Neurotherapeutics (2017) 14:161–175
DOI 10.1007/s13311-016-0488-5

REVIEW

The Spectrum of Functional Rating Scales in Neurology Clinical Trials

Pushpa Narayanaswami[1]

Published online: 27 October 2016

Abstract The selection of an appropriate outcome measure is crucial to the success of a clinical trial, in order to obtain accurate results, which, in turn, influence patient care and future research. Outcomes that can be directly measured are mortality/survival. More frequently, neurology clinical trials evaluate outcomes that cannot be directly measured, such as disability, cognitive function, or change in symptoms of the condition under study. These complex outcomes are abstract ideas or latent constructs and are measured using rating scales. Functional rating scales typically assess the ability of patients to perform tasks and roles for everyday life. Rating scales should be valid (measure what they are supposed to measure), reliable (provide similar results if administered under the same conditions), and responsive (able to detect clinically important changes over time). The clinical relevance of rating scales depends on their ability to detect a minimal clinically important difference, and should be distinguished from statistical significance. Most rating scales are ordinal scales and have limitations. Modern psychometric methods of Rasch analysis and item response theory, termed latent trait theory, are increasingly being utilized to convert ordinal data to interval measurements, both to validate existing scales and to develop new scales. Patient-reported outcomes are being increasingly used in clinical trials and have a role in clinical quality assessment. The PROMIS and NeuroQoL databases are excellent resources for rigorously developed and validated patient-reported outcomes.

Key Words Functional rating scales · outcome measures · clinimetrics · classical test theory · Rasch analysis · item response theory

Introduction

"When you can measure what you are speaking about, and express it in numbers, you know something about it", said Baron William Thomson Kelvin, after whom the Kelvin temperature scale is named. The essential question in a clinical trial consists of 4 elements, familiar to epidemiologists as the acronym "PICO": population, intervention, cointervention, and outcome. Two other features, timing and setting, are useful in specific clinical situations [1]. A clinical trial question may therefore effectively be summarized as follows: in a population of patients with the condition of interest, X, does the intervention in question, Y, as compared with Z, result in outcome O? The selection of the population, intervention, and cointervention are all intuitively important aspects of the trial design, but without selecting an appropriate outcome, and a rigorous method of measuring that outcome, the trial is futile. Reliable results cannot be obtained without an accurate and applicable outcome measure. As the number of treatment options for neurologic disorders expands, rigorous demonstration of efficacy, as measured by well-designed outcome measures, is necessary for approval by the Food and Drug Administration (FDA) or similar regulatory bodies, and, in turn, frequently defines reimbursement by third-party payers.

An obvious maxim is that the outcomes measured should be clinically relevant to the patient. The major

✉ Pushpa Narayanaswami
pnarayan@bidmc.harvard.edu

[1] Beth Israel Deaconess Medical Center/Harvard Medical School,
Neurology TCC-8, BIDMC, 330 Brookline Avenue,
Boston, MA 02215, USA

types of outcomes assessment include all-cause mortality/ survival, biomarkers, and clinical outcome assessments (COAs; Table 1). A COA "is any assessment that may be influenced by human choices, judgment or motivation and may support either direct or indirect evidence of treatment benefit. Unlike biomarkers, they rely completely on an automated process or algorithm. COAs depend on the implementation, interpretation, and reporting from a patient, clinician or observer. The four types of COAs are clinician reported outcome measures (CROs), patient reported outcome measures (PROs), observer reported outcome measures (OROs) and performance outcome measures" (http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/ucm370262.htm). Performance may be measured either by direct observation and quantification such as timed tests, or by reports of what activities can be performed from patients or caregivers.

A direct and obvious outcome, requiring little measurement is mortality/ survival time. More frequently in neurology, clinical improvement is measured by a constellation of outcomes such as improvement in weakness or degree of disability, ability to perform activities of daily living (ADLs) or instrumental ADLs, improvement in cognition or behavior, or change in symptoms specific to the condition of interest. These complex outcomes, termed "latent constructs" or "latent variables", are abstract ideas that cannot be measured directly; they are usually measured using rating scales or instruments. The scales relate a factor or variable that can be observed (the manifest variable) to the latent construct. For instance, the manifest variable of ability to run may be used to measure the latent construct of disability. A rating scale or instrument is composed of items, which are the fundamental units of measurement. Each item of the scale measures a manifest variable that is related to the latent variable, and takes on a specific value, depending on the magnitude of the latent variable in the respondent [2, 3]. Therefore, rating scales use numerical values to represent the characteristics of the outcomes being measured. The latent construct of pain can be operationalized in this paradigm into several items describing the frequency, severity, duration, character, and so on, each of which can be rated or ranked by numbers reflecting a magnitude of each item response. Rating scales are commonly used as primary or secondary outcomes in clinical trials. The accurate measurement of the effect of an intervention and detection of clinical changes depends directly on the quality of the rating scale. The quality of the rating scale also has implications for sample size estimations in study design because sample size estimations take into account the expected effect size, or difference in outcome between the control and treatment groups [4]. The science of measuring and analyzing psychological variables is referred to as psychometrics, and the term clinimetrics is applied to the design, administration, and interpretation of tests to measure clinical and epidemiologic outcomes, such as symptoms or signs, disease progression, or ADLs [5, 6]; these methods are increasingly being applied for accurate measurements of outcomes [3]. They are used to evaluate existing rating scales, as well as in the development of new scales.

Clinical trial outcomes may be reported by clinicians, patients, or observers/caregivers. CROs are assessments that are performed by investigators with some professional training related to the measurement and interpretation of the outcome. They involve interpretation of observable manifestations or phenomena related to the condition of interest. CROs may be performance measures such as the Timed Up and Go (TUG) test [7], rating scales such as the modified Ashworth scale for spasticity [8], or global ratings of change (Clinical Global Impression) [9]. CROs may also consist of readings, where the clinician collects physical data such as number of swollen joints, and so on. These tend to be less relevant to neurology. More recently, there has been an emphasis on PROs and instruments or scales that are designed to capture them. PROs are measures that are directly obtained from patients without interpretation by clinicians or caregivers [10]. Objective measures and CROs may not provide information about the effectiveness of the intervention from the patient's perspective, and may not correlate well with patients' viewpoints of health or well-being. In some diseases such as epilepsy, objective measures such as survival may not be pertinent; on the contrary, patient reports of seizure frequency and adverse events are more germane to evaluate the effectiveness of a new anticonvulsant drug. In other conditions such as restless leg syndrome, patient's perceptions of their symptoms, and the symptoms' impact on functioning are the only outcomes that can be measured because there is no clinical standard. These outcomes are typically measured using PRO instruments which measure symptom status, functional status, or quality of life. Some rating scales, such as the Unified Parkinson's Disease Rating Scale [11], combine PROs and CROs.

The term functional outcomes may be used to represent measures that reflect overall health status. Functions have been defined in this context as "the manner in which a patient can successfully perform tasks and roles required for every day that are meaningful to the patient and part of typical life" (http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/ucm370262.htm). The World Health Organization International Classification of Functioning, Disability and Health recognizes that function is a multidimensional concept operating at the level of the body and structures (physical impairment), at the level of the individual (functional activity or disability) and in societal participation (handicap) (https://www.cdc.gov/nchs/data/icd/icfoverview_finalforwho10sept.pdf). Functional status may be measured

Table 1 Types of outcome measures

Type of outcome measure	Example
Direct, objective	Mortality/survival
Biomarkers	Physiologic parameters such as blood pressure, laboratory tests on blood or other specimens or tissues
Clinical outcome assessments:*	
Clinician-reported outcomes	Modified Ashworth scale for spasticity
Patient-reported outcomes	Short Form-36 general health survey for quality of life
Observer-reported outcomes	School Function Assessment to evaluate students' participation in academic and social functions in elementary school
Performance outcomes†	Timed Up and Go test for mobility
	Amyotrophic Lateral Sclerosis Functional Rating Scale- Revised

*"Clinical Outcome Assessment (COA) is any assessment that may be influenced by human choices, judgment or motivation and may support either direct or indirect evidence of treatment benefit. COAs depend on the implementation, interpretation, and reporting from a patient, clinician or observer" (http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/ucm370262.htm)

† Performance outcomes may be measured either by direct observation and quantification such as timed tests, or by reports from patients or caregivers.

by performance based tests that are assessed by the clinician (TUG) [7], or by patient/caregiver reports of the ability to perform specific tasks or ADLs [Amyotrophic Lateral Sclerosis Functional Rating Scale-Revised (ALSFRS-R)] [12].

