ORIGINAL REPORT

TOWARDS THE SYSTEM-WIDE IMPLEMENTATION OF THE INTERNATIONAL CLASSIFICATION OF FUNCTIONING, DISABILITY AND HEALTH IN ROUTINE CLINICAL PRACTICE: EMPIRICAL FINDINGS OF A PILOT STUDY FROM MAINLAND CHINA

Jan D. Reinhardt, PhD[1,2,3#], Xia Zhang, MD, PhD[4,5#], Birgit Prodinger, PhD[2,3,6], Cristina Ehrmann-Bostan, PhD[2,3], Melissa Selb, MSc[2,6], Gerold Stucki, MD, MS[2,3,6] and Jianan Li, MD[4,5]

From the [1]Institute for Disaster Management and Reconstruction, Sichuan University and Hong Kong Polytechnic University, Chengdu, Sichuan, China, [2]Swiss Paraplegic Research, Nottwil, [3]Department of Health Sciences and Health Policy, University of Lucerne, Lucerne, Switzerland, [4]The First Affiliated Hospital of Nanjing Medical University, Nanjing, [5]Chinese Association of Rehabilitation Medicine, Beijing, China and [6]ICF Research Branch, a cooperation partner within the WHO Collaborating Centre for the Family of International Classifications in Germany (at DIMDI), Nottwil, Switzerland. [#]These two authors have contributed equally.

Objective: **The aims of this study were to evaluate the feasibility of using the International Classification of Functioning, Disability and Health (ICF) Generic Set in routine clinical practice, and of creating a functioning score based on it, and, subsequently, to examine its sensitivity to change.**

Methods: **In this prospective cohort study, data from 761 adult inpatients from 21 Chinese hospitals were analysed. Each patient was assessed at admission and discharge. Feasibility was evaluated by analysing mean assessment time. The Rasch model was used to create a metric of functioning. Sensitivity to change was analysed with mixed-effects regression and by calculating standardized effect size based on Cohen's f^2.**

Results: **Mean duration of assessment was 5.3 min, with a significant decrease between admission and discharge. After removal of the item remunerative employment, the remaining ICF Generic Set categories fitted the Rasch model well. With a mean improvement in functioning of 12.1 (95% confidence interval (95% CI): 11.5–12.6), this metric proved sensitive to change, both in terms of statistical significance ($p<0.001$) and standardized effect size (Cohen's f^2=2.35).**

Discussion: **The ICF Generic Set is feasible for use in routine clinical practice and is promising to serve as the basis for the development of a functioning score that is sensitive to change.**

Key words: functional status; sensitivity to change; psychometrics; Rasch analysis; ICF.

J Rehabil Med 2016; 48: 515–521

Correspondence address: Jianan Li, The First Affiliated Hospital of Nanjing Medical University, 300 Guangzhou Road, 210029 Nanjing, Jiangsu Province, China. E-mail: lijianan@carm.org.cn

Accepted Apr 22, 2016; Epub ahead of print May 13, 2016

INTRODUCTION

Information about people's functioning provides insights into body structures and functions, as well as how living with a health condition plays out in real life situations. Hence, functioning provides a comprehensive foundation for understanding a health condition and its impact on daily life (1, 2) and adds value to diagnostic information, e.g. in predicting length of stay in a hospital, service utilization and reimbursement (3–5). Therefore, functioning information is complementary to disease-specific information, and needs to be routinely available for operational and strategic decision-making in clinical practice, for health services management and planning, resource allocation and reimbursement, as well as for policies and programme development (6, 7).

The rehabilitation quality control system of the People's Republic of China is one of the main research priorities identified by the National Health and Family Planning Commission (previously the Health Ministry). The main objective of this quality control system is to monitor improvement in people's functioning as the primary outcome of rehabilitation services (8). However, to date, there is no commonly agreed assessment instrument to measure changes in functioning across different health conditions, contexts and populations.

To strengthen the comparability of functioning information, a framework is needed that provides a universal language of functioning, which is commonly understood across professionals and settings, and is aetiologically neutral in order to facilitate comparisons across health conditions. The International Classification of Functioning, Disability and Health (ICF) has been released by the World Health Organization (WHO) and endorsed by all its member states as the standard for describing and measuring health and functioning (9). The ICF contains an exhaustive and mutually exclusive list of categories and provides a universal language to describe functioning, is aetiologically neutral (10) and complementary to health condition specific information (11). Functioning constitutes the operationalization of health (1). To enhance the practicability of the ICF's complex and comprehensive classification scheme, sets of the most relevant ICF categories, i.e. selections of categories from the whole ICF classification to be reported for specific health conditions have been developed, based on multi-method research and international consensus processes

© 2016 The Authors. doi: 10.2340/16501977-2102
Journal Compilation © 2016 Foundation of Rehabilitation Information. ISSN 1650-1977

(12). Furthermore, an ICF Generic Set of 7 ICF categories that best described functioning and health in the general and clinical (sub-)populations was developed (13).

The ICF Generic Set serves as the minimal standard for assessing functioning in clinical practice and population-based health surveys, as well as for monitoring the impact of interventions at the clinical, service, and public health level. While the ICF Generic Set has the potential to serve as the starting point for developing a metric of functioning suitable for comparing information across the general population and clinical sub-populations, its utility for system-wide implementation in routine clinical practice needs to be examined.

