]	Subjective score, lack of clear criteria by which to assign grades.[158]
Frenchay Activities Index	5 min.[159]	Adequate.[26]	Excellent.[26]	Poor.[26]	Patients with a score of ≤15 are classified as "inactive."[158]	Yes (stroke specific).[26]	Developed specifically for stroke patients; assesses broad array of activities[41,42]; simple to administer, requires no training, suitable for use with proxy respondents.[26]	Interobserver reliability not tested; sensitivity probably limited.[41,42] Lack of standard guidelines for administration.[26]
Older Americans Resources and Services Instrumental Activities of Daily living (OARS-IADL)	45 min.	5-wk test–retest correlations for 30 elderly subjects was 0.71.[160] Correlation coefficient = 0.82.[161]	Limited.[162]	Insufficient data.	The OARS-IADL yields information about functional activity in five domains: social resources, economic resources, mental health, physical health, and activities of daily living (comprising seven physical activities of daily living and seven instrumental activities of daily living). From responses to the questions in each domain, a rater makes a judgment of the functional status in each dimension along a six-point scale where 1 = excellent functioning and 6 = totally impaired. The six-point scale can be collapsed to a dichotomous scale: individuals rated 1–3 are considered to be functioning adequately and those rated 4–6, impaired for that dimension.[161]	No; tested on community residents, patients referred to clinic because of age-related problems, and persons living in institutions.[161]	Tasks are considered necessary for community living.[161]	Includes items such as taking one's own medicine. When OARS-IADL is administered in a nursing home, the interviewer is instructed to ask the respondent whether he could perform the task if it were necessary, but the validity of this procedure has not been established. Although the OARS-IADL items seem adequate for most general purposes, other instrumental activities of daily living instruments with more detailed breakdowns of functioning will be more appropriate for hospital-based patients and for patients with chronic long-term-care needs or multiple disabilities.[163]

TABLE A1. *Continued*

Medical Outcomes Study Short Form 36 (SF-36)	<10 min.[164]	Adequate.[25]	Excellent.[25]	Excellent.[25]	Ferguson et al.[165] estimated SF-36 clinically significant change for pre- and post- (repeated measures) assessment using the Reliable Change Index (RCI).* For the total normative sample (n = 2474), physical functioning RCI = 17.07 points; Role-functioning-physical RCI = 31.26 points; Bodily pain RCI = 20.76 points; General health RCI = 24.81 points; Vitality RCI = 21.48 points; Social functioning RCI = 35.85 points; Role functioning-emotional RCI = 38.47 points; Mental health RCI = 20.01 points; Physical composite scale RCI = 7.47 norm-based t-score units; and Mental composite scale RCI = 9.70 norm-based t-score units. These values are not stroke specific.	Yes.[69,166]	Widely used in the United States[106]; brief, can be self-administered or administered by phone or interview, simple to administer[167]; standardized norms are available for several countries.[25]	Possible floor effect in seriously ill patients[41,42]; low rates of agreement between proxy respondent and patient respondent ratings.[159]
Stroke Specific Quality of life	10–15 min.	Excellent.[25]	Adequate.[25]	Adequate.[25]	Not reported.	Yes (stroke specific).[25]	Patient centered development, which may increase relevance to the patients it is intended to assess; no special training required for administration.[25]	Not well studied. Although it has been tested among severe stroke population, there are no standardized or normative values available for comparison.[25]
EuroQoL-5D	2–3 min.[168]	Adequate.[25]	Adequate.[25]	Adequate.[25]	The EQ-5D consists of five dimensions: mobility, self-care, usual activities, pain and discomfort, and anxiety and depression. Each dimension has three levels: level 1, no health problems; level 2, moderate health problems; and level 3, extreme health problems.[169]	Yes.[69,166]	Short, simple, high response rates[69]; has been evaluated for use with proxy[25]; no special training required for administration.[25]	The ability to self-complete is directly related to age and cognitive function[170]; low reliability with a proxy respondent.[166]
Stroke Impact Scale (SIS)	15–20 min.[25]	Adequate.[25]	Excellent.[25]	Poor.[25]	Changes in SIS domain scores of ~10–15 points.[30]	Yes (stroke specific).[25]	Assesses multiple domains of stroke recovery without administering multiple tests[28]; does not have significant ceiling or floor effects, validated with telephone administration[17]; proxies provide valid information.[68]	The originators of the scale report the majority of information currently available on the psychometric acceptability of this scale,[25] but the emotion domain seems to be less psychometrically acceptable than the other domains.[171]

TABLE A1. *Continued*

Assessment Name	Time to Administer	Reliability	Validity	Responsiveness	Minimal Clinically Important Difference (MCID) or Cutoff Scores	Tested for Stroke Patients?	Strengths	Weaknesses
Sickness Impact Profile (stroke-adapted version)	20–30 min.[41,42]	Adequate.[25]	Adequate.[25]	Insufficient data.[25]	Patients with a total score of >33 have poor health profiles.[172]	Yes (stroke specific).[25]	Comprehensive and well evaluated, broad range of items reduces floor or ceiling effects[41,42]; no special equipment or training is required.[25]	Time to administer somewhat long; evaluates behavior rather than subjective health[41,42]; does not assess pain, recreation, energy, general health perceptions, overall quality of life, or stroke symptoms.[173]
Family assessment device	30 min.[41,42]	Excellent.[41,42]	Excellent[41,42]	Not tested.	Cutoff score > 2.0 indicates unhealthy family.[174]	No stroke specific studies found.	Widely used in stroke, computer scoring available, excellent validity and reliability, available in multiple languages.[41,42]	Assessment subjective; sensitivity not tested; ceiling and floor effects.[41,42]

* The RCI statistic determines the magnitude of change score necessary for a given self-report measure to be considered statistically reliable. Note that the RCI alone does not indicate clinical significance. By itself, the RCI expresses only the amount of change between pre- and post-treatment scores on the SF-36 that would be statistically reliable. The SF-36 RCIs reported are relatively large, meaning that fairly substantial change scores in the SF-36 scales are needed for clinical significance. This is due in large part to variability in the size of reliability coefficients among the SF-36 scales. In general, the lower the reliability coefficient of a given SF-36 scale, the larger the RCI.[165]