Assessing the Clinical Relevance of Functional Rating Scales

The selection of a functional rating scale depends on 2 major factors: first, does the scale measure what it is supposed to, that is, does it have face validity; and, second, is it clinically relevant, that is, are the outcomes measured by the scale important to patients? Several types of validity are used to determine the ability of a scale to measure a clinically important change (see section "The Ability of a Rating Scale to Measure a Clinically Important Difference"). However, face validity is an overall subjective impression of agreement between the outcome to be measured and what is actually measured by the scale. It is a simple "face-value" assessment. A scale that measures mobility will not necessarily provide information regarding quality of life, although it may have some predictive value for this construct, because impaired mobility may indirectly affect quality of life.

Rating scales may consist of single items or multiple items. Single-item scales are easily interpreted by clinicians. For instance, stage 2 on the Hoehn and Yahr scale for Parkinson's disease provides the clinician a clear picture of bilateral symptoms, minimal disability, with involvement of posture and gait, which is distinguishable from stage 1 or 3 [13]. However, the disadvantage of these scales is their poor reliability, validity, and responsiveness (see section "The Ability of a Rating Scale to Measure a Clinically Important Difference"); reliability is poor because they are associated with considerable random error [3]. Random errors are fluctuations in measurement due to limitations of the measuring scale, and are due to chance. For instance, flipping a balanced coin 10 times will not yield 5 heads and 5 tails (Fig. 1). Validity of a single-item scale is low because it is impossible to measure complex latent constructs such as disability or quality of life with a single question. Single-item scales can also be limited by their subjectivity where there is no reference framework or context (they are too vague), and hence different subjects use their own frames of reference [3]. For instance, the terms "slight increase", "more marked increase", or "considerable increase" in muscle tone on the modified Ashworth scale are susceptible to subjective interpretation by the clinician, although further definitions of these terms are provided [8]. The ability of single-item scales to detect change (responsiveness) is also problematic because each division of the scale encloses a broad range of the variable being measured; consequently, sensitivity to detect small changes is low [3].

Multiple-item scales consist of a set of items, each with multiple ordered response categories which are assigned a numerical value. The scales combine numerical values from all of items into a composite value, called the raw score, summed score, or scale score. This composite score is a measure of the latent variable that the scale is meant to estimate [3]. There are advantages to this approach; reliability is improved because the combination of multiple items reduces random error. Because the variables are broken down into multiple smaller components, validity and responsiveness are also improved [3]. The main disadvantage is that they are not as easily interpreted as single-item scores. The Barthel Index for ADL assessment is an example of a multiple item score, ranging between 0 and 100 across 10 items [14]. The composite score is useful to compare patients but does not

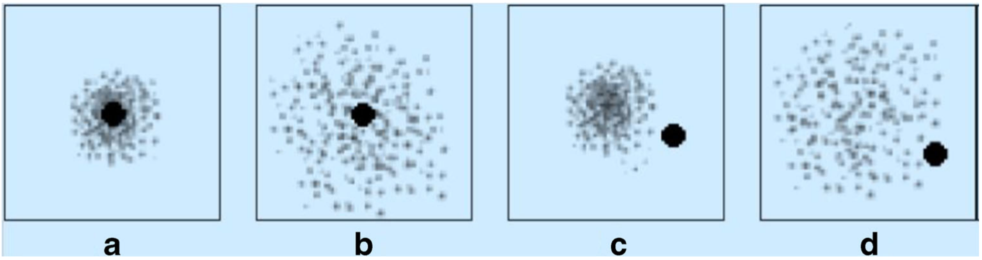

Fig. 1 Random error and systematic error. The black center represents the "truth". The individual dots represent the results of studies. (A) The dots do not fall on the black center is because of a small degree of random error. (B) The dots are more widely dispersed because of a larger degree of random error. (C, D) The dots or results of individual studies are off target because of systematic error. In (C) there is a small degree of random error, while (D) shows a larger degree of random error (from Gary Gronseth, MD, with permission)

provide details of the patient's level of independence with respect to specific activities. A score of 20 does not explicitly inform the clinician what activities can and cannot be performed. The scores obtained from multiple-item scales can be used for group comparisons but not for absolute measurements of a single subject [3]. The selection of single-item *versus* multiple-item scales, however, is dependent on the specific context. Longitudinal assessments or population studies often use the simpler single-item scales because of the time burden and greater potential for missing responses with multiple-item scales [15].

Rating scales can also be classified as generic measures and condition specific measures. Generic scales measure aspects of health-related functioning, such as mobility (TUG test) [7], quality of life [Short Form Health Survey (SF-36)] [16], and physical and social health (Sickness Impact Profile) [17]. Condition-specific scales, as the name implies, measure changes in 1 or more aspects of the specific condition of interest, reflecting aspects of function that are associated with that condition. Several such scales are used in neurologic disorders, including the Unified Parkinson's Disease Rating Scale [11], ALSFRS-R [12], and Kurtzke Extended Disability Status Scale for multiple sclerosis [18], to name a few. Not uncommonly, generic rating scales such as SF-36 are used to complement the information obtained by condition-specific scales. The choice of a measure, generic *versus* condition-specific entails some thought regarding their scope: generic measures are broad, providing information about several aspects of health and can be used across conditions as long as they have face validity for the outcome of interest. They can compare health status in patients receiving different interventions, or with different diseases. However, they are not sensitive to small and potentially clinically important treatment changes. They do not capture the outcome of greatest importance for the particular condition being studied because they do not isolate the constructs that are most relevant to that condition. Condition-specific scales, on the other hand, measure changes in outcomes that are most characteristic of the condition. However, they are much narrower in focus and may miss some effects of the intervention [19]. Because complications represent undesirable outcomes associated with treatment of a specific condition, scales measuring treatment complications are usually condition specific [19]. Different condition-specific scales for the same condition are often available. In this situation, several factors may influence the choice of an outcome rating scale, resulting in heterogeneity of outcome measurement across clinical trials. These diverse outcome measures cannot be compared readily across studies of the same intervention except by using statistical methods such as transforming scores to standardized effect sizes such as Cohen's d, which estimate the magnitude of the effect using a numeric value for each study (see below) [20]. A Cohen's d of 0.2 is considered a small effect size; 0.5, a medium effect; and 0.8, a large effect size [20]. The development of a core outcome set for a specific clinical condition has also been suggested as a way of addressing heterogeneity in outcome measures. These sets are recommended as the minimum outcomes to be measured in clinical trials of the condition [21]. However, outcomes need not necessarily be restricted to those in the core outcome set. Although core outcomes will allow the results of different trials to be compared or combined, the hope is that researchers will continue to study additional outcomes. For example, a core set of measures has been proposed for adult and juvenile idiopathic inflammatory myopathies [22]. This core set consists of measures in the domains of global activity, muscle strength, physical function, laboratory assessment, and extramuscular disease [22]. This core set measure has been used in clinical trials to enroll patients (inclusion criteria) and for outcomes assessment [23]. The Core Outcome Measures in Effectiveness Trials (COMET) initiative, launched in 2010, brings together researchers interested in the development of core outcome sets (http://www.comet-initiative.org/). The COMET database is an excellent repository of available core outcome sets by specialty and disease (http://www.comet-initiative.org/studies/search).