The objective of this study was to evaluate the usefulness of the ICF Generic Set to assess functioning in routine clinical practice based on a pilot study in Mainland China. The specific aims were: (*i*) to evaluate the feasibility of using the ICF Generic Set in routine clinical practice; (*ii*) to identify whether it is possible to aggregate information across categories contained in the ICF Generic Set into a functioning score; and (*iii*) to examine its sensitivity to change.

METHODS

Study design and setting

This is a prospective cohort study. Rehabilitation departments of 21 provincial level hospitals located in major cities of 11 different regions of Mainland China administered the ICF Generic Set to patients admitted between 20 May and 30 June 2013 at admission and discharge. The study was approved by the ethics review board of Nanjing Medical University.

Participants

Patients with different health conditions requiring physical rehabilitation were recruited for this study. Based on their International Classification of Diseases (ICD-10) diagnosis at admission, patients were assigned to 13 different health condition groups: bone and joint disease, brain tumour, cerebral palsy, fracture of extremity, fracture of trunk, limb dysfunction, muscle disease and pain, nervous system disease, peripheral nerve injury, spinal cord injury, other spondylopathies, stroke, and traumatic brain injury. For multivariable analysis, these groups were further collapsed as considered clinically meaningful: (*i*) a musculoskeletal health condition group including, e.g. patients with limb dysfunctions or bone and joint diseases; (*ii*) a neurological health condition group including, e.g. patients with brain tumour or stroke; and (*iii*) spinal cord injury (SCI) and traumatic brain injury (TBI). Included were patients with definite medical diagnosis, who were admitted to the rehabilitation departments of 1 of the study hospitals and who had provided written informed consent. Originally, 994 patients were recruited for this study. From those, children and patients with missing admission or discharge data were excluded. This study includes 761 adult patients for whom complete data at both admission and discharge was available.

Measures and procedures

The ICF Generic Set consists of 7 ICF categories: *Energy and drive functions* (b130), *Emotional functions* (b152), *Sensation of pain* (b280), *Carrying out daily routine* (d230), *Walking* (d450), *Moving around* (d455), and *Remunerative employment* (d850) and was administered as a clinical measure using the generic ICF qualifier as a rating scale. The response options were: 0=no problem, 1=mild problem, 2=moderate problem, 3=severe problem, and 4=complete problem. Health professionals received online training in administering the ICF Generic Set. This online training took 4 h and consisted of an introduction to the ICF model, classification and qualifiers, as well as an introduction to the ICF Generic Set, the study design and aims. Each patient was evaluated by the same health professional at admission and discharge and the duration of the assessment in minutes was also recorded. Assessments were carried out at most 2 days after admission and 2 days before discharge, respectively.

Data analysis

Feasibility of using the ICF Generic Set in routine clinical practice. The mean duration of the assessment, in minutes, was analysed at admission and discharge. To analyse the difference in duration between discharge, as well as the amount of variance due to nesting of the measurements for patients in different evaluators and hospitals, we fitted a mixed-effects model featuring random intercepts for hospitals, evaluators, and patients and calculated intraclass correlation coefficients (ICC). The above model showed superior fit compared with all nested models based on likelihood ratio tests.

Aggregation of information across categories contained in the ICF Generic Set into a functioning score. To identify whether the total score over all categories contained in the ICF Generic Set provided a valid and objective measure of functioning, we treated the ICF categories as items and tested the fit of these items to the Rasch model (14, 15). In the context of the Rasch model "item" is a commonly used term. In the context of the metric of functioning in this study, an ICF category employed together with the ICF qualifier as a rating scale is considered an item. The Rasch model is a probabilistic model that builds upon the assumptions of local independence, unidimensionality, and invariance. Whether the data meets these assumptions is tested in an iterative process. Both item difficulty and person ability are located on the latent functioning trait. The ICF category Remunerative employment (d850) was removed from the attempt to create a functioning summated score for several reasons. Information about the extent of the problem people have in participating in remunerative employment cannot be assessed, but only inferred when a person is in a hospital. Hence, it is more meaningful to record the employment status of a person. As employment status would not change during a hospital stay given the Chinese health and social security system, the information would not be sensitive to change, thus the research team agreed on excluding this ICF category from further analysis.

To avoid dependency of the data as a result of repeated measurements, we selected 2 random samples of patients from admission and discharge, so that each patient was considered only once in each of the data-sets, while ensuring that the time-points were equally represented. This selection allowed for a cross-validation of the results of the Rasch analysis. The data-sets are referred to below as Sample A, the development sample, and Sample B, the validation sample.

