J Rehabil Med 2005; 37: 193–201

REVIEW ARTICLE

CLINIMETRICS IN REHABILITATION MEDICINE: CURRENT ISSUES IN DEVELOPING AND APPLYING MEASUREMENT INSTRUMENTS[1]

Joost Dekker, Annet J. Dallmeijer and Gustaaf J. Lankhorst

From the Department of Rehabilitation Medicine and Institute for Research in Extramural Medicine, VU University Medical Center, Amsterdam, The Netherlands

Clinimetrics in rehabilitation medicine, i.e. the field of developing, evaluating and applying measurement instruments, has undergone considerable progress. Despite this progress, however, several issues remain. These include: (i) selection of an instrument out of the wide range available; (ii) using an instrument in a variety of diagnostic groups; (iii) using an instrument in individual patients, as opposed to a group of patients; and (iv) the use of instruments in clinical practice. This paper reviews these issues, as well as current attempts at resolving them. Illustrative examples are given. It is concluded that solutions seem to be available, but considerable research effort is required to make these a reality. Clinimetrics in rehabilitation medicine remains a field with challenging opportunities for research.

Key words: measurement, review, diagnostic groups, clinical practice.

J Rehabil Med 2005; 37: 193–201

Correspondence address: Joost Dekker, Department of Rehabilitation Medicine, VU University Medical Center, PO Box 7057, NL-1007 MB Amsterdam, The Netherlands. E-mail: j.dekker@vumc.nl

Submitted December 1, 2004; accepted February 25, 2005

INTRODUCTION

Measurement in rehabilitation medicine concerns functioning and disability: impairments of body structures and functions, activity limitations and participation restrictions. Measurement may also concern environmental and personal factors that affect functioning and disability, but this article will focus on measurement of functioning and disability. The measurement of functioning and disability generally has 1 of 3 aims: diagnosis, prognosis or evaluation (1). In diagnosis, the aim of measuring is to discriminate between subjects. For example, in stroke patients one may wish to discriminate between patients with good or poor bladder function. In prognosis, the aim is to discriminate between subjects on a longitudinal basis. One example is the measurement of bladder function or other bodily functions at admission in order to discriminate between stroke patients who will or will not be able to live independently in 6 months time. In evaluation, the aim of measurement is to evaluate changes in functioning and disability over time. This may be illustrated with the monitoring of walking ability, as an indicator of progress during rehabilitation. A more complex example is a clinical trial evaluating the differential change in walking ability in groups of patients being treated with different exercise regimens.

For these 3 purposes – diagnosis, prognosis and evaluation – a wide range of measurement instruments is available. The methodology for developing and evaluating these instruments is becoming increasingly sophisticated. Traditional clinimetric methods for evaluating reproducibility, validity and feasibility (2) have been supplemented with methods to evaluate responsiveness (3) and interpretability (4), thereby extending the evaluation of the quality of measurement instruments. Next to classical test theory, item response theory has been introduced, which offers new options in developing and using measurement instruments. Furthermore, clinicians are increasingly inclined to introduce measurement in clinical practice. Thus, the field of clinimetrics in rehabilitation medicine seems to be developing rapidly.

Despite these encouraging developments, several issues have not yet been resolved in a satisfactory way. These issues are primarily related to the development and application of measurement instruments. They include: (i) selection of an instrument out of the wide range available; (ii) using an instrument in diverse diagnostic groups; (iii) using an instrument in individual patients, as opposed to a group of patients; (iv) the use of instruments in clinical practice. The goal of the present article is to summarize these issues and to present current ideas about potential solutions.

SELECTION OF A MEASUREMENT INSTRUMENT

A wide range of instruments is available to measure components of functioning and disability. Even when focussing on a specific aspect of functioning or a specific category of patients, one is typically confronted with a disturbingly wide range of options. In a way, the situation in measuring health resembles the

[1] Partly based on a lecture by Gustaaf Lankhorst at the International symposium “Measurement and evaluation of outcomes in rehabilitation”, September 2004.

© 2005 Taylor & Francis Group Ltd. *ISSN 1650–1977*
DOI 10.1080/16501970510035179

pre-Napoleonic era, when a variety of measures of length were in use, thus creating confusion and impediments to trade. In the field of rheumatology, for example, more than 100 measures of "patient outcomes" were identified (5); and this is clearly a selection only, because neither biomedical nor biomechanical nor work-related measures were included in this review. With such high numbers of instruments available, the question of how to select a measurement instrument becomes of paramount importance. Because of the sheer number, it is not an easy task to select the instrument that is best suited to a particular purpose, even when one is aware of all the instruments available. Furthermore, explicit and transparent criteria for selecting an instrument should be available.

Systematic reviews of measurement instruments

A potential solution is to perform a systematic review of measurement instruments. In a systematic review, one aims to identify all measurements which are available for a specific purpose, using systematic searches in electronic databases and using explicit criteria to include or exclude instruments. This procedure results in a set of selected instruments, which subsequently are described and evaluated. Descriptive information on the instruments includes the goal of measurement, the nature of the measurement instrument (e.g. questionnaire, rating of performance, measurement of physical properties such as force or pressure), the specific populations for which the instrument was developed, the format of the measurement instrument (e.g. number of items, response options, minimum and maximum score) and issues related to feasibility (e.g. time needed to perform the measurement, required equipment and training).

