INTERNATIONAL CLASSIFICATION OF FUNCTIONING DISABILITY AND HEALTH

Scales could be developed based on simple clinical ratings of International Classification of Functioning, Disability and Health Core Set categories

Eva Grill\textsuperscript{a}, Gerold Stucki\textsuperscript{a,b,c,*}

\textsuperscript{a}Institute for Health and Rehabilitation Sciences, Ludwig Maximilian University, Munich, Germany
\textsuperscript{b}Department of Physical Medicine and Rehabilitation, University Hospital Munich, Ludwig Maximilian University, Munich, Germany
\textsuperscript{c}Swiss Paraplegic Research, Nottwil, Switzerland

Accepted 12 February 2008

Abstract

Objective: The aim of the study was to examine whether clinical ratings of International Classification of Functioning, Disability and Health categories can be integrated into parametric scales, which provide a reliable estimation of specified patient problems and rehabilitation goals using the example of mobility of the upper and lower extremities in the acute hospital situation.

Study Design and Setting: Psychometrical study based on data from a prospective multicentric cohort study in patients with musculoskeletal conditions in the acute hospital.

Results: Two hundred thirty-four patients were included (mean age 56, 50% female. Forty-four percent with a diagnosis involving lower extremities, 18% with a diagnosis involving the hip, 18% with a diagnosis involving upper extremities, 16% with a diagnosis involving the spine). After adjustment for differential item functioning two separate scales for upper and lower extremity mobility could be constructed. The constructed scales had 10 (upper) and eight (lower) items displaying adequate to good fit.

Conclusion: The results of this study indicate that it could be possible to develop scales based on categories of the International Classification of Functioning, Disability and Health Core Sets. This may be a promising approach for areas where psychometrically sound measures are not available. © 2009 Elsevier Inc. All rights reserved.

Keywords: Outcome assessment (health care); Item response theory; Classification; Physical function; Health status measurement; Mobility limitation

1. Introduction

Human functioning is a central concept in medicine relevant to the individual and the society. Its measurement is therefore essential for research on effective patient care [1]. In rehabilitation practice and research it is the basis for the assessment of patients problems and the setting of rehabilitation goals, the assignment to services and interventions, intervention management, and evaluation of intervention outcomes [2,3].

Measures of functioning that are useful in the hand of health professionals across health care systems and services need to be simple and based on a universal language of functioning. Measures that are useful for the specification and evaluation of rehabilitation goals need to be tailored to patients’ problems. In rehabilitation research, measures should be etiologically neutral to allow comparisons across health conditions and they should closely reflect rehabilitation goals to be valid and assure responsiveness.

No currently available measures fulfill these requirements. In general, current measures are not simple and are thus often used only for research or health statistics but not in daily clinical practice. Lacking a universal language or classification of functioning current measures vary widely with respect to included items [4], to assumed underlying constructs and accordingly, for example, to the definition and naming of scales and subscales [5].

Currently widely used measures in rehabilitation such as the Functional Independence Measure [6] or the Barthel Index [7] have been developed to measure, for example, resource utilization or to assign patients to appropriate levels of rehabilitation services. However, they are neither comprehensive assessments of functioning [8] nor invariant across different health conditions [9]. Most importantly,
they are not specific enough when assessing patients problems and defining rehabilitation goals. Also, most of the current measures used in rehabilitation are categorical measures and hence not qualified for parametric statistics such as calculating means, standard deviations, or even change scores [10].

Thus, there is the need to develop a practical approach which enables health professionals and researchers to develop interval scaled clinical measures which are simple to rate and interpret and useful to specify and follow patients problems and rehabilitation goals in clinical practice. For this, we need first a unifying framework and classification which comprehensively covers the human experience in relation to functioning irrespective of the underlying condition and in relation to a wide range of personal and environmental factors. Second, we need to show whether clinical ratings of the categories of such a universal classification can be integrated into parametric scales, which provide a reliable estimation of specified patient problems and rehabilitation goals.