Since no prior information exists on the factor structure of the ICF Generic Set, an exploratory bifactor analysis followed by a confirmatory bifactor analysis (CFA) on the polychoric correlation matrix was carried out to identify the factor structure for subsequent consideration in the Rasch analysis. Exploratory bifactor analysis assumes the presence of a single general factor and multiple independent specific factors and no specification on which factor the items should load. The number of factors considered in the bifactor analysis was determined based on permuted parallel analysis (16, 17). In CFA, root-mean-square error of approximation (RMSEA) <0.10, the comparative fit index (CFI) and the Tucker Lewis Index (TLI) >0.95, and χ^2 fit statistic of $p>0.05$ (non-significant) were used as criteria for accepting the inclusion of an item (18). The CFA was undertaken with R version 3.1.2 (19) and Rasch analysis with RUMM2030 (20).The Partial Credit Model (PCM) was chosen after a likelihood ratio test was performed with the output of the initial analyses to identify which version of the polytomous Rasch model (Rating Scale or Partial-Credit) was appropriate (21, 22). For each item the so-called item location was obtained, i.e. the overall difficulty of the item on the same scale. In addition, item thresholds, i.e. equal probability points between 2 adjacent response options, were estimated for each item. Thresholds should be ordered to be interpretable. RUMM2030 uses pairwise conditional estimation of item parameters. Person parameters are calculated given the item parameter estimates, using weighted maximum likelihood (20). Items with significant individual item χ^2 probability values at the overall significance level of 0.05 (and Bonferroni correction for the number of items) indicate misfit to the Rasch model. The overall fit of the data to the Rasch model was checked by the overall χ^2 of the items (22, 23). Item misfit can be influenced by local dependency, multidimensionality, and differential item functioning (DIF).

Local dependency is an assumption of the Rasch model, which was tested with Yen's Q_3 statistic. Q_3 is the correlation between item residuals of the Rasch analysis (24). The parametric bootstrapping procedure implemented by Christensen et al. was used to calculate the critical value for Yen's $Q_{3,*}$ (difference between Q_3 and the mean correlation) (25). Local dependent items were combined into a testlet, which is basically a super-item combining the locally dependent items (26). Under the testlet design, the threshold order is no longer expected (27).

Unidimensionality was assessed by comparing, using *t*-tests, the persons' abilities estimated separately for the items with positive vs negative loadings on the first residual component from the PCA. The hypothesis of unidimensionality is rejected if the number of significant *t*-tests is significantly larger than 5%. If this is not the case the analysis supports the assumption of unidimensionality (28).

Invariance is another assumption of the Rasch model, which was examined by testing for DIF across age groups (above or below 53 years), genders, health conditions groups (bone and joint disorders, neurological diseases, and SCI or TBI) and time of assessment (admission vs discharge) with analysis of variance (ANOVA) tests based on an overall significance level of 0.05 and Bonferroni correction. Items demonstrating DIF were split into specific questions for each of the groups showing DIF. A final PCM with all the split and non-split items was re-estimated.

Measurement quality was checked by examining the targeting of the functioning scale to the samples for both admission and discharge assessments. In brief, an examination of the spread of persons and items locations will inform us if items are covering the areas on the calibrated scale measuring the ability of the persons (29). Reliability was studied with the person separation index (PSI) from the Rasch analysis, which is similar to reliability as defined in classical test theory, except that it is based on person parameters from the Rasch model rather than the total raw score over all items. The value of PSI depends on both the study population and measurement errors. Values of 0.70 or greater are considered adequate at the group level (23, 30).

After confirming the findings of the development sample in the validation sample, a user-friendly scale from 0 to 100 was created for each sample A and B.

Sensitivity to change. Persons abilities at all time-points were estimated using the item estimates of the validation sample (31). Ability estimates were then transformed to a scale ranging from 0 (no problem) to 100 (complete problem). First, unadjusted mean scores and 95% confidence intervals (95% CI) were compared between admission and discharge. Secondly, a linear mixed-effects model featuring random intercepts for subjects and hospitals and thus accounting for the clustered structure of the data was calculated to predict functioning scores by time-point, unadjusted and adjusted for demographics and diagnosis. In a series of likelihood ratio tests comparing all possible nested models the above model showed superior fit. As residuals were not normally distributed we used robust standard errors based on the Huber-White Sandwich Estimator. Thirdly, responsiveness was assessed by calculating the local effect size of time of assessment (admission vs discharge) based on Cohen's f^2 (32, 33), which is an appropriate measure of effect size for mixed models and reflects the proportion of additional residual variance explained by the independent variable in question (i.e. time of assessment). Cohen's f^2 compares the variance explained by the full model, i.e. the one including time-point, as opposed to a null model with a constant and random effects only (R^2_{ab}) with the variance explained by a model without time (R^2_a). Random effects of the null model and the model without time-point are fixed at values of the full model. Cohen's f^2 is then calculated by applying the following formula:

$$f^2 = \frac{R^2_{ab} - R^2_a}{1 - R^2_{ab}}$$

Values between 0.15 and 0.34 are considered moderate and values above 0.35 are considered large by convention (32, 33). The statistical analyses other than Rasch analysis was performed with Stata 13 (Stata Corporation, TX, USA).

RESULTS

Sample descriptive information is provided in Table I. Almost two-thirds of the patients were male and most were 50 years or older. Stroke was the most common diagnosis, followed by SCI, other spondylopathies, bone and joint disease, and TBI. The mean length of hospital stay was 18 days (standard deviation (SD) 8.2 days). The longest mean length of stay was found for patients with TBI (23 days, SD 6.5) and shortest for patients with bone and joint diseases (14 days, SD 6.5). Table II shows the distribution of response options for individual ICF categories of the ICF Generic Set at admission and discharge, while Fig. 1 shows a heat map of functioning profiles (medians) across diagnostic groups at admission.