In order to evaluate the selected instruments, information on clinimetric properties is extracted from the studies identified during the systematic search. Recently, a checklist has been developed which facilitates the systematic evaluation of clinimetric properties of measurement instruments (6, 7). This checklist focuses on questionnaires and contains items on validity, reproducibility, responsiveness, interpretability and feasibility (practical burden). For illustrative purposes, some of these items will be summarized here. For further information, the reader is referred to the original publications (6, 7).

The concept of validity refers to the degree to which an instrument measures what it is supposed to measure. The checklist focuses on content and construct validity. Criterion validity was not included in the checklist: a gold standard is frequently not available in rehabilitation, which precludes evaluation of criterion validity (i.e. the degree to which the scores on an instrument correspond to the scores on the gold standard). Content validity examines the extent to which the domains of interest are comprehensively sampled by the measurement instrument. In order to rate content validity, the methods used for item selection and item reduction are evaluated: because the questionnaires are supposed to address disability as experienced by patients, a positive rating for content validity is given when patients were involved in the process of item selection and reduction. Construct validity refers to the extent to which scores on a particular instrument relate to other measures in a manner that is consistent with theoretically derived hypotheses concerning the constructs that are being measured. In the checklist, construct validity is considered to be adequately tested if hypotheses were specified and the results of the studies on construct validity correspond with these hypotheses.

Reproducibility is the extent to which an instrument is free of measurement error. The checklist focuses on test-retest reliability and agreement. Reliability concerns the degree to which patients can be distinguished from each other, using a particular measurement instrument. Agreement concerns the degree to which scores on repeated observations correspond with each other. In the checklist, statistics and cut-off values for reliability and agreement to be considered adequate are defined.

Responsiveness refers to an instrument's ability to detect important change over time in the concept being measured. Responsiveness can be conceptualized as longitudinal validity: does the instrument measure changes in the concept that it is supposed to measure? Testing responsiveness is analogous to testing construct validity: hypotheses on changes in the concept being measured should be formulated and tested, using the instrument being studied. In the checklist, responsiveness was considered adequately tested if hypotheses were specified and when the results were in correspondence with these hypotheses.

Concerning feasibility (practical burden), the checklist focuses on time required for administration and ease of scoring. In the checklist, criteria for ease of administration and ease of scoring are provided.

Interpretability can be defined as the degree to which one can assign meaning to quantitative scores: information is required on the clinical meaning of scores and on which difference between scores can be regarded as clinically meaningful. In the checklist, interpretability is rated positive if information is presented on a minimal clinically important difference (MCID); or if information is presented that could facilitate the interpretation of scores (e.g. distribution of scores in subgroups of patients, information on the relationship of scores to well-known functional measures or clinical diagnoses, distribution of scores before and after treatment).

Illustration of systematic review, using the checklist

The checklist was used in the evaluation of questionnaires on shoulder disability (6). The systematic search and selection of instruments resulted in 28 studies referring to 16 shoulder disability questionnaires. Descriptive information and detailed information on the clinimetric properties of these questionnaires was provided, using the checklist. Furthermore, a table summarizing the quality assessment was provided. From that summary table, it was concluded that 1 specific questionnaire received most positive ratings, i.e. overall, this questionnaire seems to have the best clinimetric properties. This questionnaire was the Disabilities of Arm, Shoulder and Hand Scale (DASH, 7). However, as pointed out by the authors, the best scale is not always best for a particular purpose. For example, for the evaluation of shoulder surgery, a questionnaire specifically developed for shoulder surgery (8) may be preferred over the DASH, which addresses shoulder disability in general instead of specifically shoulder operations. Similarly, if one focuses on diagnostic discrimination, a questionnaire with a particularly high score for reliability (i.e. the Simple Shoulder Test, 9) may be more appropriate than the DASH: the latter instrument seems to be an all-purpose instrument, while the former might be more appropriate if diagnostic discrimination is the primary and overriding goal of measurement. Thus, in selecting an instrument, one cannot simply select the instrument with the best overall rating. Instead, in the context of a specific clinical or research setting, one should select the instrument that is best suited for the particular purpose in that specific context. The process of selecting a measurement instrument starts with clearly specifying the particular purpose of measurement. In the next step, the information provided in the systematic review may facilitate the selection process: the detailed and systematic description of the instruments and their clinimetric properties facilitates the choice of an instrument for use in a specific setting.

Future developments

The checklist developed by Bot et al. (7) is by no means perfect. However, based on previous checklists and current thinking in clinimetrics, it seems to be the most up-to-date checklist currently available. Further development of this or similar checklists providing transparent and systematic criteria for the evaluation of measurement instrument is clearly indicated. Furthermore, the availability of the current or future checklists may also improve the standards for reporting on clinimetric studies. As pointed out by Bot et al. (7), the quality of reporting on clinimetric studies is currently rather poor. Essential information for the evaluation and selection of measurement instruments was frequently found to be missing. Using the checklist, future authors may improve the quality of their reports on clinimetric studies.