The International Classification of Functioning, Disability and Health (ICF) [11] has now emerged as the accepted universal language and framework for rehabilitation [12] to describe and classify functioning, health and disability. The ICF is organized into four components, Body Functions, Body Structures, Activities and Participation, and Environmental Factors subdivided into over 1,400 different, hierarchically arranged categories. With both the opportunities of ICF and newly developed methods of item response theory combined it is now possible to examine whether it is in principle possible to develop clinical measures which are tailored to patients problems by simply integrating clinical ratings of relevant ICF categories with the ICF qualifier. The ICF provides qualifier which range from 0 to 4 for each category to account for the severity of the problem. The ICF qualifier can be used as a direct measurement of human functioning by an expert based on given definitions and defined anchors.

When developing clinical measures tailored to patients problems and rehabilitation goals one can rely on parts of the ICF which comprehensively cover aspects of functioning for specific health conditions [13] or health care situations such as the acute hospital or early acute rehabilitation facilities [14]. Those parts, the so-called ICF Core Sets, were developed in an international effort supported by the World Health Association.

If then, for example, a scale is needed to assess mobility of the upper or lower extremities in the acute hospital, ICF categories can be selected out of one of the ICF Core Sets in the acute hospital situation. By applying item response theory, one can then examine whether the selected categories indeed cover a common underlying trait (such as mobility), thus forming a scale, which of the selected categories have a reasonable fit in relation to the assumed trait, and whether the selected categories cover the spectrum of ability one is likely to encounter in typical patients.

The objective of this paper is to examine whether clinical ratings of ICF categories can be integrated into parametric scales that provide a reliable estimation of specified patient problems and rehabilitation goals using the example of mobility of the upper and lower extremities in the acute hospital situation.

The specific aims were

1 to examine whether a subset of categories relevant for the specified problems and selected from the Acute ICF Core Set for Musculoskeletal Conditions constitutes a scale with Rasch properties invariant to selected person factors and
2 to identify which categories of the subset can and should be included in a clinical scale to enable simple and valid measurement in clinical practice.

2. Methods

2.1. Study design and patients

The data collection for this psychometrical study in patients with musculoskeletal conditions was carried out from January to October 2005 as part of a larger prospective cohort study in the acute hospital. Patients were recruited consecutively from the departments of orthopedics, rheumatology, and surgery of 20 acute hospitals across Switzerland. All patients provided written informed consent. The study was approved by the institutional ethics committee of each participating hospital.

2.2. Measures

The ICF is a multipurpose classification belonging to the WHO (World Health Organization) family of international classifications and is designed to record and organize a wide range of information about health and health-related domains (WHO). The ICF has two parts each divided into two components: Part one covers Body Functions (b), Body Structures (s), and Activities and Participation (d). Part two covers contextual factors including the components Environmental Factors (e) and Personal Factors. In the ICF Classification the letters b, s, d, and e, which refer to the components of the classification, are followed by a numeric code starting with the chapter number (one digit) followed by the second level (two digits), and the third and fourth level (one digit each). To give an example, the category Walking is a second level category with the code d450. This indicates that Walking belongs to the Activities and Participation component (d), and within this component is part of the chapter Walking and moving (d4). Other categories of this chapter are, for example, Moving around (d455) and Moving around using equipment (d465).

We collected information on the second level categories of the acute ICF Core Set for musculoskeletal conditions [15] and additional categories, which had been included because of their particular relevance for physical therapists [16].
Categories were graded with the qualifier 0 if a patient had no limitation and 1 if the patient had a limitation. If the information was not sufficient to allow grading the observation was set to missing. This approach has shown acceptable reliability [17]. Grading ICF categories yields a patient-specific categorical profile of limitation. These grades, however, do not form a single scale.

2.3. Data collection

Assessment was carried out at a minimum of 24 and a maximum of 48 h after admission for an acute event or after surgery. Patients were recruited and assessed by physical therapists trained in the application and principles of the ICF.