Feasibility of using the ICF Generic Set in routine clinical practice

Over all timepoints the mean duration of assessment was about 5 min and 30 s (SD 4.2; min–max: 1–30 min). With 6.2 min (SD 4.7; min–max: 1–30 min) the mean duration of the assessment was higher at admission than at discharge (4.9 min; SD 3.6; min–max: 1–25 min). According to the mixed-effects model the difference in duration between admission and discharge was significant at $p<0.001$. ICC analysis showed that 15% of the variance in duration was due to different hospitals, 41% due to different evaluators, and 28% due to nesting of measurements in different patients.

Table I. *Descriptive information on sample demographics, and diagnostic groups at admission (*n *= 761)*

Variable	Distribution
Gender, % (*n*)	
Male	65.3 (497)
Female	34.7 (264)
Age, mean (SD)	53.49 (16.34)
18–29 years, % (*n*)	9.6 (73)
30–49 years, % (*n*)	29.2 (222)
50–64 years, % (*n*)	34.3 (261)
65–90 years, % (*n*)	26.9 (205)
Diagnostic group, % (*n*)	
Bone and joint disease	4.6 (35)
Brain tumour	1.7 (13)
Cerebral palsy	0.9 (7)
Fracture (extremity)	2.8 (21)
Fracture (trunk)	1.7 (13)
Limb dysfunction	1.8 (14)
Muscle disease and pain	1.6 (12)
Nervous system disease	2.9 (22)
Peripheral nerve injury	1.6 (12)
Spinal cord injury	12.1 (92)
Other spondylopathies	7.4 (56)
Stroke (hemiplegia)	56.1 (427)
Traumatic brain injury	3.7 (28)
Others	1.2 (9)

SD: standard deviation.

Table II. *Distribution of response options and mean item scores at admission and discharge*

ICF item		Time[a]	No problem *n* (%)	Mild problem *n* (%)	Moderate problem *n* (%)	Severe problem *n* (%)	Complete problem *n* (%)	*p*-value[b]	Mean (Median)
b130	Energy and drive	1	105 (13.80)	145 (19.05)	170 (22.34)	234 (30.75)	107 (14.06)	$p<0.05$	2.12 (2)
	functions	2	184 (24.18)	238 (31.27)	186 (24.44)	104 (13.67)	49 (6.44)		1.47 (1)
b152	Emotional	1	267 (35.13)	256 (33.68)	163 (21.45)	51 (6.71)	23 (3.03)	$p<0.05$	1.09 (1)
	functions	2	382 (50.26)	251 (33.03)	94 (12.37)	19 (2.50)	14 (1.84)		0.73 (1)
b280	Sensation of pain	1	245 (32.24)	189 (24.87)	183 (24.08)	127 (16.71)	16 (2.11)	$p<0.05$	1.32 (1)
		2	350 (46.05)	283 (37.24)	106 (13.95)	16 (2.11)	5 (0.66)		0.74 (1)
d230	Carrying out	1	44 (5.78)	106 (13.93)	163 (21.42)	246 (32.33)	202 (26.54)	$p<0.05$	2.60 (3)
	daily routine	2	99 (13.01)	178 (23.39)	195 (25.62)	175 (23.00)	114 (14.98)		2.04 (2)
d450	Walking	1	122 (16.03)	85 (11.17)	94 (12.35)	125 (16.43)	335 (44.02)	$p<0.05$	2.61 (3)
		2	184 (22.08)	126 (16.56)	107 (14.06)	128 (16.82)	216 (28.38)		2.09 (2)
d455	Moving around	1	122 (16.05)	63 (8.29)	82 (10.79)	105 (13.82)	388 (51.05)	$p<0.05$	2.76 (4)
		2	168 (22.11)	84 (11.05)	99 (13.03)	132 (17.37)	277 (36.45)		2.35 (3)
d850	Remunerative	1	148 (26.67)	29 (5.53)	16 (2.88)	33 (5.95)	329 (59.28)	$p=0.235$	2.66 (4)
	employment	2	156 (28.11)	32 (5.77)	16 (2.88)	51 (9.19)	300 (54.05)		2.55 (4)

[a] 1=Admission; 2=Discharge. [b] Marginal homogeneity test was used for indicating statistical significance between admission and discharge.
ICF: International Classification of Functioning, Disability and Health.

Aggregation of information across categories contained in the ICF Generic Set into a functioning score

For both samples A and B, the permuted parallel analysis indicated the presence of 2 factors. The bifactor analysis showed higher factor loadings on the general factor than on specific factors for all items (Sample A: range 0.49–0.90; Sample B: range 0.54 – 0.94), except for *Sensation of pain* (b280) (Sample A and B) and *Emotional functions* (b152) (Sample B). The CFA confirmed the results of exploratory analysis with low factor loading on the general factor for *Sensation of pain* (b280) (0.10 in both samples A and B). Both samples A and B achieved RMSEA <0.10, and CFI and TLI >0.95, but none produced a non-significant χ^2 fit (RMSEA$=0.07$, $p=0.001$ Sample A and RMSEA$=0.09$, $p<0.001$ Sample B). The initial Rasch analysis presented non-significant pairwise differing ability estimates for the lower limit of confidence intervals (Sample A: 5.26%, (3.67–6.84); Sample B: 4.47% (2.99–5.93)).