Standardization of measurement

At present, only a few systematic reviews of measurement instruments are available. It is our expectation that – similar to the growth of systematic reviews of clinical trials, observational studies and diagnostic research – more systematic reviews of measurement instruments will become available in the near future, thereby facilitating the selection of instruments from the wide range of instruments described in the literature. The findings in these reviews may also contribute to standardization of measurement in rehabilitation. The current heterogeneity in measurement instruments is an impediment for comparison and synthesis of research findings in rehabilitation. The same applies to clinical practice: communication about patients is hampered by the diversity in measures used. Clearly, a certain degree of standardization in measurement may facilitate communication in rehabilitation medicine. The findings in systematic reviews on measurement instruments may provide important input to the process of standardization. Current attempts at defining which components of functioning and disability should be assessed in various diagnostic categories (11) can be supplemented with the results of systematic reviews on measurement instruments: once consensus has been achieved on which dimensions are to be assessed in a certain category of patients, one can than proceed to a certain degree of standardization of measurement instruments. The results of reviews on measurement instruments based on transparent and systematically applied clinimetric criteria provide essential input to this process.

Measurement always serves a specific purpose. The present call for standardization is made within the general context of acknowledging that a specific measurement instrument might be appropriate for some purposes, but not for others. Thus, to assess particular components of functioning and disability, specific instruments are required. For example, a timed performance test can be used to assess observed aspects of walking ability, while a questionnaire is used to assess subjective walking ability. Furthermore, when assessing a particular component of functioning and disability, the goal of measurement may be diagnosis, prognosis or evaluation: the clinimetric properties of the instrument might make it more suitable for one of these goals, but not for another. For example, reproducibility and validity are cardinal criteria for a diagnostic purpose, while for the evaluation of treatment responsiveness is most important. A certain degree of standardization of measurement instruments is clearly indicated, but this should not obscure the fact that specific instruments are required to fulfil specific measurement purposes.

APPLYING MEASUREMENT INSTRUMENTS IN VARIOUS DIAGNOSTIC GROUPS: DIMENSIONALITY OF MEASUREMENT INSTRUMENTS

Generic versus disease specific measures

Instruments can be categorized as either generic or specific measures. Generic measures intend to measure the same construct (activity limitations or participation restrictions) across different patient groups, while specific measures are developed for application in one diagnostic group only. The use of generic measures has several advantages, including the reduced need for developing and testing different instruments for all patients groups separately, and uniformity of measurement in rehabilitation facilities (which is expected to facilitate communication between rehabilitation professionals). An important advantage is that, when using generic measures, the burden of different diseases and disabilities can be compared among patient groups and, in some cases, with the healthy population. Although it is seems inevitable that the outcomes of generic measures provide less specific information about each patient group, it is also suggested that well-designed generic measures yield results that are at least as good as disease specific instruments (2).

Measurement instruments, whether generic or disease specific, usually consist of 1 or more subscales (domains), where items are summed to form a total subscale score. It is important that the subscales measure 1 clearly defined underlying construct, such as mobility or communication, preferably based on the domains of the International Classification of Functioning, Disability and Health (12). All items of the same (sub)scale are supposed to measure 1 construct, and should therefore be related to the construct that is intended to be measured. This implies that all the individual items of the same subscale should be moderately correlated with each other and that each item should be correlated with the total scale score it belongs to and not (or only weakly) to any other subscale (2).

Dimensional structure of measurement instruments

In rehabilitation medicine typically several patient groups with varying disabilities and disease characteristics are treated, which may explain the popularity of generic measures. However, prior to applying generic measures in a variety of patient populations, the clinimetric properties of generic measurement instruments should be investigated in each patient group separately. Apart from studying clinimetric properties, such as reproducibility and interpretability, it is important to investigate the dimensional structure of the instrument in each patient group separately. In order to be able to calculate sum scores from the items, the dimensions of the measurement instrument have to be consistent across groups. It should therefore be tested whether the items correlate with the same subscale scores (i.e. the dimensions they belong to) in all patient groups. If items behave differently (i.e. do not measure the supposed construct) it should be reconsidered whether this item can be used in this patient group. This is especially important in rehabilitation research because pooled analyses are frequently performed, evaluating outcome in a diagnostically mixed group of patients. Obviously, the above also applies when using disease specific instruments in other patient groups than that they have been developed for. It may be possible that the same instrument can be used in other patient populations, but this should be tested in advance.

Traditional methods to investigate the dimensional structure of measurement instruments are factor analysis (or principal component analysis) and determining internal consistency of the dimensions (subscales) by calculating Cronbach's alpha. However, these methods have some recognized limitations with the use of dichotomous and ordinal data (13). Another approach that is increasingly applied in rehabilitation medicine for investigating dimensional structure and scalability of measurement scales is Rasch analysis.