2.4. Statistical analysis

Analyses make use of the theory of the Rasch unidimensional measurement model [18]. This theory explains how items on a scale perform and how the tested persons’ ability relates to the properties of the scale. The Rasch model allows validating a scale with dichotomous or polytomous items in a way that it is possible to sum up single item values to a summary score, therefore, to calculate change scores and effect sizes in a statistically meaningful way.

The basic assumption of the Rasch model is that it is more likely to pass an easy item as compared to a difficult item, and that a person with high ability is more likely to pass an item than a person with low ability. The probability of passing an item is defined as a logistic function of the difference between person ability \( \theta \) and item difficulty \( b \).

\[
P_{ni} = \frac{\exp(\theta_n - b_i)}{1 + \exp(\theta_n - b_i)}
\]

\( P_{ni} \) is the probability that person \( n \) will pass an item \( i \), \( \theta \) denotes the person ability, \( b \) denotes the item difficulty in an interval scaled continuum, theoretically ranging from \(-\infty \) to \( +\infty \).

If the observed responses of an item set fit the Rasch model, that is, if easy items are passed by more patients and more able persons pass more items, then the item set is unidimensional and constitutes a Rasch scale. This can be verified by specific goodness-of-fit statistics. It can then be assumed that the items of the scale measure one single trait. This is confirmed by analysis of local independence of items. If the underlying single trait (the so called Rasch factor) is removed, there should be no residual associations left. Additionally, another important property to be examined is if a scale functions consistently across different subgroups of patients, that is, if an item is equally easy or difficult for men and women, for younger and older persons. Thus each item is examined for differential item functioning (DIF) across levels of person factors.

To show if there are subsets of an ICF Core Set, which are unidimensional, a sequence of analyses was carried out. Because the dimensionality of the ICF arguably is very complex and the goal of this study was to investigate whether it is possible to construct meaningful scales representing rehabilitation goals, we restricted our further analyses to ICF categories from the component Activities and Participation.

When first creating a scale, it is reasonable to choose items according to their inherent common concept. Thus we concentrated on two separate sets of categories, one set containing categories relevant to mobility of the upper extremities, a second set of categories relevant to mobility of the lower extremities. Table 1 shows the involved ICF categories and the two subsets created by the two a priori criteria. Additionally, the person factors diagnosis, age and sex were categorical covariates for further examination of DIF.

The two sets of categories were then fitted separately to the Rasch model. Goodness of fit was examined by overall fit statistics and individual item-fit statistics. We used three over-all fit statistics; the two item-person interaction statistics (for item and for person fit) are z-distributed, that is, a mean of zero and a standard deviation of 1 indicate perfect fit to the model, the item-trait interaction statistic has a \( \chi^2 \) distribution. Good fit would be indicated by a nonsignificant value of the item-trait interaction statistic, indicating that the rank of the items is constant irrespective of the level of the underlying trait. The individual item-fit statistics indicate how well an item fits into the scale. We reported fit residuals as the total of individual person and item deviations from the model that should lie within the range of \( \pm 2.5 \) and the \( \chi^2 \) statistic, which should be nonsignificant after adjustment for multiple testing.

To show if the scales function independently from person factors, we examined if at any given level of person ability the probability to pass an item is equal across strata of a personal factor. To put this into perspective, when examining the category walking (the item), a woman with a certain level of limitation in lower extremity mobility (the person ability) will have the same probability to be able to walk (to pass the item) as compared to a man with this level of limitation, if the Rasch assumptions hold. Thus, each item is examined for DIF across age and gender strata. The person-response residuals for each item show how much a person at a certain ability level departs from the response expected at this level of ability. The ANOVA statistic of the person factor and class intervals as factors should yield a nonsignificant result after appropriate adjustment for multiple testing. If there is uniform DIF of some of the items, that is, the detected difference in response across strata is the same regardless of the level of ability; those items can be adjusted to vary by person factor. In contrast, nonuniform DIF, that is, varying differences in response across strata, cannot be accounted for. In this study, we examined DIF for diagnosis, dividing the sample into patients with conditions affecting upper or lower extremities, spine, or hip, for age, dividing the sample into two age groups, and gender.
As a final test of unidimensionality, person-item deviation residuals were examined by principal components analysis [19].