Most studies utilize more than 1 outcome measure. Statistically, this results in multiple comparisons between groups. The *p*-value provides an estimate of the probability that the observed difference between 2 groups is due to chance (given the assumption that there is no difference between groups, i.e., the null hypothesis). If this probability is high (conventionally set at > 0.05, the significance level, α), we conclude that the observed difference is due to chance. If this probability is low (≤ 0.05), the observed difference is unlikely due to chance. A p-value of 0.05 implies a 5 % probability that findings are due to chance, yielding a false-positive result. Thus, the type I (false-positive) error rate is equal to α, the significance level. The likelihood of finding a positive result with a p-value of ≤ 0.05 by chance alone increases as the number of comparisons increase [24, 25]. This means that if several statistical comparisons are performed, we may obtain a statistically significant result with $p \leq 0.05$ by chance alone. Therefore, when multiple comparisons are performed, the p-value that reflects a significant result should be decreased proportionately. There are several formulae for applying such corrections. The Bonferroni correction is one such method: the conventional p-value of 0.05 is divided by the number of outcomes to obtain the "cut-off" p-value. If there are 5 outcomes, the p-value for significance is 0.05/5 or 0.01 [24]. However, the sample size needed to detect a difference between groups at $p < 0.01$ is often considerably larger than for $p < 0.05$. Therefore, specifying a primary outcome measure apriori is important. For the primary outcome measure, a standard significance level of $p < 0.05$ is used [25]. When interpreting secondary outcomes, it is important to bear in mind that a "statistically significant" p-value may not be truly significant unless there is an explicit statement regarding correction for multiple comparisons.

The use of confidence intervals (CIs) around the point estimate is also useful to assess the statistical significance of an outcome [26, 27]. The point estimate is the estimate of the measure of effect, often expressed as risk difference, relative risk (RR), or odds ratios. This estimate is based on a study of a sample of subjects from the population, and it will vary if the study were to be repeated with different samples from the population. The CIs provide a range of values within which the estimate of the true effect would fall were we to repeat the study several times. A 95 % CI is often chosen; this is simply the range that includes the true point estimate 95 % of the time. It provides information regarding the precision of the study to exclude an effect and also about the direction of the effect, which a p-value does not provide [27]. If the CI does not include the estimate for no effect (0 in the case of risk difference, 1 in the case of RR or odds ratios), it can be assumed that there is a statistically significant effect [26]. For example, consider the results of 2 hypothetical studies showing the same RR of 2. The RR is the ratio of the probability of the outcome in the intervention (or exposed) group compared with that in the nonintervention (or nonexposed) group. The first hypothetical study has a 95 % CI of 1.5 to 2.5. This provides several pieces of information. First, the point estimate (the RR of 2) indicates a favorable effect of treatment. Second, the CI is narrow, indicating precision of the point estimate. Finally, the CI does not span unity, indicating that the beneficial effect is statistically significant. Consider the second hypothetical study with the same RR of 2, but with 95 % CI 0.8 to 6. Although the RR of 2 indicates a beneficial effect of treatment, the CI spans unity, indicating that the result is not statistically significant. In addition, the wide CI indicates imprecision in the point estimate, and therefore cannot exclude potentially important clinical effects on both sides (both a beneficial effect and a harmful effect of the therapy). Finally, consider a study with RR 0.98 with a narrow CI of 0.97 to 0.99. This is statistically significant, as the CI does not cross unity. However, the clinical effect is likely modest because it is so close to unity.

Methods for Measurement of Functional Outcomes: Objective *versus* Subjective, Masked *versus* Unmasked: Why Does it Matter?

It has been said that "anything we measure is an abstraction of reality; anything we measure is measured with error" [2]. Bias or systematic error of a clinical trial is a study's tendency to measure the intervention's effect on the outcome incorrectly (Fig. 1). The problem lies in the study design. Statistical methods cannot correct systematic errors. The method of outcome measurement is one factor that determines the risk of bias of a study. Other factors include randomization, prospective or retrospective data collection, and so on [1]. The relationship of outcome measurement to the assessor's knowledge of the subject's treatment status is an important determinant of the risk of bias of the study. Objective outcomes are those that are unlikely to be affected by observer expectancy bias (the assessor's bias causing them to subconsciously influence subjects). Survival is an objective outcome; as also results of a laboratory test. If the outcome measure is not objective, the assessment of outcome should be masked or blinded (the assessor is not aware of the treatment allocation of the subject) in order to reduce observer expectancy bias. If blinding is not possible (usually in trials of surgical procedures where there may be obvious clues to the type of intervention), having an independent outcome assessor, that is, an outcome assessor who is other than the treating clinician, can also be used to reduce the risk of bias. Trials where the outcome measure is objective or double blinded have the lowest risk of bias [1]. Although a laboratory assay is usually considered an objective outcome measure, it may be limited because it is indirect. Indirect or surrogate outcomes, focusing on biological or physiological factors related to the disease of interest, (e.g., results of imaging or laboratory testing) are not uncommonly used in clinical trials as a representative measure of clinical outcomes. The disadvantage of surrogate measures is that they may not predict clinically important outcomes [1].

Several measurement issues influence the utility of rating scales. The information source may be the clinician, patient, or caregiver. The information collected may be a report, that is, addressing how specific tasks are done, or may be obtained by direct patient assessment. An example of the former is ALSFRS-R where items assess patient reports of their ability to carry out specific tasks [12]. The mini-mental status examination is an example of direct patient assessment [28]. Rating scales using reports may use patient reports or proxy reports, for example family or other caregiver information. Because proxy reports and direct patient reports may not be similar, consistency for all subjects and across serial measurements is important. Because the mode of administration (face-to face, telephone, web-based) can provide different results, this should also be consistent across subjects and across repeated measurements [2]. Hence the choice and method of measurement of the outcome is important not only to ascertain that we are measuring clinically relevant outcomes, but also to make sure that we are measuring the relationship between the intervention and outcome accurately. Finally, practical considerations that influence the choice of a rating scale include ease of use, time taken to administer the scale, the need for special training to administer it, the mode of administration (paper, web-based, telephone, etc.), and costs.

The Ability of a Rating Scale to Measure a Clinically Important Difference

The factors that influence the ability of a rating scale to measure a clinically important difference are its validity, reliability, and responsiveness.

Validity

Validity is defined as the extent to which the scale measures what it is intended to measure. It has been stated that "validating a scale is a process by which we determine the degree of confidence we can place on inferences we make about people based on their scores from that scale" [29]. Although the importance of validity is obvious, it is not easy to establish. There are several types of validity; however, 3 main types are used to evaluate rating scales: criterion validity, content validity, and construct validity. Often, they boil down to face validity as the heart of this operation.

Criterion validity refers to the correlation of the scale with a gold standard or a previously established measure of the domain of interest (the criterion). The difficulty with establishing criterion validity is the requirement for a gold standard. Criterion validity may be concurrent or predictive. In concurrent validity, the criterion and the scale are assessed and correlated at the same time [2]; a correlation between the scale being validated and an established rating scale, both administered to the same group of subjects on consecutive days is an example of concurrent validity. In predictive validity the scale is ratified against a criterion that is collected at a future time point [2]. A correlation between high-school students' standardized testing scores and future admission to an Ivy League school, for instance, is predictive validity.

Content validity measures the extent to which the items in the scale comprehensively cover the concepts of interest [2]. The concepts, or domains, of a rating scale may be expressed in terms of symptoms, functioning (physical, psychological, and social), or overall health perceptions or quality of life. Physical functioning may be evaluated as capacity (what patients think they can do) or performance (what they can actually do) [30]. Each of these domains should be measured by separate subscales [31]. Content validity is difficult to implement because it is often not possible to sample the entire domain as the concepts cannot be clearly delineated [31]. Often, content validity boils down to face validity, a judgment call. An important aspect of content validity is the interpretability of the scale, especially in the case of PROs or observer-rated outcomes. It has been recommended that rating scales be developed such that they should not require reading skills beyond that of a 12 year old [29].

Construct validity tests whether a scale measures the intended construct [2]. A construct is an abstract principle that conceptualizes the latent (unobservable) variable that is being measured. Health-related quality of life (HRQoL) is a latent variable; in SF-36, a frequently used scale to measure HRQoL, 8 constructs are used to define this variable: physical activity, social activity, bodily pain, mental well-being, usual role activities, fatigue, and perceptions of general health. These 8 constructs are measured by 36 items [16]. However, as the latent variable cannot be observed directly, and has neither criterion (gold standard) nor content (measurable domain), construct validity has to be tested indirectly. One method of testing construct validity is to administer the rating scale to 2 groups of subjects, one of which is predicted to have the construct or variable being measured, and another, not to have the variable. If the scale has construct validity, the group with the construct will be expected to score higher than the group without. However, in practice, scales should be able to discriminate in the middle ranges of the characteristic, not just the extremes [29]. Another way to assess construct validity is to use correlation. If 2 scales are expected to measure the same construct, they should be highly correlated (convergent validity). If they are measuring different latent variables, they should not be correlated (discriminant validity) [2]. Special statistical methods such as factor analysis can also be used to determine if the items on a rating scale correspond to an underlying construct [2].