The Rasch measurement model is based on item response theory (14). It converts ordinal scales into an interval measure, which expresses the difficulty of items and ability of the subjects on 1 measurement continuum (logit or log odds unit, 14). The Rasch model can be used to explore whether all items of the scale measure a single construct (unidimensionality of a scale). If items do not fit the model, it indicates that these items do measure a different construct. In general, Rasch analysis can be applied for evaluating dimensions and scalability of newly developed and existing instruments, but can also be used to convert an ordinal scale into an interval measure for the statistical analysis. Because item difficulties are expressed on the same measurement continuum, Rasch analysis can also be applied to investigate variations in item difficulties among groups. The hierarchy of items, i.e. the location of the items on the interval scale, is assumed to be invariant across groups. Variation in item difficulty between groups is referred to as differential item functioning (DIF). Different sources of DIF can be identified, such as age, gender or culture. However, disease can also be a source of DIF that should be taken into account when comparing the outcomes of different patient groups, or when pooling data in a (statistical) analysis. If item difficulties vary between groups, identical sum scores of different (patient) groups are likely to result from different item profiles and thus different levels of functioning which, again, hampers comparison between groups. This is a fairly new field of research and disease as a source of DIF has not yet been fully investigated.

Findings in rehabilitation medicine

Examples of generic measures that are frequently used in rehabilitation are the Barthel Index (BI), that measures physical disability with 10 ordinal items, and the Functional Independence Measure (FIM) for measuring disability in the motor (13 items) and cognitive domain (5 items). The BI was originally developed for patients with neuromuscular and musculoskeletal disorders but has been used in several other diagnostics groups. However, some studies showed that the BI is not always suitable to measure physical functioning because of its lack of unidimensionality (15, 16). The dimensional structure of the FIM has been investigated extensively by several researchers using a variety of methodologies. Results from factor analysis showed that the 2-factor structure, that was proposed by Linacre et al. (17), could be confirmed in most patient groups, although in some groups more than 2 dimensions could be distinguished (18). However, other studies applying the more stringent Rasch analysis showed however that the motor scale is not unidimensional in all patient groups (19, 20).

In a recent Dutch study on functional prognosis in neurological disorders, results of different patient groups were pooled to investigate shared determinants of functional outcome. In order to be able to perform a pooled analysis of the different patient groups involved in the study, DIF (among patient groups) was investigated in several instruments using Rasch analysis. Among others, the SF-36 Physical Functioning scale (10 items on a 3-point rating scale) and the FIM were used to describe the functional outcome in these patient groups. As an example, results of the DIF analysis in patients with stroke, multiple sclerosis and amyotrophic lateral sclerosis are shown in Fig. 1 (unpublished results).

In this figure the item difficulties (expressed in logits) are shown for each group separately. It shows that the overall hierarchy is comparable among groups, but that some item

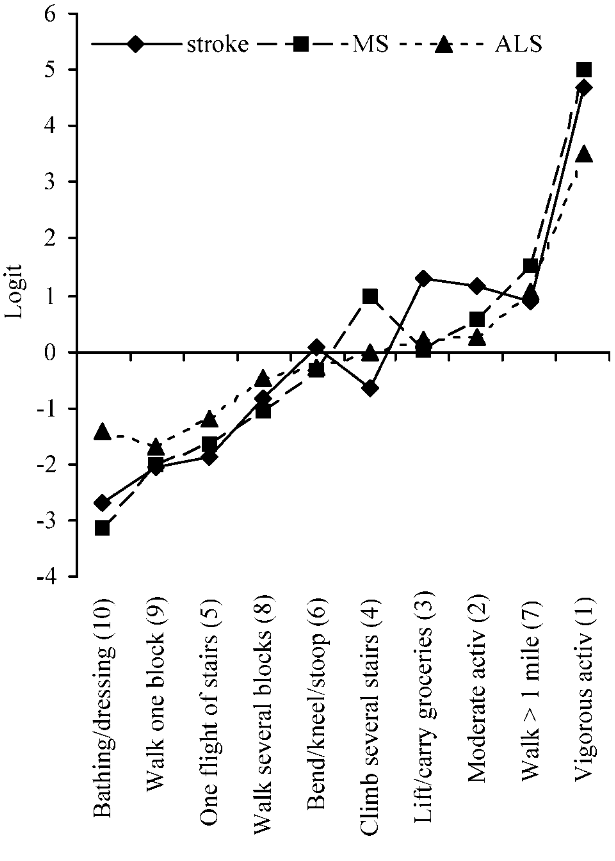

Fig. 1. Item difficulties (in logits) of the SF-36 Physical Functioning scale for patients with stroke, multiple sclerosis (MS) and amyotrophic lateral sclerosis (ALS).

difficulties differ considerably. Comparison of item difficulties among groups identified DIF in all group comparisons, but the number of items showing DIF and the extent of DIF were rather small. In Fig. 2 an example of a DIF plot is given, showing the item difficulties for patients with stroke and MS. Three out of the 10 items showed DIF. In contrast, results of the DIF analysis performed on the FIM motor scale in patients with stroke, MS and TBI showed less promising results; 7 out of 11 fitting items showed DIF.