A person separation index can be used to show how patients have been spread out along the scale defined by the items, thus indicating how well the scale differentiates between patients. The person separation index is close to 0 when all the persons are in a similar location and approaches 1 the more they spread across the item continuum. Significance level was set at 0.05. The Bonferroni method was applied as adjustment for multiple testing, yielding a significance level of 0.05/ \( k \) where \( k \) is the number of tests carried out simultaneously.

Rasch analysis was carried out with the RUMM2020 package (Perth, RUMM Laboratory), bivariate statistical analyses were carried out with SAS V9.1 (Cary, NC).

### 3. Results

A total of 234 patients with a mean age of 56 (median 58, range 18–91 years) were included in the analyses, 50% were female. Forty-four percent had a diagnosis involving mainly lower extremities such as knee replacement, rupture of tendons or fractures, 18% had a diagnosis involving the hip, mainly hip replacement, 18% had a diagnosis involving mainly upper extremities, and 16% had a diagnosis involving the spine. Prevalence of limitation for the 17 ICF categories of the component Activities and Participation is shown in Table 1.

The category most frequently limited was Lifting and carrying objects with 87% of the patients limited; the category least frequently limited was Family relationships with 6% of the patients limited.

#### 3.1. Scale for upper extremity mobility

Table 1 shows the 12 categories initially considered for the upper mobility scale. Of those, four were also considered for the lower mobility scale.

Almost all categories displayed insufficient fit to the Rasch model, resulting in highly significant item-trait interaction. There was uniform and nonuniform DIF by diagnosis, that is, the scale was not working the same across different health conditions. Consequently, we split the data set into three groups, patients with diagnoses involving upper extremities, patients with diagnoses involving lower extremities and patients with diagnoses involving the spine. These groups did not overlap. We then reanalyzed the scale within the group of persons with diagnoses involving only upper extremities. Category d445 Hand and arm use had to be excluded from analysis because all patients were limited in this category. Category d760 Family relationships had to be removed subsequently because of insufficient item fit. Mean item fit of the remaining 10 categories was \( \chi^2 = 25.053 \) at 30 degrees of freedom, \( P = 0.722 \). No category showed significant item-trait fit statistics (\( P = 0.003 \)), that is, there was no sign of misfit, and all of the item fit residuals were between –2.5 and 2.5. The category most likely to be limited was d510 Washing oneself, the category least likely to be limited d420 Transferring oneself (see Table 2 for the item statistics and locations).

To examine DIF we carried out two-way analyses of variance with age group and gender. There was neither uniform nor nonuniform DIF for age or gender. To put this into context, this indicates that item difficulty was stable irrespective of age or gender.
Principal components analysis showed that there were no residual associations, that is, no second factor after the Rasch factor was removed, giving evidence for good fit. Person-item locations are shown in Fig. 1. This indicates that the categories displayed a good spread over the full range of person ability without major floor or ceiling effects. Person separation index was 0.83, which is very good.

3.2. Scale for lower extremity mobility

Table 1 shows the 9 categories initially considered for the lower mobility scale. Of those, four were also considered for the upper mobility scale.

Almost all categories displayed insufficient fit to the Rasch model, resulting in highly significant item-trait interaction. As in the upper mobility categories, there was uniform and nonuniform DIF by diagnosis. Consequently, we split the data set and reanalyzed the scale within the group of persons with diagnoses involving only lower extremities and the hip. Category d460 Moving around in different locations had to be removed because of insufficient fit.

The remaining eight categories displayed a significant item-trait fit statistic. Category d420 Transferring oneself showed significant uniform DIF for age. We subsequently split the category for age.