The DIF and local dependency assumptions were not met. More specifically, all items, except *Emotional functions* (b152) showed

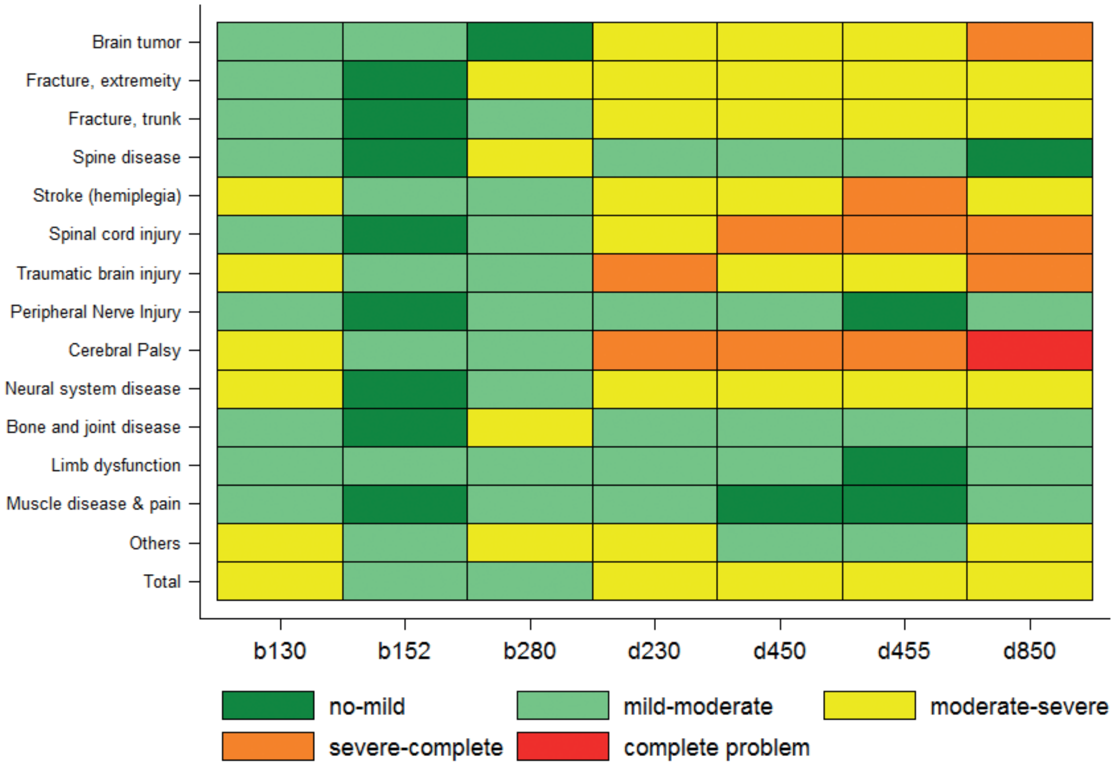

Fig. 1. Heat map of functioning profiles (medians) across diagnostic groups at admission. b130: energy and drive; b152: emotional functions; b280: sensations of pain; d230: managing daily routine; d450: walking; d455: moving around; d850: remunerative employment.

DIF for health conditions group. In addition, *Sensation of pain* (b280) showed DIF for gender and age groups. According to the critical value of 0.12 for Yen's Q_3, the following items showed local dependency in both samples A and B: *Energy and drive functions* (b130) and *Emotional functions* (b152), *Emotional functions* (b152) and *Sensation of pain* (b280), *Carrying out daily routine* (d230) and *Walking* (d450), *Walking* (d450) and *Moving around* (d455). Two testlets (super-items) were created: Body Functions testlet: *Energy and drive functions* (b130), *Emotional functions* (b152) and *Sensation of pain* (b280) and Activities and Participation testlet: *Carrying out daily routine* (d230), *Walking* (d450) and *Moving around* (d455). The testlet design showed unidimensionality in both samples A and B (Table III). The functioning scale showed good model fit after adjusting for DIF related to health condition group in the Activity and Participation testlet and related to time of assessment group in the Body Functions testlet. Table III shows item locations and fit statistics, the split strategies of the 2 testlets and the targeting of the scale. The reliability of the scale as indicated with the PSI was just below 0.7 for both samples A and B (Table III).

Sensations of pain (d280) showed the most significant DIF for health condition groups in both samples A and B. Therefore, we carried out an additional analysis where *Sensation of pain* (b280) was not included into the Body Functioning testlet. The overall fit statistic (Sample A: $\chi^2_{df=54}=106.48$, $p<0.001$; Sample B: $\chi^2_{df=54}=106.48$, $p<0.001$), individual item fits and the PSI (WITH extremes: 0.61 Sample A, 0.61 Sample B) showed poorer fit to the Rasch model.

A negligible floor effect occurred (at admission 2%, at discharge 5.6%).

Targeting of persons at admission and discharge in relation to the items is also shown in Fig. 2. When comparing the distribution of item thresholds with the persons' ability, functioning items did not discriminate well between persons with a very low/high level of difficulties.