Future developments

To perform a pooled analysis, or when comparing results among patient groups, adjustments for DIF can be applied, as recently described by Tennant et al. (21). Using this procedure, items showing DIF are split up as disease specific items (see for further explanation Tennant et al., 21). To what extent adjustments for DIF among patient groups is required (or necessary) in the different generic measurement instruments frequently used in rehabilitation is, to our knowledge, not yet investigated. It is, however, possible that DIF among patient groups causes misfit of the data to the Rasch model. Other sources of DIF, such as gender, age group or culture, should also be further explored in future studies. Adjustment for DIF may lead to improvement in measurement properties, such as improved discrimination between groups and better responsiveness, but this also remains to be investigated.

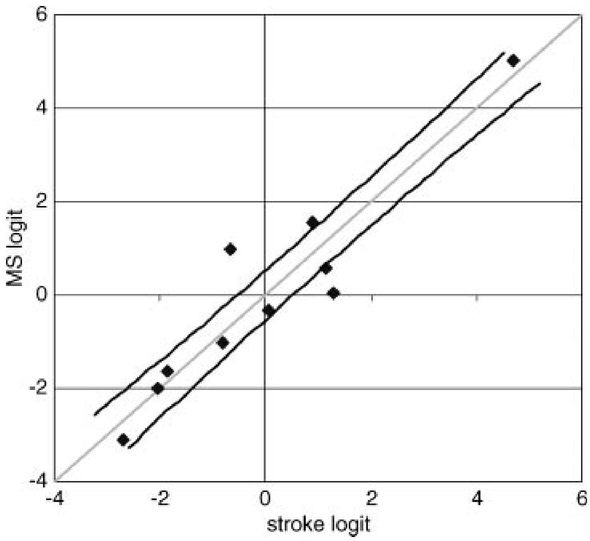

Fig. 2. Example of differential item functioning (DIF) plot for the SF-36 Physical Functioning scale in patients with stroke and multiple sclerosis (MS). Item difficulties for stroke are plotted on the x-axis and for MS on the y-axis. An identity line is drawn through the origin with a slope of 1. The area between the 2 other lines indicates the 95% confidence interval. Items outside this area demonstrate DIF (unpublished results).

MEASUREMENT OF INDIVIDUAL PATIENTS VERSUS A GROUP OF PATIENTS

Measurement in rehabilitation research typically concerns a group of patients. An important reason for doing so is that measurement error is reduced by taking the average of the measurements in the group of patients: increasing the number of observations reduces the error in the outcome of the measurement. In clinical practice, however, one is interested in measuring an individual patient: a measurement instrument may be used to get an objective and quantitative value of impairments of the body structures and functions, activity limitations and participation restrictions in an individual patient.

Error of measurement

When measuring individual patients, the requirements for the quality of the instruments used are higher than in the research setting (22, 23). This is in contrast to common opinion in clinical practice: clinicians tend to think that the quality requirements for measurement in clinical practice are lower than in research settings. However, taking the average of observations in a group reduces the error of measurement; when measuring an individual patient, one is confronted with the full, non-attenuated error of measurement.

Thus, it has been frequently stated that the reliability coefficient of instruments used in groups should be at least 0.70, while the reliability coefficient of instruments used in individual patients should be at least 0.90 (2, 13). Although the general truth of this rule of thumb can be questioned, it is true that in order to be reasonably certain that the score of an individual falls within certain limits one needs a rather precise instrument with little measurement error, i.e. the reliability of the instrument should be relatively high. Conversely, if the clinician uses an instrument that has adequate reliability in the group setting, he should be aware of the fact that the measurement of individuals is associated with a larger error. Because the reliability of many existing instruments, although adequate in the group setting, does not meet the high standards of the individual setting, clinicians should be aware that measuring individual patients is generally associated with a relatively high degree of error and the results should be interpreted with some caution.

In addition to uncertainty about the actual measurement value in a diagnostic or prognostic situation, measurement error limits the ability to detect clinical change in a patient. When evaluating change in an individual patient, real clinical change may be obscured by measurement errors in the instrument used. An instrument with a large error of measurement (i.e. with low test-retest reliability) may fail to detect real clinical improvement in an individual patient. The statement made above about the need for a relatively high reliability when assessing an individual patient applies also to the responsiveness of an instrument, i.e. the ability of an instrument to detect clinical change (3). When evaluating change in an individual patient, the requirements about the responsiveness of the instrument used are higher than in the group setting. Failure to demonstrate improvement in an individual patient may be a true observation, but it may also be the result of using an instrument which responsiveness is not good enough to demonstrate change in individual patients.

Adaptive or tailored testing

A potential solution for these problems is the development of so-called adaptive or tailored testing, based on item response theory (IRT, 2, 14). This approach consists of developing a disability scale, comprising a large number of items (e.g. 200 items) which form a hierarchy: a hierarchy ranging from items indicating a low level of disability to items indicating severe disablement. If these items form a perfect hierarchy (as shown by IRT-techniques), one can use a few items to screen for the global level of disability; if the global level of disability is known, one can then administer that part of the scale that corresponds to the patient's level of disability and thereby determine the exact level of disability. For example, in a patient who functions relatively well (as shown by the screening items), one administers items from the low disability end of the scale; for a patient in whom the screening items signal poor functioning, one can administer items from the high disability end of the scale. In determining the exact disability level, a rather large but still feasible number of items (e.g. 30 items) can be used: because one has to administer only items from the end of the scale that corresponds to the level of disability of the patients, all other items can be disregarded; the perfect hierarchy of the scale ensures that the patient will or will not pass the disregarded items. This large but feasible number of items reduces measurement error and ensures precise measurement. Theoretically, this approach of adaptive or tailored testing offers the possibility to reduce measurement error considerably, thereby allowing measurement with little error in individual patients. Dijkers (24, 25) has demonstrated the value of adaptive testing in simulation studies on the FIM, but the rather low number of items in the FIM limits the value of adaptive testing using the FIM. Thus, the actual value of this approach in rehabilitation practice remains to be demonstrated.