Mean item fit of the resulting nine categories was $-0.587$ (SD 0.782), and person fit was $-0.383$ (SD 0.311). Item trait interaction indicated that the scale was invariant for patients at different levels of disability ($\chi^2 = 29.754$ at 27 degrees of freedom, $P = 0.325$). No category showed significant item-trait fit statistics ($P = 0.003$), that is, there was no sign of misfit, and all of the item fit residuals were between $-2.5$ and $2.5$. The category most likely to be limited was d450 Walking, the category least likely to be limited d760 Family relationships (see Table 3 for the item statistics and item locations).

Principal components analysis showed that there were no residual associations, that is, no second factor after the Rasch factor was removed, giving evidence for good fit.
Person-item locations are shown in Fig. 2. This indicates that the categories displayed a good spread over the full range of person ability without major floor or ceiling effects. Person separation index was 0.51, which is moderate.

4. Discussion

The result of our study is proof of concept for the development of problem tailored clinical measures by simply integrating clinical ratings of relevant ICF categories using the ICF qualifier. The illustrated process to develop clinical measures for the mobility of upper and lower extremity can potentially serve as a practical approach, which enables health professionals and researchers to develop parametric clinical measures tailored to the needs of their patients and services.

We found that the two scales were internally consistent, covered the whole spectrum of item difficulty and did not have any relevant ceiling or floor effect in the population studied. There was no relevant correlation between the values of the two indices further indicating that the two scales measure different concepts.

There have been several attempts to develop separate scales for upper and lower body mobility [20–22]. It may be due to inconsistencies that those separate scales performed not as well as they should as compared to a single functioning scale [23]. Developing measures based on the ICF Core Sets with the help of Modern Test Theory has several advantages. ICF Core Sets define first what to measure. They provide an exhaustive and conceptually sound choice of variables. This choice can be modified or enlarged according to the specific purpose, to meet the assessment needs and to define rehabilitation needs in the acute hospital. If data fit to the Rasch model, those ordinal categories can be added to a true interval scale or index, which is invariant to patient and observer characteristics and provides estimates of patient ability which are invariant to the items used. It seems possible to develop ICF-based measures by directly applying ICF categories with some simple specifications of the ICF categories. The direct

Table 3
Fit of categories to the Rasch model, location order, lower extremities/hip scale

<table>
<thead>
<tr>
<th>ICF code</th>
<th>Location</th>
<th>Fit Residual</th>
<th>Chi Square</th>
<th>Probability</th>
<th>% Limited</th>
</tr>
</thead>
<tbody>
<tr>
<td>d760</td>
<td>-4.737</td>
<td>-0.063</td>
<td>5.459</td>
<td>0.141</td>
<td>5.0</td>
</tr>
<tr>
<td>d240</td>
<td>-3.662</td>
<td>0.963</td>
<td>6.507</td>
<td>0.089</td>
<td>17.9</td>
</tr>
<tr>
<td>d420</td>
<td>-2.143</td>
<td>-0.582</td>
<td>1.320</td>
<td>0.724</td>
<td>30.6*</td>
</tr>
<tr>
<td>d420</td>
<td>-0.534</td>
<td>-1.461</td>
<td>3.933</td>
<td>0.269</td>
<td>71.1**</td>
</tr>
<tr>
<td>d465</td>
<td>0.112</td>
<td>-1.291</td>
<td>4.665</td>
<td>0.198</td>
<td>71.8</td>
</tr>
<tr>
<td>d415</td>
<td>1.242</td>
<td>-0.773</td>
<td>4.329</td>
<td>0.228</td>
<td>86.2</td>
</tr>
<tr>
<td>d4551</td>
<td>2.718</td>
<td>-0.327</td>
<td>1.568</td>
<td>0.667</td>
<td>94.4</td>
</tr>
<tr>
<td>d410</td>
<td>3.182</td>
<td>-1.004</td>
<td>1.119</td>
<td>0.773</td>
<td>93.2</td>
</tr>
<tr>
<td>d450</td>
<td>3.822</td>
<td>-0.747</td>
<td>0.855</td>
<td>0.836</td>
<td>96.7</td>
</tr>
</tbody>
</table>

These categories represent the finalized scale, ordered by difficulty of its items (family relationships was the category least probably to be limited, walking the category most probably to be limited).