Sensitivity to change

Mean scores of ability estimates from the above Rasch analysis transformed to a scale ranging from 0 to 100 were 52.9 (95% CI 52.1–53.8) at admission and 40.8 (95% CI 39.8–41.8) at discharge. This difference was significant at $p<0.001$ in the unadjusted mixed-effects model and the model adjusted for demographics and diagnostic groups (both models featuring random intercepts for hospitals and subjects). The mean improvement estimated by the adjusted model was 12.1 (95% CI 11.6–12.6). Cohen's f^2 was estimated as 2.35.

DISCUSSION

In this study the application of the ICF Generic Set for the collection of information about patients' functioning in routine clinical practice was evaluated for the first time. While the feasibility in terms of duration of the assessment was good on average, there was some variation, in particular due to different

Table III. *Individual item locations, and fit statistics, including targeting, unidimensionality, reliability, local dependency, and differential item functioning (DIF) for both samples A and B*

	DIF strategy		Sample A Individual item fit statistic				Sample B Individual item fit statistic			
Testlets	First DIF	Second DIF	Location	SE	FR	*p*-value	Location	SE	FR	*p*-value
Part A: Individual item location and fit statistics										
Body Functions testlet	Admission		0.395	0.029	1.104	0.217	0.476	0.030	1.031	0.694
(Energy and drive functions (b130), Emotional functions (b152), Sensations of pain (b280))	Discharge		1.228	0.034	1.476	0.576	0.712	0.033	1.347	0.914
Activities and Participation testlet	Bone and joint disorders		–0.078	0.029	–1.159	0.015	–0.075	0.031	–1.235	0.025
(Carrying out daily routine (d230),	Neurological diseases	Admission	–0.665	0.029	–0.341	0.021	–0.524	0.021	–0.234	0.004
Walking (d450), Moving around (d455))		Discharge	–0.267	0.032	–0.672	0.333				
	Spinal Cord Injury (SCI) or Traumatic Brain Injury (TBI)		–0.612	0.039	–0.672	0.367	–0.590	0.039	–0.480	0.247
Part B: Targeting, unidimensionality and overall fit statistic										
Item-trait interaction – χ^2										
Value			79.568				65.073			
df			54				45			
p-value			0.013				0.027			
Reliability – PSI, WITH extremes			0.659				0.645			
Items, mean (SD)			0.000 (0.715)				0.000 (0.584)			
Fit residual, mean (SD)			–0.036 (1.067)				0.086 (1.078)			
Persons, mean (SD)			–0.126 (0.726)				–0.129 (0.705)			
Unidimensionality, % (95% CI)			2.36 (1.28–3.44)				2.11 (1.08–3.12)			

SD: standard deviation; 95% CI: 95% confidence interval; SE: standard error of measurement; FR: fit residual; DIF: differential item functioning; PSI: person separation index.

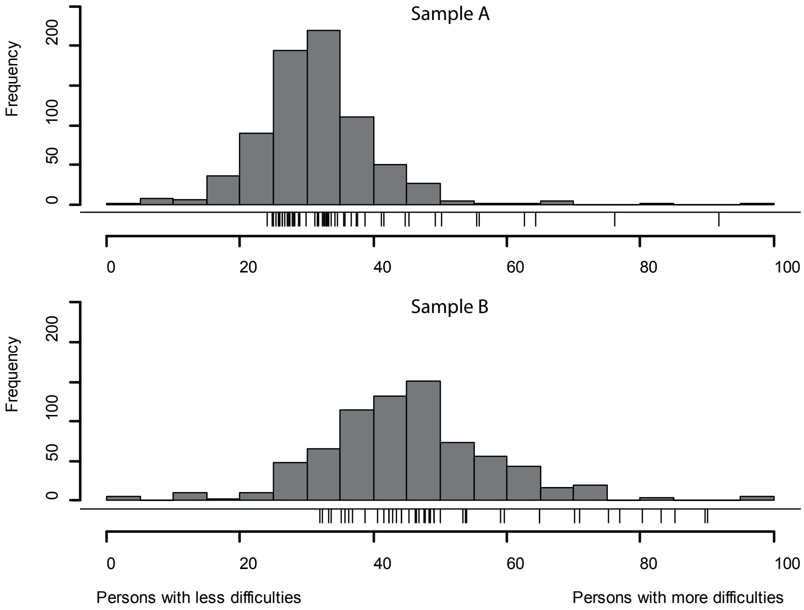

Fig. 2. Histogram of functioning abilities (*grey columns*) and item thresholds (*small vertical lines*) for both samples A and B.

evaluators. After applying a testlet design and considering DIF for the health conditions group and time of assessment, the ICF Generic Set categories fitted the Rasch model reasonably well for both samples A and B. The testlet approach proved to be helpful to deal with local response dependence. By creating 2 super-items, we were able to generate a total score of persons' functioning by summing up the initial response options of the ICF Generic Set. This metric proved sensitive to change due to rehabilitation treatment, both in terms of statistical significance as well as standardized effect size.

As the variation in the duration of assessment was due mainly to different evaluators, we assume that differing receptiveness of the online training may have played a role. This assumption is supported by our finding that assessment time was significantly reduced at discharge evaluation compared with admission, indicating a learning effect.