Individualized measures

The trend towards patient-oriented rehabilitation has induced the development of individualized measures (or patient-specific measures). In patient-oriented rehabilitation, it is emphasized that the patient has a strong say in defining the problems that should be addressed during rehabilitation. In this approach, individualized measures are used, which are adapted to the problems of a specific patient. In an individualized measure, the patient defines the nature of the problem; and the patient subsequently rates the severity of the problem. It is deemed important that the patient describes the nature of the problem, in his or her own words and in the context of his or her own daily experiences. In clinimetric terms, this procedure is expected to enhance the validity of the measurement of disability: by letting the patient define the nature of the problem, one presumably measures disability exactly as experienced by the patient. In traditional instruments with standardized items (such as the Sickness Impact Profile or the SF 36), a selection of potential problems is described, using common wordings; thus, there is a risk that the patient's specific problem is not mentioned or the problem is described inadequately. Individualized measures try to circumvent this, by letting the patient define the problem.

Canadian Occupational Performance Measure

An example of an individualized measure is the Canadian Occupational Performance Measure (COPM, 26). In a semi-structured interview with an occupational therapist, the patient identifies problems in activities of self-care, productivity and leisure. The patient then prioritizes these problems and selects the 5 most important problems. The patient rates both performance (i.e. ability to perform the activity) and satisfaction (i.e. satisfaction with activity) of these problems on a 10-point scale. The performance ratings are then added to a summary score, as are the satisfaction ratings.

The divergent validity of the COPM was studied (26). Divergent validity refers to the ability of an instrument to differentiate the concept under study from other constructs. In support of the divergent validity of the COPM, it was found that for 81 problems out of 443 problems identified with the

COPM, no corresponding item could be found in 2 traditional instruments (i.e. the SIP68 and the Disability Impact Scale (DIP)). Examples included problems with sitting, caring for loved ones such as grandchildren and spouse, and personal appearance. Furthermore, correlations between scores on the COPM and the SIP were low. Thus, there is some support for the hypothesis that the COPM, as an individualized measure, assesses problems that are not assessed with traditional instruments.

Not completely unexpected, it was found that the reproducibility of the COPM left something to be desired (unpublished data). Patients were assessed twice, with an interval of 7 days, by 2 different therapists. Only about two-thirds of the problems prioritized by the patients at the first assessment were also prioritized at the second assessment. Furthermore, the reproducibility (intraclass correlation coefficient) of the performance score was moderate; the same applied to the satisfaction score. Thus, it seems that there is a risk that in a individualized measure like the COPM, the lack of standardized items leads to a reduced reproducibility of the measurement results. The semi-structured interview and the process of prioritizing problems leave room for considerable variation (error) among patients and among test occasions.

As expected, the responsiveness of the COPM appeared to be rather good (unpublished data). Patients were assessed before and after occupational therapy. The COPM was sensitive in detecting improvement as reported by patients on a transition index (criterion responsiveness). In addition, improvement on the COPM correlated with improvement on other measures such as the Sickness Impact Profile (construct responsiveness).

These results suggest that the COPM, as an individualized instrument, indeed measures aspects of patients' problems which are not assessed by traditional instruments consisting of standardized items. Similarly, it seems that having patients define the nature of the problem indeed results in a responsive measure. However, probably as a result of the individualized nature of this instrument, the COPM is not the best instrument for the purpose of comparing patients and distinguishing among patients (reliability).

USING MEASUREMENT INSTRUMENTS IN CLINICAL PRACTICE

Increasingly, clinicians are inclined to use measurement instruments in clinical practice. This may be due to an intrinsic interest in measurement, which yields a quantitative estimate of impairments of the body structures and functions, activity limitations and participation restrictions. On the other hand, clinicians are under extrinsic pressure from the management of their institute to introduce measurement in clinical practice, especially to evaluate rehabilitation outcome. Given the increasing use of measurement in clinical practice, be it intrinsically or extrinsically motivated, a critical appraisal of this trend seems to be in place.

Clinical assessment versus measurement in research projects

Clinical assessment of a patient in rehabilitation medicine is different from measurement for research applications. In clinical rehabilitation we are dealing with patients with permanent disabilities as a result of disease or injury. Clinical assessment aims at identifying problems and potential solutions. It does not necessarily include measurement. The purpose of clinical assessment of a patient is to identify his/her activity limitations and restrictions in participation, to identify impairments that underlie the activity limitations and to find options for treatment of these conditional impairments. This is typically done by history taking, including a checklist of activities and participation and by physical examination, sometimes supplemented with additional examinations (X-ray, gait analysis). On the basis of this assessment a rehabilitation diagnosis is made, rehabilitation goals are defined and a rehabilitation program is designed (what is desirable? what is possible?).