Reliability

Reliability is the reproducibility, repeatability, or stability of a scale. A scale with high reliability gives similar results if administered under consistent conditions assuming that the underlying variable it measures has not changed. Reliability is important because it is determines the ability of a scale to detect a true difference between measurements. Reliability is also a prerequisite for validity because unless the measure is consistent, it is not possible to be certain that the scores accurately measure the variable of interest. A measure cannot be more valid than it is reliable. The reliability of a scale can be measured by the degree of random error associated with its measurements [2, 32]. There are several types of reliability, each addressing a different source of random error. Most clinicians are familiar with inter-rater and intrarater reliability, both being forms of test–retest reliability. An additional form of reliability in multiple outcome scales is internal consistency.

Reliability can be expressed as [33]:

$$\frac{\text{True variance}}{\text{True variance} + \text{error variance}}$$

Therefore, the greater the variance due to measurement error, the lower is the reliability.

Test–Retest Reliability (Intrarater and Inter-rater)

Intrarater reliability evaluates the reproducibility of a scale when administered to the same individuals after a period of time. The correlation between the scores of each trial provides a measure of intrarater or test–retest reliability. The interval between test and retest is an important consideration when determining intrarater reliability. If the interval is too short, memory effects may falsely increase reliability. If it is too long, there may be a change in the underlying variable, which will change the scores, with apparent poor reliability, when the real issue is that the scale has responded to a true change in the underlying variable [2]. For performance-based measures, such as TUG [7], or other CROs which require an external rater, both inter- and intrarater reliability are important. Intrarater reliability indicates how consistently a rater administers and scores an outcome measure, whereas inter-rater reliability indicates how well 2 raters agree in the way they administer and score an outcome measure. Thus inter-rater reliability measures the extent of concordance among raters. This is important when multiple raters are administering a scale, as in CROs that are measured by investigators in a multicenter trial. Unless there is concordance between raters, that is, 2 raters are in agreement when the scale is applied to the same subject under similar conditions, comparisons between scores obtained by the two raters will not be valid. Inter-rater reliability thus provides a measure of the homogeneity between raters.

Intraclass correlation coefficients (ICCs) such as the reliability coefficient provide an assessment of reliability [34]. The ICC provides information on the degree to which repetition of the same test gives the same results under the same conditions in the same subjects [34]. A reliability coefficient of 1 indicates perfect reliability without measurement error, and 0 indicates no reliability. It is to be noted, however, that ICC is calculated for continuous data, not ordinal data. The kappa (κ) statistic (Cohen's κ) is commonly utilized to assess inter-rater agreement in scales with nominal data. Some agreement can occur by chance alone, and the kappa statistic indicates the agreement beyond that expected by chance. The usual interpretation of the kappa statistic is as follows: < 0, less than chance agreement; 0.01 to 0.2, slight agreement; 0.21 to 0.4, fair agreement; 0.41 to 0.6, moderate agreement; 0.61 to 0.8, substantial agreement; > 0.81, near-perfect agreement [35]. To evaluate inter-rater reliability of nominal scales across more than 2 raters, Fleiss' Kappa is used [36]. Ordinal measures may be tested for reliability using weighted kappa. Weighted kappa emphasizes disagreements; in an ordinal scale with responses of "not at all satisfied", "slightly satisfied", "neutral", "very satisfied", and "extremely satisfied", we may be more interested in the extreme responses of "not at all satisfied" and "very satisfied" than in the minor differences between "slightly satisfied" and "neutral". A weighted kappa assigns larger weights to larger disagreements between ratings using quadratic methods [37]. With a weighted kappa, disagreement of "slightly satisfied" and "neutral" will count as partial agreement, but a disagreement of "not at all satisfied" and "very satisfied" would be considered no agreement [38].

Internal Reliability or Internal Consistency

Internal reliability or consistency measures the extent to which items in a scale or subscale are correlated, and therefore measuring the same construct (i.e., are homogeneous). It measures how closely each item in the scale is related to the overall scale [2, 31]. The measure of internal consistency is referred to as Cronbach's α, which is a type of ICC. Alpha is close to 1 when the items of the scale are highly correlated. However, this also implies that items in the scale may be redundant because they are measuring the same construct. Additionally, α can increase falsely with increasing items on a scale, reflecting redundant items and not necessarily implying high internal consistency [2]. It is suggested that an ideal α is 0.7 to 0.8 [39]. Another method to assess internal consistency of items in a scale is to measure item to total score correlation.

If the item score does not correlate with the total score minus the item, it is not an optimal item [2].

Responsiveness

The third factor that influences the accuracy of functional rating scales is their ability to detect clinically important changes over time [31, 40]. Responsiveness is also referred to as a measure of longitudinal validity or sensitivity to change. When functional outcome scales are utilized to evaluate differences between 2 groups of subjects at one time point, reliability and validity are sufficient. However, for measures that are designed to evaluate changes over time, responsiveness is also a requirement [34, 40]. An example of a nonresponsive item is "I have attempted suicide" from the Sickness Impact Profile, a general measure of health status. A person who responds yes at the first administration of the scale would logically answer yes in all subsequent administrations. Therefore, this item is not responsive to changes in emotional function, or suicide risk [17, 40].

Responsiveness should be assessed by testing predefined hypotheses regarding the expected differences over time in 2 groups. However, it is important to first clarify the definition of responsiveness: are we evaluating the scale's ability to detect changes over time in general, detect clinically important changes over time, or detect changes in the true value of the construct? These are related, but not identical, concepts. A general change is any change, regardless of whether it is clinically relevant. Often, this is a statistically significant difference over time [34]. This is frequently equated to the concept of sensitivity to change. Detection of a clinically meaningful change implies an explicit judgment to define a clinically meaningful change. Finally, a change in the true value of the construct is a further refinement of the previous definitions where it not only requires a judgment of what is an important change, but also a gold standard for the variable being measured; this is not easy for latent variables [41]. Responsiveness can be expressed mathematically as [2, 29]:

$$\text{Responsiveness} = \frac{\text{Variance due to change}}{\text{Variance due to change} + \text{error variance}}$$

Methods for measuring responsiveness are complex. It is possible to compute statistically the smallest detectable change of an outcome measure. The smallest detectable change refers to the smallest within-subject change in score that can be interpreted as a real change with $p < 0.05$. Various methods have been used to quantify responsiveness such as the effect size statistic (e.g., Cohen's d) and standardized response mean. The general concept is to divide the difference in means between measure points (e.g., baseline and postintervention) by a measure of variance (e.g., pooled SD for Cohen's). However, because these estimates are also linked to the magnitude of change in response to the intervention, they are limited measures of responsiveness [42]. Other methods of quantifying responsiveness are available; these are beyond the scope of this review [43, 44].

The *floor* and *ceiling effects* of an outcome measure also influence its responsiveness. Floor or ceiling effects are said to be present if more than 15 % of respondents achieve the highest or lowest possible score, respectively [45]. If many respondents initially achieve the highest or lowest score possible, a change cannot be detected over time. Floor or ceiling effects also influence reliability, because the subjects with the lowest or highest scores cannot be differentiated from each other. It is also possible that there is limited content validity at the extreme ends of the scale [31].

Interpretability of Rating Scales

The interpretability of an outcome measure is defined as the extent to which one can assign qualitative meaning to the quantitative scores obtained by the measure [32].