Sensation of pain (d280) showed low factor loading on the general factor in both exploratory and confirmatory analysis. This may be due to several reasons. Firstly, the local response dependence among items could affect the CFA loadings of some of the items, since CFA does not account for local response dependence. The problem may therefore be that the loadings of some of the other items are too strong and not that the loading of sensation of pain is too weak. Secondly, the item may have been viewed in terms of a symptom, i.e. the presence of an unpleasant feeling indicating actual or potential tissue damage (34) or as an impairment, i.e. the absence of the ability to feel pain, e.g. due to a sensory complete SCI. Thirdly, in contrast to the other items pain is subjective and cannot be assessed with "objective" methods. Fourthly, some patients may have received pain medication, leading to a reduction in covariance of pain and the other items. As this ICF category has been identified as one of the variables relevant in the collection of minimal information on functioning, it is important to collect the respective information and include it in the functioning score. This was supported by the unidimensionality test in the initial Rasch analysis. Further research is warranted to clarify whether sensation of pain really constitutes a different dimension or can be integrated in a metric of functioning, e.g. by improving instructions for assessment.

Furthermore, we did not include *Remunerative employment* (d850) in the Rasch model, since the category did not change over time. This may be due to the setting, as inpatient rehabilitation in the Chinese context rarely includes vocational rehabilitation. Nevertheless, this domain should be assessed as it occurs to be relevant, particularly with regard to community follow-up.

With an improvement of approximately 12 points between admission and discharge and a narrow confidence interval, the metric of functioning based on the 5 categories of the ICF Generic Set was highly responsive to change from a statistical point of view. However, future research is needed to establish a minimal clinically important difference (35), for instance to help determine the sample size of randomized controlled trials that aim to use functioning as a primary outcome.

Several limitations of our study need to be considered. The generalizability of our findings may be limited due to the short time interval during which this study was conducted. Specifically, this study may reflect selection bias, since complete data were solely available for patients with shorter lengths of stay. Furthermore, most of the patients had neurological conditions, which may limit generalizability to other diagnostic groups. Follow-up studies should consider a longer time interval to fully incorporate all types of cases in physical inpatient rehabilitation. Moreover, hospitals that participated in the study were mostly from well-developed urban areas of China, so that we are unable to draw conclusions for lower resource regions.

Despite the limitations of this study, we would also like to highlight some of the study's strengths. The study features a large sample involving 21 hospitals treating patients with different types of diagnoses. Adjusted models were estimated to accommodate this nested design and further compared with simpler models regarding model fit. While further research is warranted to address inter-rater reliability and concurrent validity with respect to established measures of functional health, such as the Barthel Index, the study design was tailored to answer our initial research questions.

In conclusion, the ICF Generic Set is promising for the collection of functioning information during routine clinical practice. The findings of this pilot study support its feasibility in the clinical setting and its utility as a potential basis for a metric score of functioning that is sensitive to change.

ACKNOWLEDGEMENTS

This study would not have been possible without the support of numerous persons and institutions. We would like to acknowledge the tremendous support of the National Health and Family Planning Commission of the People's Republic of China and the Chinese Association of Rehabilitation Medicine. We would also like to thank the participating hospitals and the clinical raters: Sheng Bi from China PLA General Hospital, Weijing Liao from Zhongnan Hospital of Wuhan University, Xiaoling Huang from Tongji Hospital of Huazhong University, Yi Wu from Shanghai Huashan Hospital, Changjie Zhang from Xiangya 2nd Hospital of Zhongnan University, Zulin Dou from the 3rd Affiliated Hospital of Sun Yat Sen University, Guozhi Huang from Guangzhou Zhujiang Hospital, Zhuoming Chen from the 1st Affiliated Hospital of Jinan University, Weiqun Song from Beijing Xuanwu Hospital, Lijuan Ao from the 2nd Affiliated Hospital of Kunming Medical University, Dongfeng Huang from the 1st Affiliated Hospital of SunYat Sen University, TiebinYan from the 2nd Affiliated Hospital of Sun Yat Sen University, Qiang Wang from the Affiliated Hospital of Qingdao University, Wenhua Chen from Shanghai 1st People's Hospital, Mouwang Zhou from the 3rd Affiliated Hospital of Beijing University, Chenqi He from West China Hospital of Sichuan University, Shouwei Yue from Qilu Hospital of Shandong University, Qing Xie from Shanghai Ruijin Hospital, Zhiqiang Zhang from Shengjing Hospital of China Medical University, Jianxian Wu from the 2nd Affiliated Hospital of Anhui Medical University, and Hui Ding and Xiao Lu from the 1st Affiliated Hospital of Nanjing Medical University.