Trying to combine measurement and clinical assessment is not always easy. The Rehabilitation Activities Profile is a clinical assessment tool (RAP, 27) with the domains: Communication, Mobility, Personal care, Occupation and Relationships. It can be used as a checklist with or without a 0–3-point severity rating. Using the quantitative version turned out to be rather time-consuming and did not increase satisfaction in RAP-teams (28). The qualitative version, however, is widely used in The Netherlands. On the other hand, Wikander et al. (29) have reported the successful use of the FIM for both team communication and assessment/evaluation. In a randomized controlled trial patients in the FIM group more patients regained continence before discharge than in the control group. There was also a greater improvement in well being in the FIM group.

Outcomes measurement

Outcomes measurement came up during the 1980s as a result of increasing healthcare costs, although it was also expected to improve quality of care and patient outcome (30). The challenge was accepted by the US rehabilitation community, which resulted in the development of the Functional Independence Measure (FIM). The FIM was soon used in many rehabilitation facilities in the US and Europe. Traditional clinimetric properties (e.g. reliability) are reported to be good (31).

What is certainly important in this respect about FIM are its instruction and certification courses. However, even trained FIM users have been found to be biased in their judgement of FIM items, when they had knowledge of scores of other team members on other items (32). An important question regarding outcomes research is whether patients have any benefit from the use of outcome measures on a routine basis in clinical practice. When the data are being gathered and used as part of a quality of care system, this is probably the case. However, rehabilitation teams are often under pressure to do outcome measurements as part of "best practice". There seems to be no good reason to do that, because the time used for measurement might be at the expense of treatment time. Outcome measurement is sometimes recommended to improve accountability of rehabilitation providers. Here caution should be applied. Clinicians might be tempted to change their case mix rather than improving the quality of rehabilitation care, in order to meet demands about outcome performance.

Thus, the trend towards increased use of measurement instruments in clinical practice should be seen with some reservation. Clinical assessment of a patient in rehabilitation medicine involves much more than the mere application of measurement instruments and it might not always be wise to combine measurement and clinical assessment. The use of outcome measures in clinical practice only for management purposes is not to be recommended: outcome measures should ideally be used in the context of improvement of quality of care.

CONCLUSION

It has been argued that, despite promising developments, several issues concerning the development and application of measurement instruments in rehabilitation medicine remain to be resolved. This paper describes and illustrates these issues, as well as current attempts at solving these issues. (i) In several fields of research, the range of instruments available is disturbingly wide: systematic reviews of measurement instruments may facilitate the selection of an adequate instrument from all instruments available. Furthermore, a certain degree of standardization may facilitate both synthesis of results in research and communication in clinical practice. (ii) Application of an instrument in a variety of diagnostic groups requires that the dimensional structure of the instrument and the difficulty of separate items is comparable across these diagnostic groups. Statistical techniques, including factor analysis and Rasch analysis, are available to test this. If item difficulty varies substantially among diagnostic groups, statistical corrections are possible, but it remains to be demonstrated that these procedures indeed improve the quality of the measurement instrument. (iii) When measuring individual patients (as opposed to a group of patients), one is confronted with a relatively high measurement error. A potential solution is to use highly reliable instruments (i.e. instrument with little measurement error): in rehabilitation medicine, these instruments are not frequently available, however. Another solution could be so-called adaptive or tailored testing. This kind of instrument still has to be developed in rehabilitation medicine. When measuring individual patients, individualized or patient specific measures can be used. The validity and responsiveness of such a measure seem to be rather high, but this seems to be achieved at the expense of a relatively low reliability. (iv) A critical appraisal of the introduction of measurement in clinical practice seems to be indicated. Clinical assessment of a patient is not equivalent to applying measurement instruments: clinical assessment may or may not include measurement. The use of outcome measures only for management purposes is not to be recommended: in clinical practice, outcome measures are ideally used in the context of improvement of quality of care.

In summary, in the development and application of measurement instruments in rehabilitation medicine several issues remain to be solved. It is concluded that solutions for these issues seem to be available, but considerable research effort is required to make these potential solutions a reality. Clinimetrics in rehabilitation medicine remains a field with challenging opportunities for research.

ACKNOWLEDGEMENT

This article is partly based on a lecture by Gustaaf Lankhorst at the International symposium "Measurement and evaluation of outcomes in rehabilitation", given in September 2004.