There is a tendency for clinicians to sometimes focus exclusively on the *p*-value to interpret the results of a study. Clinical outcomes should be interpreted in a clinical context. The oft-repeated phrase "statistical significance is not the same as clinical significance" encapsulates this point. Statistical significance is a mathematical parameter which does not shed much light on the question of clinical importance. Whether or not an observed treatment effect is clinically importance requires clinical judgment. It is important also to recognize that statistical significance and clinical importance may be discordant. Statistical significance may be achieved by a large enough sample size, which reduces random error and provides the precision to detect small changes in outcomes. Therefore, changes in functional outcome measures must be interpreted in the context of the minimal clinically important difference (MCID). The MCID should be established a priori in clinical trials. This is usually a somewhat subjective judgment call. MCID may be defined as "the smallest difference in a score of a domain of interest that patients perceive to be beneficial and that would mandate, in the absence of troublesome side effects and excessive costs, a change in the patient's management" [46]. The term minimally important difference has been used to focus on relevant changes in patient experiences. Minimally important difference is defined as "the smallest difference in score in the outcome of interest that informed patients or informed proxies perceive as important, either beneficial or harmful, and which would lead the patient or clinician to consider a change in the management" [47, 48].

Interpretability of functional measures addresses the question: what changes in scores correspond to a large, moderate, or small benefit? Does a change of 10 points on a quality-of-life scale mean that the subject is less distressed and experiencing higher levels of well-being? Mean changes in scores are less informative for this determination than the proportion of subjects who achieve a good outcome as measured by a prespecified degree of change in outcome scores. Additionally, changes in outcomes between groups of patients (treatment *vs* control) are different from within-patient changes over time (pre- and postintervention). Hence, population MCID and individual MCID are not equivalent [49–52]. MCID are commonly used for sample size estimations.

MCIDs may be estimated using several methods; these methods generally fall into 2 categories. The first is the anchor-based method, wherein the scores on the measure of interest (target) is calibrated with another independent measure of clinical change, the anchor [50]. Given a range of anchor scores that corresponds to the MCID, a target score corresponding to that value is calculated. Alternatively, receiver operating characteristic (ROC) curves can be used. Each respondent is classified according to the anchor measure as experiencing a change or not. Specificity and sensitivity of the target measure scores are calculated based on the anchor, and ROC curves constructed. The optimal cut point on the target measure is the one with the least number of misclassifications compared with the anchor measure. Misclassifications are false-positives (respondents mistakenly classified as experiencing a change) and false-negatives (respondents mistakenly classified as not experiencing a change) [50, 53]. Changes in scores on instruments measuring dyspnea, fatigue, and emotional function in patients with chronic heart and lung disease calibrated against patient global ratings of change have found that a mean score change of 0.5 on a 7-point Likert scale corresponds to the MCID [46]. In order to apply anchor-based methods, an interpretable anchor must be available, and there must be a clear association between the anchor and target [50]. Multiple anchor methods compare a target with several anchors; these methods tend to provide more information when trying to interpret complex latent variables such as quality of life, but are more difficult to interpret. Single-anchor methods, in addition to establishing the MCID, also provide differences in scores on the target instrument that establish a “cut-off” for small, moderate, and large changes in the outcome variable [50].

The second method of assessing clinical change uses distribution-based statistics, where the score is interpreted in terms of the relationship between the magnitude of effect and a measure of variability, such as standard deviation (SD). These methods most commonly use effect sizes such as Cohen’s d or standard mean differences, discussed above [20]. If the means of both groups are equal, the effect size is 0. An effect size of 0.5 indicates that the difference between two compared groups is 0.5 SD. In general, 0.5 SD of a continuous outcome measure is a clinically meaningful difference [25]. The reader is referred to Brozek et al. [49] and Crosby et al. [54] for a discussion of other methods. Distribution-based methods are less labor-intensive than anchor-based methods. Their main drawbacks are that measures of variability such as SD change between studies, and there is no innate meaning to the numerical effect size [49].

Methodological Aspects of Rating Scales: Ordinal Scales and Measurements

The assumption when using rating scales is that they provide a method of measuring, quantifying, and therefore comparing outcomes in a numerical fashion. Measurement is a process of quantifying the magnitude of anything in comparison to a standard. For instance, a weight of 5 lbs indicates a standard of 1 lb against which the object is compared, and expressed as a number, 5, times the unit, 1 lb. Therefore, a fundamental requirement of measurements is to define the standard of comparison and unit [55]. In this sense, measurement involves numbers that can be used in calculations. The numbers maintain their values and relationships with each other after mathematical operations such as multiplication or division.

Rating scales map out the underlying construct being measured as a line (continuum) varying from lower to higher, where people can be located. A person’s total score determines his or her location on the continuum. This process is referred to as “scaling” [56]. However, ordinal rating scales do not provide measurements in the scientific sense of the word. They assign numbers to rank-ordered items, that is, they rank subjects in an order from high to low, or vice versa, on the variable of interest. According to classical measurement theory (see next section), these are not measurements in the scientific sense of the word because there is no standard unit of reference [57]. The ranks are not calibrated against a standard, because there is no objective “reference” standard. Hence, by definition, ordinal scales are subjective measurements [57]. The Likert scale, an ordinal scale familiar to most people, consists of a statement relevant to the topic to which there are up to 5 to 7 response choices to rate the degree to which respondents agree or disagree with the statement. These choices are scored numerically on an ordinal scale. Although the responses are ranked by numbers, the difference between categories of responses is not measurable and may not be equal. Therefore, the difference between the ranks for “always”, “often”, and “sometimes” is not necessarily the same. The only information provided by ranking is that one value is greater or less than the other. However, by assigning sequential numbers to each category, the presumption is that the intervals between them are equal. This issue also comes into play with multiple-item rating scales. Because the distance between ranks is not equal, sum scores cannot be computed accurately. However, numerical ratings for each item are frequently summed to provide a raw score. These summed scores are usually treated as interval, linear data and mean scores are computed, assuming that the intervals between rankings are equal. Thus, this process converts a lower, ordinal level of measurement to a higher, interval level of measurement. There is controversy whether this is appropriate; some authorities support the view that summed scores are not measurements because of the lack of a unit [3, 58–60], whereas others believe that ordinal scores are forms of measurement, although less robust [59] and that they approximate interval data [61]. Ordinal scales make another assumption: that the distance between the summed scores is consistent across the range of the scale, which may not be accurate. For instance, a change in 5 points in the lower end of a scale may not have the same implication as a change in the same number of points across the mid- or higher end of the scale [56].

Another limitation of ordinal scales is the subjectivity introduced by the respondent's preferences and values. For instance, a question regarding quality of life may mean different things to different subjects based on what they consider adequate or good, and 2 responses may be different even if the underlying construct being measured is the same. Finally, ordinal scales, although useful to make comparisons between groups, cannot be used reliably to compare changes of individual patient scores longitudinally because of their lack of precision (test–retest reliability). This is because the random error for an individual can be substantial, resulting in wide CIs around the scores [45].

It is not uncommon in clinical trials to see dichotomized ordinal scales, where a cut-off point is established between ranges of scores. For instance, the modified Rankin scale is often dichotomized into 2 clinical categories with arbitrary thresholds of not disabled (0–2) and disabled (3–5). Although this provides ease of clinical interpretation, dichotomizing scores results in loss of data on the more subtle categories of disability [3]. Because each grade of the scale reflects fairly distinct functional levels, a change of even 1 grade in either direction may be meaningful clinically. This distinction is lost when the scores are dichotomized [3, 62].

Interval outcome scales have continuous, linear data. Unlike ordinal scales, the distance between successive categories is known and assumed to be equal in interval scales. An example of such interval data is the Fahrenheit scale, where the interval between 50 and 55 degrees is the same as that between 90 and 95 degrees. However, interval scales do not have a true zero. Therefore, it is not appropriate to say that 80 degrees is twice as hot as 40 degrees [57, 59, 61]. Ratio scales are interval measures with a true zero. Height and weight, where there is a true zero, or lack of measure of the variable, can be measured by ratio scales. The ratio between 5 feet and 4 feet stays the same, even when converted to inches. The visual analog scale is considered by some authorities to represent an interval or even ratio scale, whereas others argue that it is nonlinear [63, 64].

The recommended statistical methods for analyzing results from rating scales are a matter of some deliberation. It has been suggested that parametric statistical methods (e.g., *t* test, that assume normal distribution of the sample, and that data are derived from an interval scale) be used only with interval or ratio data and nonparametric tests (e.g., Wilcoxon rank-sum test) be used for ordinal data, although this is also subject to some controversy. The argument for using parametric statistics for ordinal data, which is not an uncommon practice, is that the sum scores approximate interval data and parametric statistics can handle the shortcomings of ordinal scales [3, 65].