*Within stratum of persons aged 55 and below.

**Within stratum of persons aged over 55.

Fig. 2. Person-item location distribution for the lower extremities scale. See Fig. 1 for interpretation.
qualification of ICF categories is the most representative and world wide applicable approach not requiring any other information than the globally available ICF taxonomy. This is particularly relevant for the application of the ICF for research and clinical practice in the developing countries where more sophisticated and self-administered measures generally are not available. Based on our results it seems possible not only to use categorical profiles in international large surveys but to come up with measurements of aspects of human functioning allowing for a more differentiated analysis.

Although it was not a goal of this paper to examine the dimensionality of ICF components, we found that it would not have been possible to create a single scale containing all categories of the ICF Activities and Participation component. Rasch analysis of the whole set of categories displayed severe misfit indicating more than one single underlying construct (data not shown). Two separate scales, however, proved to be consistent.

Location of the categories across the indices showed high face validity. On the scale for upper extremity mobility, Washing oneself was the item most likely limited, followed by Caring for body parts. On the scale for lower extremity mobility, Walking and Changing body position were the categories most likely to be limited.

Several limitations merit consideration. The two exemplary indices for the upper and lower extremity were not consistent across different health conditions. Thus, they had to be tested separately in patients with upper and lower extremity conditions. Sample size of patients with lower extremity conditions was sufficient [24]. Sample size for upper extremity conditions was relatively small which limited the power of the goodness-of-fit tests.

Although it may seem counterintuitive that walking turned out to be more difficult than climbing stairs, it has to be kept in mind that with 97% and 94% of patients limited in walking and climbing stairs, respectively, those two categories were close enough, thus, the property of the sample would not allow to truly differentiate the two.

In addition, stability of the scale, that is, item difficulty, over time should be further evaluated by examining the item fit at admission to and at discharge from hospital.

Although improved functioning arguably is the main issue of rehabilitation, the two scales proposed by this analysis do not take patient preferences into account. The current study is not able to answer the question if and how patients’ goals have been met. However, because there have been recent approaches to account for the methodological challenges of goal attainment scaling, our study is a case in point. There is the need for further proof that the direct ICF-based measurement, for example, using the ICF Core Set categories is generally possible across health conditions [13,26] and settings from the acute hospital to the community [14,27].

The development of ICF-based standards for what to measure and the development of ICF-based measures can be seen as communicating iterative process. Measures as shown are constructed based on the informed selection of potentially relevant categories. However, when studying the performance of these categories in the context of the measure that is being developed one can also learn much with respect to the further refinement of the ICF Core Sets. Because ICF Core Sets should, for reasons of practicability, be as short as possible without loosing on comprehensiveness, it is most helpful to identify redundant items and to identify the best category among several alternative categories. Similarly, when reducing the length of a Core Set it is important not to cut out a category that is positioned at one of the two ends of the measured spectrum or as the only category within certain range and therefore makes a major contribution to ability differentiation. Therefore, Rasch analysis is an important tool for the further development of ICF Core Sets.

The results of this study indicate that it could be possible to develop scales based on categories of the ICF Core Sets. This may be a promising approach for areas where psychologically sound measures are not available.

Acknowledgment

This study was supported by grants from the Friedrich-Baur-Foundation of the Ludwig Maximilian University Munich and the Swiss Physiotherapeutic Society.

We thank the contributing hospitals and the study coordination centre at the Institute of Physical Medicine, University Hospital Zurich, Switzerland, for data collection and support.

References