REFERENCES

1. Guyattt G, Kirschner B, Jaeschke R. Measuring health status: what are the necessary measurement properties. J Clin Epidemiol 1992; 45: 1341–1345.
2. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use (3rd edn). Oxford: Oxford University Press; 2003.
3. Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PMM. On assessing responsiveness of health related quality of life instruments: guidelines for instrument evaluation. Qual Life Res 2003; 12: 349–362.
4. 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.
5. Katz PP. Introduction to special patient outcomes in rheumatology issue of arthritis care & research. Arthritis Rheum (Arthritis Care Res) 2003; 49: 1–4.
6. Bot SD, Terwee CB, van der Windt DA, Bouter LM, Dekker J, de Vet HC. Clinimetric evaluation of shoulder disability questionnaires: a systematic review of the literature. Ann Rheum Dis 2004; 63: 335–341.
7. Bot SD, Terwee CB, van der Windt DA, Bouter LM, Dekker J, de Vet HC. Psychometric evaluation of self-report questionnaires – the development of a checklist. In: Proceedings of the Second Workshop on Research Methodology; 2003, June 25–27, Amsterdam, NL. Amsterdam: VU University Amsterdam; 2003, p. 161–168.
8. Dawson J, Fitzpatrick R, Carr A. Questionnaire on the perceptions of patients about shoulder surgery. J Bone Joint Surg Br 1996; 78: 593–600.
9. Dawson J, Fitzpatrick R, Carr A. The assessment of shoulder instability. The development and validation of a questionnaire. J Bone Joint Surg Br 1999; 81: 420–426.
10. Lippitt SB, Harryman DTI, Matsen FAI. A practical tool for evaluation of function: the simple shoulder test. In: Matsen FA III, Fu FH, Hawkins RJ, eds. The shoulder: a balance of mobility and stability. Rosemont, Illinios: The American Academy of Orthopaedic Surgeons; 1993, p. 501–518.
11. Stucki G, Grimby G. Applying the ICF in medicine. J Rehabil Med 2004; 36 (suppl 44): 5–6.
12. WHO. International Classification of Functioning, Disability and Health. Geneva: WHO; 2001.
13. Nunnally JC, Bernstein IH, eds. Psychometric theory. 3rd edn. New York: McGraw-Hill; 1994.
14. Tesio L. Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. J Rehabil Med 2003; 35: 105–115.

15. Kucukdeveci AA, Yavuzer G, Tennant A, Suldur N, Sonel B, Arasil T. Adaptation of the modified Barthel Index for use in physical medicine and rehabilitation in Turkey. Scand J Rehabil Med 2000; 32: 87–92.
16. Laake K, Laake P, Ranhoff AH, Sveen U, Wyller TB, Bautz-Holter E. The Barthel ADL index: factor structure depends upon the category of patient. Age Ageing 1995; 24: 393–397.
17. 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.
18. Stineman MG, Jette A, Fiedler R, Granger C. Impairment-specific dimensions within the Functional Independence Measure. Arch Phys Med Rehabil 1997; 78: 636–643.
19. Granger CV, Hamilton BB, Linacre JM, Heinemann AW, Wright BD. Performance profiles of the functional independence measure. Am J Phys Med Rehabil 1993; 72: 84–89.
20. Kucukdeveci AA, Yavuzer G, Elhan AH, Sonel B, Tennant A. Adaptation of the Functional Independence Measure for use in Turkey. Clin Rehabil 2001; 15: 311–319.
21. Tennant A, Penta M, Tesio L, Grimby G, Thonnard JL, Slade A, et al. Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model: the PRO-ESOR project. Med Care 2004; 42 (suppl 1): I37–I48.
22. Roebroeck ME, Harlaar J, Lankhorst GJ. The application of generalizability theory to reliability assessment: an illustration using isometric force measurements. Phys Ther 1993; 73: 386–395.
23. de Vet HCW, Terwee CB, Bouter LM. Current challenges in clinimetrics. J Clin Epidemiol 2003; 56: 1137–1141.
24. Dijkers MP, Yavuzer G. Short versions of the telephone motor Functional Independence Measure for use with persons with spinal cord injury. Arch Phys Med Rehabil 1999; 80: 1477–1484.
25. Dijkers MP. A computer adaptive testing simulation applied to the FIM instrument motor component. Arch Phys Med Rehabil 2003; 84: 384–393.
26. Dedding C, Cardol M, Eyssen IC, Dekker J, Beelen A. Validity of the Canadian Occupational Performance Measure: a client-centred outcome measurement. Clin Rehabil 2004; 18: 660–667.
27. Van Bennekom CAM, Jelles F, Lankhorst GJ. Rehabilitation Acitivies Profile. Disabil Rehabil 1995; 17: 169–175.
28. Jelles F, Van Bennekom CAM, Lankhorst GJ, Bouter LM, Kuik DJ. Introducing an innovatieve method in team conferences. Disabil Rehabil 1996; 18: 374–379.
29. Wikander B, Ekelund P, Milsom I. An evaluation of multidisciplinary intervention governed by Functional Independence Meausre (FIM) in incontinent stroke patients. Scand J Rehab Med 1998; 30: 15–21.
30. Ellwood PM. Outcomes management – a technology of patient experience. N Engl J Med 1988; 318: 1549–1556.
31. Hamilton BB, Laughlin JA, Fiedler, RC, Granger CV. Interrater reliability of the 7-level Functional Independence Measure. Scand J Rehabil Med 1994; 26: 115–119.
32. Wolfson AM, Doctor JN, Burns SP. Clinician judgements of functional outcomes: how bias and perceived accuracy affect rating. Arch Phys Med Rehabil 2000; 81: 1567–1574.