Measurement Theories: Classical Test Theory and Latent Trait Theories

A discussion of functional rating scales is incomplete without a brief discussion of measurement theories. Measurement theories are mathematical models of the factors that affect the scores generated by rating scales. They can be used to develop rating scales and also to evaluate existing rating scales. They explain the relationship of the scale scores to the latent construct, and evaluate if the quantitative conceptualization of the latent construct has been successfully operationalized [3]. Traditional psychometric methods use the classical test theory (CTT) [3, 66]. This theory evaluates rating scales in terms of their reliability, validity, and responsiveness. It assumes that the observed score (O) is the true score (T) plus measurement error (E), that is, O = T + E [67]. However, the values of T and E cannot be computed, and only the observed score is obtained and analyzed. The theory itself cannot be tested because the true score and measurement error cannot be measured separately, unless some assumptions are made to define error scores [3, 68]. The true score can then be calculated as T = O – E. Additionally, the construct being measured is independent of the rating scale used; however, both the true and observed scores are rating scale dependent, that is, a respondent may have unequal scores for disability on 2 different rating scales, although they both measure the same underlying construct [69]. Conversely, features of the rating scale itself, such as reliability, are respondent dependent because they are assessed using responses from subjects [68]. The other limitation of CTT is the conversion of ordinal scores to interval-like summed scores, and the inherent issues with such an approach, discussed in the previous section.

In order to overcome the limitations of CTT and to transform ordinal scales to interval measurements, new psychometric methods, termed latent trait theory (LTT) or modern test theory have been proposed. The 2 major types of LTT are item response theory (IRT) [70] and Rasch analysis [71]. These are also mathematical models and provide statistical methods to analyze rating scale scores and evaluate the reliability and validity of the scales. These methods, like CTT, also aim to measure the relationship between the true score and the observed score, but they do so by addressing the relationship between a person's measurement on the underlying latent construct and the probability of selecting one of the response options of an item on the rating scale [56]. These methods evaluate the person's location on an interval-level continuum or "line" rather than the total score which is ordinal. This "person location" on the continuum determines the person's response to an item. This means that the response to an item is determined by the patient's level of ability related to that item. A person who is severely disabled will logically be more likely to respond "no" to a question regarding running. By the same token, a person's response to an item is related to the level of difficulty of the item. A person who is moderately disabled is more likely to answer "no" to playing sports than to walking. The theories use a probability of response to indicate that a true prediction is not possible. Finally, in these theories, the focus is no longer the total score; instead, individual item scores are evaluated [56].

The fundamental difference between IRT and the Rasch model is the way they approach the data. IRT prioritizes the data, and if a model does not fit the data, another model that best explains the data is used. Rasch measurement on the other hand, prioritizes the model, and if the data do not fit the model, an exploratory approach is used to explain why the data do not fit the model (Fig. 2) [3, 56]. The advantages of these models over CTT models are the ability to obtain interval measurements from ordinal scales, to obtain measurements at the person level, as well as group comparisons, and to use subsets of items from a scale as needed [56]. Subsets of ordinal scales cannot be used separately without compromising reliability and validity. The disadvantage is that they are complex, require some training and are usually performed using special software programs. Over the last several years, several preexisting neurologic rating scales have been evaluated and revised using LTT models [72–75].

PRO Measures: Special Considerations in Development and Interpretation

PROs are becoming increasingly popular in clinical trials because of the recognition that CROs may not measure outcomes that are directly relevant to patients. In clinical practice, we routinely use patient reports to evaluate the effectiveness of our interventions; clinical trials should, ideally, be no different.

Historical Aspects of PROs

The impetus for using PROs has its basis in the Health Insurance Experiment. The Health Insurance Experiment is the only long-term study of the effects of cost-sharing on use of medical services, healthcare quality, and outcomes. It was supported by the Department of Health and Human Services and conducted by the RAND Corporation between 1972 and 1984. The study randomized 3958 participants to 1 of 14 health-insurance models, providing either free care or 3 different types of cost sharing. A validated medical history questionnaire captured participants' outcomes with respect to physical health, role functioning, and health perceptions, an early use of a PRO in a clinical trial [76]. Subsequently, the Medical Outcomes Study evaluated practical measures for monitoring patient outcomes in clinical practice [77]. Over 70 measures of well-being and functioning were developed and validated for this study, setting the stage for the use of these measures in future clinical trials [78]. One of the best known PRO measures developed from the Medical Outcomes Study is SF-36 (http://www.rand.org/health/surveys_tools/mos/36-item-short-form.html). As the pharmaceutical research community recognized the value of PROs including quality of life outcome measures, the FDA, recognizing the need for stringent development and application of PROs, began the process of setting standards for PROs used in drug trials [79].

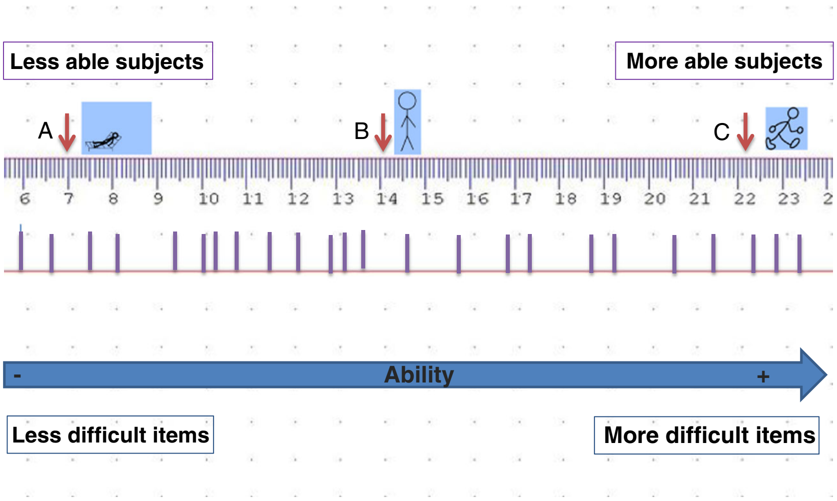

Fig. 2 Rasch model interval "ruler". The Rasch model conceptualizes a measurement scale where subjects are ranked by their ability from low to high, and items on a rating scale are ordered by the level of difficulty from easiest to most difficult. The thick lines on the ruler indicate the location of items of increasing difficulty from left to right. (A), (B), and (C) represent 3 subjects with increasing levels of physical ability

Development of PROs

Development of PROs is reviewed in depth by the FDA and in other reviews [79, 80]. Briefly, specific issues pertaining to PROs include the need to obtain patient input in the development phase using interviews, surveys, or focus groups. The questions should be presented in a format that is easily understood. Once developed, preliminary PRO instruments undergo cognitive testing, wherein participants similar to those in whom the surveys are intended to be used complete the survey and then answer questions about their interpretation of the questions and how they selected their responses. Preliminary surveys are modified as necessary based on the results of cognitive testing This helps to ensure the face validity of the instruments [79, 80].

PROs and Medical Product Labeling

Despite regulatory encouragement for the use of PROs in clinical trials, drug approval based on PROs remains challenging. A recent study found that only 3/40 (7.5 %) of approvals by the Office of Hematology and Oncology Products of the FDA received PRO-related labeling between 2010 and 2014 [81]. Some factors underlying this may include challenges in study design using PROs and the quality of PRO instruments used.

PROs: What's Next?

In order to make valid, reliable PRO instruments easily available to researchers, the Patient-Reported Outcome Measurement Information System, (PROMIS) was launched in 2005 through a National Institutes of Health Roadmap Initiative (http://www.healthmeasures.net/explore-measurement-systems/promis). Using rigorous methods, PROMIS has developed item banks (a collection of items that assess a specific trait, e.g., pain). These instruments are available to researchers free of cost. NeuroQol, also available for free, is a series of PRO instruments for pediatric and adult neurological conditions (http://www.healthmeasures.net/explore-measurement-systems/neuro-qol). Both PROMIS and NeuroQol can be applied as computerized tests or on paper and have been translated into languages other than English. However, these PROs are likely underutilized at the present time. A MEDLINE search via PubMed for "NeuroQol" on 29 September 2016 revealed only 6 studies using NeuroQol PRO measures between 2003 and 2016.