REFERENCES

1. Stucki G, Kostanjsek N, Ustün B, Cieza A. ICF-based classification and measurement of functioning. Eur J Phys Rehab Med 2008; 44: 315.
2. Cieza A, Stucki G. Understanding functioning, disability, and health in rheumatoid arthritis: the basis for rehabilitation care. Curr Opin Rheumatol 2005; 17: 183–189.
3. Clohan DB, Durkin EM, Hammel J, Murray P, Whyte J, Dijkers M, et al. Postacute rehabilitation research and policy recommendations. Arch Phys Med Rehabil 2007; 88: 1535–1541.
4. Madden R, Marshall R, Race S. ICF and casemix models for healthcare funding: use of the WHO family of classifications to improve casemix. Disabil Rehabil 2013; 35: 1074–1077.
5. Hopfe M, Marshall R, Riewpabioon W, Tummers J, Kostanjsek N, Üstün B. Improving casemix systems by integrating functioning information. Proceedings of the WHO-FIC Annual Meeting; Cape Town, South Africa, 2011.
6. Kostanjsek N, Rubinelli S, Escorpizo R, Cieza A, Kennedy C, Selb M, et al. Assessing the impact of health conditions using the ICF. Disabil Rehabil 2011; 33: 1475–1482.
7. Kostanjsek N, Escorpizo R, Boonen A, Walsh NE, Üstün TB, Stucki G. Assessing the impact of musculoskeletal health conditions using the International Classification of Functioning, Disability and Health. Disabil Rehabil 2011; 33: 1281–1297.
8. Stucki G, Qiu Z, Li J, Li J, Wu X. Towards the system wide implementation of the ICF in rehabilitation in China. Chin J Rehabil Theory Pract 2011; 17: 5–10.
9. World Health Organization. International Classification of Functioning, Disability and Health: ICF. Geneva: World Health Organization; 2001.
10. Cieza A, Stucki G. The International Classification of Functioning Disability and Health: its development process and content validity. Eur J Phys Rehab Med 2008; 44: 303–313.
11. Escorpizo R, Kostanjsek N, Kennedy C, Robinson Nicol MM, Stucki G, Ustun TB. Harmonizing WHO's International Classification of Diseases (ICD) and International Classification of Functioning, Disability and Health (ICF): importance and methods to link disease and functioning. BMC Public Health 2013; 13: 742.
12. Selb M, Escorpizio R, Kostanjsek N, Stucki G, Ustün B, Cieza A. A guide on how to develop an international classification of functioning, disability and health core set. Eur J Phys Rehabil Med 2014; 51: 105–117.
13. Cieza A, Oberhauser C, Bickenbach J, Chatterji S, Stucki G. Towards a minimal generic set of domains of functioning and health. BMC Public Health 2014; 14: 218.
14. Fischer GH, Molenaar I, editors. Rasch models – foundations, recent developments, and applications. New York: Springer; 1995.
15. Christensen KB, Kreiner S, Mesbah M, editors. Rasch models in health. London, UK: Wiley; 2012.
16. Reise SP, Morizot J, Hays RD. The role of the bifactor model in resolving dimensionality issues in health outcomes measures. Qual Life Res 2007; 16: 19–31.
17. Jennrich RI, Bentler PM. Exploratory bi-factor analysis. Psychometrika 2011; 76: 537–549.
18. Brown TA. Confirmatory factor analysis for applied research. New York: Guilford Press; 2006.
19. R Development Core Team. R: a language and environment for statistical computing. Vienna; 2011. Available from: http://www.R-project.org.
20. Andrich D, Sheridan B, Luo G. Rasch models for measurement: RUMM2030. Perth, Western Australia; 2010.
21. Masters GN. A Rasch model for partial credit scoring. Psychometrika 1982; 47: 149–174.
22. Wright BD, Masters GN. Rating scale analysis. Chicago: MESA Press; 1982.
23. Andrich D. Rasch models for measurement. Sage university paper series on quantitative applications in the social sciences, series number 07/068. Newbury Park, California: Sage Publications; 1988.
24. Yen WM. Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. Appl Psychol Meas 1984; 8: 125–145.
25. Christensen KB, Makransky G, Horton M. Critical values for Yen's Q3: Identification of local dependence in the Rasch model using residual correlations. Copenhagen: University of Copenhagen, Department of Biostatistics; 2015.
26. Wainer H, Kiely GL. Item clusters and computer adaptive testing: a case for testlets. J Educ Meas 1987; 24: 185–210.
27. Andrich D. Item discrimination and Rasch-Andrich thresholds revisited. Rasch Meas Transact 2006; 20: 1055–1057.
28. Smith EV, Jr. Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. J Appl Meas 2002; 3: 205–231.
29. Bond TG, Fox CM. Applying the Rasch Model: fundamental measurement in the human sciences (2nd edition). Mahwah, New Jersey: Lawrence Erlbaum Associates; 2007.
30. Andrich D. An index of person separation in latent trait theory, the traditional KR. 20 index, and the Guttman scale response pattern. Educ Res Perspect 1982; 9: 95–104.
31. Mallinson T. Rasch Analysis of repeated measures. Rasch Meas Transact 2011; 25: 1317.
32. Cohen JE. Statistical power analysis for the behavioral sciences (2nd edn). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.; 1988.
33. Selya AS, Rose JS, Dierker LC, Hedeker D, Mermelstein RJ. A Practical guide to calculating Cohen's f(2), a measure of local effect size, from PROC MIXED. Front Psychol 2012; 3: 111.
34. Loeser JD, Treede RD. The Kyoto protocol of IASP basic pain terminology. Pain 2008; 137: 473–477.
35. Angst F, Aeschlimann A, Stucki G. Smallest detectable and minimal clinically important differences of rehabilitation intervention with their implications for required sample sizes using WOMAC and SF-36 quality of life measurement instruments in patients with osteoarthritis of the lower extremities. Arthritis Rheum 2001; 45: 384–391.