Clinical Implications

In the world of clinical research, the choice of an appropriate functional rating scale is an important aspect of trial design because the results of a clinical trial are only as accurate as the rating scales employed to measure the outcomes of interest. This has implications for patient care and subsequent research. Therefore, the quality of existing rating scales should be carefully assessed before selection for a trial, and new rating scales should be rigorously validated. Medical researchers may not be familiar with clinimetric methods, but should realize the importance of obtaining the necessary expertise at the stage of study design. An excellent review with easily understandable examples provides researchers with a basic familiarity with clinimetry [56].

With reference to the increasing use of PROs in clinical trials, the FDA recommends an endpoint model for their use, wherein the use of the PRO as a primary, secondary, or exploratory endpoint is prespecified [79]. This helps to interpret the results of the PRO measure accurately for statistical and clinical significance. The FDA requirements for PROs are stringent and recommend developing a new PRO or modifying an existing PRO to ensure that the PRO is appropriate to measure the concept of interest [79]. Although laudable, this is not always practical. The European Medicines Agency also recommends an endpoint model for measures of HRQoL [82], and has recently provided guidelines on the use of PROs for outcomes other than HRQoL in oncology studies [83].

What is the role of outcome measurement in the realm of clinical practice? As far back as 1988, Ellwood [84] suggested the concept of "outcome management", defined as the use of patient experiences to inform medical decision-making by clinicians, payers, and patients [84]. Ellwood draws parallels between good clinical practice and a clinical trial, envisioning clinical practice in 3 steps: selecting the appropriate intervention, routine measurement of patient reported and clinical outcomes, and evaluating data to inform decision making [84]. In today's world of electronic health records, tracking outcome measures in clinical practice is feasible and relatively simple, and PROs are an important component of this process. Real-world data regarding natural history of disease, effectiveness of treatments (in contrast to efficacy in the rigorously controlled clinical trial setting), long-term adverse events and evolution of patient preferences are some of the data that can be obtained by incorporating these outcome measures in practice.

As part of quality-improvement initiatives that form a major aspect of healthcare reform in the USA, performance measures are collected and compared to benchmarks; these often include clinical outcomes such as falls, readmission rates, or laboratory testing (e.g., HbA1c in diabetics). PROs are now being increasingly recognized for their role in performance measurement. A recent review suggests best-practice methods to assess quality of care using PROs [85]. A word of caution, however: patient reports of quality of life, or other domains such as role functioning may not directly measure the results of the intervention and may be affected by unrelated patient factors; hence, accurate measurement of performance solely on the basis of PROs is challenging. Additionally, the collection of multiple outcome measures in the setting of a clinic visit is limited by time constraints; there are already several burdensome aspects of reporting in clinical practice, driven by legislative and regulatory mandates.

Conclusions

Whether used in research trials or as performance measures in clinical practice, delineating CROs and PROs that are effective and efficient is a challenge, especially when multiple outcome measures for a specific domain or disorder are available. Research efforts should be invested to validate existing measures and determine the best performing ones, or to develop measures when efficient and effective measures are lacking. The PROMIS and NeuroQoL databases are excellent, freely available resources for PRO instruments. Use of the appropriately selected PRO instruments from these databases has the advantages of ensuring PRO instrument quality and being able to compare outcomes across studies and across clinical care settings. However, PROs that measure focused effects of an intervention rather than broad concepts of health need to be developed to capture disease- and intervention-specific effects.

Acknowledgments I thank Gary Gronseth, MD, Professor of Neurology, Kansas University Medical Center, KS, USA, for suggestions on the content and structure of the manuscript, and Elizabeth Raynor, MD, Associate Professor of Neurology, Harvard Medical School/ Beth Israel Deaconess Medical Center, for reviewing the manuscript and providing suggestions.

Required Author Forms Disclosure forms provided by the authors are available with the online version of this article.

References

1. American Academy of Neurology. Clinical Practice Guideline Process Manual. In: American Academy of Neurology, editor. 2011 Edition ed. St. Paul, MN, 2011.
2. Frytak JR, Kane RL. Measurement. In: Kane RL, editor. Understanding Health Care Outcomes Reasearch, Second Edition. Sudbury, MA: Jones and Bartlett Publishers; 2006. pp. 83-120.
3. Hobart JC, Cano SJ, Zajicek JP, Thompson AJ. Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. Lancet Neurol 2007;6:1094-1105.
4. Hobart JC, Riazi A, Lamping DL, Fitzpatrick R, Thompson AJ. How responsive is the Multiple Sclerosis Impact Scale (MSIS-29)? A comparison with some other self report scales. J Neurol Neurosurg Psychiatry 2005;76:1539-1543.
5. de Vet HC, Terwee CB, Bouter LM. Current challenges in clinimetrics. J Clin Epidemiol 2003;56:1137-1141.
6. Feinstein AR. An additional basic science for clinical medicine: IV. The development of clinimetrics. Ann Intern Med 1983;99:843-848.
7. 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.
8. Bohannon RW, Smith MB. Interrater reliability of a modified Ashworth scale of muscle spasticity. Phys Ther 1987;67:206-207.
9. Guy WE. ECDEU Assessment Manual for Psychopharmacology. In: NIMH Psychopharmacology Research Branch DoERP, editor. Rockville, MD, 1976. p. 218-222.
10. Patrick DL, Burke LB, Powers JH, et al. Patient-reported outcomes to support medical product labeling claims: FDA perspective. Value Health 2007;10(Suppl. 2):S125-S137.
11. Fahn S, Elton RL, Committee MotUD. Unified Parkinson's Disease Rating Scale. In: Fahn S, Marsden C, Calne DB, Goldstein M, editors. Recent Developments in Parkinson's Disease. 2. Florham Park, NJ: Macmillan Health Care Information; 1987. pp. 153-164.
12. Cedarbaum JM, Stambler N, Malta E, et al. The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function. BDNF ALS Study Group (Phase III). J Neurol Sci 1999;169:13-21.
13. Hoehn MM, Yahr MD. Parkinsonism: onset, progression and mortality. Neurology 1967;17:427-442.
14. Mahoney FI, Barthel DW. Functional evaluation: the Barthel Index. MD State Med J 1965;14:61-65.
15. Sloan JA, Aaronson N, Cappelleri JC, Fairclough DL, Varricchio C, Clinical Significance Consensus Meeting G. Assessing the clinical significance of single items relative to summated scores. Mayo Clin Proc 2002;77:479-487.
16. Ware JE, Jr., Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992;30:473-483.
17. Bergner M, Bobbitt RA, Kressel S, Pollard WE, Gilson BS, Morris JR. The sickness impact profile: conceptual formulation and methodology for the development of a health status measure. Int J Health Serv 1976;6:393-415.
18. Kurtzke JF. A new scale for evaluating disability in multiple sclerosis. Neurology 1955;5(8):580-583.
19. Atherly A. Condition-Specific Measures. In: Kane RL, editor. Understanding Health Care Outcomes, Second Edition. Sudbury, MA: Jones and Bartlett Publishers; 2006. pp. 165-183.
20. Cohen J. Statistical power analysis for the behavioral sciences. 2nd edition ed. Hillsdale, NJ: Lawrence Earlbaum Associates; 1988.
21. Williamson PR, Altman DG, Blazeby JM, et al. Developing core outcome sets for clinical trials: issues to consider. Trials 2012;13:132.
22. Miller FW, Rider LG, Chung YL, et al. Proposed preliminary core set measures for disease outcome assessment in adult and juvenile idiopathic inflammatory myopathies. Rheumatology (Oxford) 2001;40:1262-1273.
23. Oddis CV, Reed AM, Aggarwal R, et al. Rituximab in the treatment of refractory adult and juvenile dermatomyositis and adult polymyositis: a randomized, placebo-phase trial. Arthritis Rheum 2013;65:314-324.
24. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. BMJ 1995;310:170.
25. Zlowodzki M, Bhandari M. Outcome measures and implications for sample-size calculations. J Bone Joint Surg Am 2009;91(Suppl. 3):35-40.
26. Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about therapy or prevention. B. What were the results and will they help me in caring for my patients? Evidence-Based Medicine Working Group. JAMA 1994;271:59-63.
27. du Prel JB, Hommel G, Rohrig B, Blettner M. Confidence interval or p-value?: part 4 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2009;106:335-339.

28. Folstein MF, Folstein SE, McHugh PR. “Mini-mental state”. A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 1975;12:189-198.
29. Streiner D, Norman G. Health Measurement Scales: A Practical Guide to their Development and Use. 2nd ed. New York: Oxford University Press; 1995.
30. Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. JAMA 1995;273:59-65.
31. Terwee CB, Bot SD, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007;60:34-42.
32. Lohr KN, Aaronson NK, Alonso J, et al. Evaluating quality-of-life and health status instruments: development of scientific review criteria. Clin Ther 1996;18:979-992.
33. Carmines EG, Zeller RA. Reliability and Validity Assessment. Beverly Hills, CA: Sage Publications; 1979.
34. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chronic Dis 1987;40:171-178.
35. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-174.
36. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971;76:378-382.
37. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968;70:213-220.
38. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med 2005;37:360-363.
39. Nunnally JC, Bernstein IH. Psychometric Theory. New York: McGraw Hill; 1994.
40. Guyatt GH, Deyo RA, Charlson M, Levine MN, Mitchell A. Responsiveness and validity in health status measurement: a clarification. J Clin Epidemiol 1989;42:403-408.
41. Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PM. On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. Qual Life Res 2003;12:349-362.
42. O'Connor RJ, Cano SJ, Thompson AJ, Hobart JC. Exploring rating scale responsiveness: does the total score reflect the sum of its parts? Neurology 2004;62:1842-1844.
43. Cohen J. A power primer. Psychol Bull 1992;112:155-159.
44. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Med Care 1989;27(3 Suppl.):S178-S189.
45. McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? Qual Life Res 1995;4:293-307.
46. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials 1989;10:407-415.
47. Schunemann HJ, Puhan M, Goldstein R, Jaeschke R, Guyatt GH. Measurement properties and interpretability of the Chronic respiratory disease questionnaire (CRQ). COPD 2005;2:81-89.
48. Schunemann HJ, Guyatt GH. Commentary—goodbye M(C)ID! Hello MID, where do you come from? Health Serv Res 2005;40: 593-597.
49. Brozek JL, Guyatt GH, Schunemann HJ. How a well-grounded minimal important difference can enhance transparency of labelling claims and improve interpretation of a patient reported outcome measure. Health Qual Life Outcomes 2006;4:69.
50. Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR, Clinical Significance Consensus Meeting G. Methods to explain the clinical significance of health status measures. Mayo Clin Proc 2002;77:371-383.
51. Hemingway H, Stafford M, Stansfeld S, Shipley M, Marmot M. Is the SF-36 a valid measure of change in population health? Results from the Whitehall II Study. BMJ 1997;315:1273-1279.
52. Lydick E, Epstein RS. Interpretation of quality of life changes. Qual Life Res 1993;2:221-226.
53. Ward MM, Marx AS, Barry NN. Identification of clinically important changes in health status using receiver operating characteristic curves. J Clin Epidemiol 2000;53:279-284.
54. Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol 2003;56:395-407.
55. Hobart J, Cano S. Rating Scales for Clinical Studies in Neurology—Challenges and Opportunities. US Neurology [Internet] 2008;4(1).
56. Hobart J, Cano S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. Health Technol Assess 2009;13:iii, ix-x, 1-177.
57. Kampen J, Swyngedouw M. The Ordinal Controversy Revisited. Qual Quant 2000;34:87-102.
58. Merbitz C, Morris J, Grip JC. Ordinal scales and foundations of misinference. Arch Phys Med Rehabil 1989;70:308-312.
59. Michell J. Measurement: a beginner's guide. J Appl Meas 2003;4: 298-308.
60. Wright BD, Linacre JM. Observations are always ordinal; measurements, however, must be interval. Arch Phys Med Rehabil 1989;70: 857-860.
61. Michell J. Measurement scales and statistics: a clash of paradigms. Psychol Bull 1986;100:398-407.
62. Kasner SE. Clinical interpretation and use of stroke scales. Lancet Neurol 2006;5:603-612.
63. Stevens SS. On the theory of scales of measurement. Science 1946;103:677-680.
64. Kersten P, White PJ, Tennant A. Is the pain visual analogue scale linear and responsive to change? An exploration using Rasch analysis. PLOS ONE 2014;9:e99485.
65. Baker B, Hardyck C, Petronovich L. Weak measurement vs. strong statistics: an empirical critique of S.S. Stevens proscriptions on statistics. Educ Psychol Meas 1966;29:291-309.
66. DeVellis RF. Classical Test Theory. Med Care 2006;44:S50-S59.
67. Novick MR. The axioms and principal results of classical test theory. J Math Psychol. 1966;3:1-18.
68. Hambleton RK, Jones RW. Comparison of classical test theory and item response theory and their applications to test development. Educ Measure 1993;12:38-47.
69. Lord FM. The relation oftest score to the trait underlying the test. Educ Psychol Measure 1953;13:517-548.
70. Lord FM. The relation of the reliability of multiple choice tests to the distribution of item difficulties. Psychometrika 1952;17:181-194.
71. Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. In: Danish Institute for Educational Research C, editor. Chicago, IL: The University of Chicago Press; 1960.
72. Sadjadi R, Conaway M, Cutter G, Sanders DB, Burns TM, Group MGCM-QS. Psychometric evaluation of the myasthenia gravis composite using Rasch analysis. Muscle Nerve 2012;45:820-825.
73. Roalf DR, Moore TM, Wolk DA, et al. Defining and validating a short form Montreal Cognitive Assessment (s-MoCA) for use in neurodegenerative disease. J Neurol Neurosurg Psychiatry 2016 Apr 12 [Epub ahead of print].
74. Carlozzi NE, Schilling SG, Lai JS, et al. HDQLIFE: the development of two new computer adaptive tests for use in Huntington disease, Speech Difficulties, and Swallowing Difficulties. Qual Life Res 2016;25:2417-2427.
75. Burns TM, Sadjadi R, Utsugisawa K, et al. An international clinimetric evaluation of the MG-QOL15, resulting in slight

revision and subsequent validation of the MG-QOL15r. Muscle Nerve 2016 May 24 [Epub ahead of print].

76. Brook RH, Ware JE, Jr., Rogers WH, et al. Does free care improve adults' health? Results from a randomized controlled trial. N Engl J Med 1983;309:1426-1434.

77. Tarlov AR, Ware JE, Jr., Greenfield S, Nelson EC, Perrin E, Zubkoff M. The Medical Outcomes Study. An application of methods for monitoring the results of medical care. JAMA 1989;262:925-930.

78. Hays RD, Sherbourne CD, Mazel RM. User's Manual for the Medical Outcomes Study (MOS) Core Measures of Health-Related Quality of Life. Santa Monica, CA: RAND Corporation; 1995.

79. Guidance for Industry: Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. Food and Drug Administration, 2009.

80. Rothrock NE, Kaiser KA, Cella D. Developing a valid patient-reported outcome measure. Clin Pharmacol Ther 2011;90:737-742.

81. Gnanasakthy A, DeMuro C, Clark M, Haydysch E, Ma E, Bonthapally V. Patient-reported outcomes labeling for products approved by the Office of Hematology and Oncology Products of the US Food and Drug Administration (2010–2014). J Clin Oncol 2016;34:1928-1934.

82. Reflection Paper on the Regulatory Guidance for the use of Health-Related Quality of Life (HRQL) Measures in the Evaluation of Medicinal Products. European Medicines Agency, London, 2005.

83. Appendix 2 to the Guideline on the evaluation of anticancer medicinal products in man:The use of patient-reported outcome (PRO) measures in oncology studies. European Medicines Agency, London, 2016.

84. Ellwood PM. Shattuck lecture—outcomes management. A technology of patient experience. N Engl J Med 1988;318:1549-1556.

85. Basch E, Spertus J, Dudley RA, et al. Methods for developing patient-reported outcome-based performance measures (PRO-PMs). Value Health 2015;18:493-504.