BMJ Open

# ICF Generic Set as new standard for the system wide assessment of functioning in China: a multicentre prospective study on metric properties and responsiveness applying item response theory

Cristina Ehrmann,[1,2,3] Birgit Prodinger,[1,2,3] Gerold Stucki,[1,2,3] Wenzhi Cai,[4] Xia Zhang,[5] Shan Liu,[4] Shouguo Liu,[6] Jianan Li,[6] Jan D Reinhardt[1,2,7]

**To cite:** Ehrmann C, Prodinger B, Stucki G, *et al*. ICF Generic Set as new standard for the system wide assessment of functioning in China: a multicentre prospective study on metric properties and responsiveness applying item response theory. *BMJ Open* 2018;**8**:e021696. doi:10.1136/bmjopen-2018-021696

► Prepublication history for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2018-021696).

Received 12 January 2018
Revised 18 October 2018
Accepted 25 October 2018



For numbered affiliations see end of article.

**Correspondence to**
Dr Jan D Reinhardt;
reinhardt@scu.edu.cn

## ABSTRACT

**Objectives** To examine metric properties and responsiveness of the International Classification of Functioning, Disability and Health (ICF) Generic Set when used in routine clinical practice to assess functioning.

**Design** Prospective multicentre study.

**Setting** 50 hospitals from 20 provinces of Mainland China.

**Participants** 4510 adult inpatients admitted to the departments of Pulmonology, Cardiology, Neurology, Orthopaedics, Cerebral Surgery or Rehabilitation Medicine.

**Main outcome measures** The ICF Generic Set (ICF Generic 6 Set) applied with an 11-point numeric rating scale (0-no problem to 10-complete problem) was fit to the Partial Credit Model (PCM) to create an interval score of functioning.

**Results** PCM assumptions were found to be fulfilled after accounting for Differential Item Functioning. With an average improvement by 7.86 points of the metric ICF Generic 6 score (95% CI 7.53 to 8.19), the ICF Generic 6 Set proved sensitive to change (Cohen's $f^2$=0.41). Ceiling and floor effects on detecting change in functioning were cancelled or reduced by using the metric score.

**Conclusion** The ICF Generic 6 Set can be used for the assessment of functioning in routine clinical practice and an interval score can be derived which is sensitive to change.

## Strengths and limitations of this study

- This study introduces a new International Classification of Functioning, Disability and Health (ICF)-based standard for collecting reliable information on patients' functioning in routine clinical practice in hospitals across China.
- The metric ICF Generic 6 Set score derived in this study can be used to compare functioning across health conditions, clinical departments, hospitals and over time.
- The non-random selection of hospitals in this study may, however, limit generalisability of the results.

## INTRODUCTION

For an optimal planning of treatments and documenting outcomes of interventions, diagnostic information should be complemented by information on functioning, that is physiological and mental functions, activities of daily living and participation in society.[1 2] As indicated by previous research, diagnosis alone cannot sufficiently predict relevant health outcomes such as hospitalisation,[3] length of stay,[4] social integration,[5] and mortality.[6 7] Information on functioning, in turn, has been demonstrated to be of added value in predicting those outcomes.[6–9] Moreover, interval scales of functioning can be used to quantify the impact of interventions within and across patient populations.[10]

The International Classification of Functioning, Disability and Health (ICF) is the reference for a systematic documentation of meaningful domains of functioning such as memory, pain, walking, self-care and social interactions, which are units of classification and called categories.[1] The list of more than 1400 ICF categories is organised into two parts, each with two components: Functioning and Disability with the components Body functions and structures (b and s) and Activities and participation(d) and Contextual factors with the components Environmental factors (e) and Personal factors. Personal factors have not yet been classified. The ICF categories are designated by the letters b, s, d and e, followed by a numeric code starting with the chapter number (first level, one digit), followed by the second level (two digits) and the third and fourth levels (one digit each). A detailed description of functioning using the ICF usually involves

the selection of second-level, third-level or fourth-level categories. In order to facilitate the assessment of functioning in clinical, research and other health-related settings, ICF Core Sets[11] and the ICF Generic Set have been developed.[12] While an ICF Core Set is a subset of the ICF codes for describing patient functioning in populations with a specific health condition or in a specific setting, the ICF Generic Set defines a minimum set of information on functioning that should be collected across health conditions and clinical settings as well as in the community.[12] The ICF categories of the Generic Set are: from the component Body Functions: (1) energy and drive functions (b130), (2) emotional functions (b152), (3) sensation of pain (b280) and from the component Activities and Participation: (4) carrying out daily routine (d230), (5) walking (d450), (6) moving around (d455) and (7) remunerative employment (d850). These ICF categories address four out of the eight World Health Survey domains of functioning. They were shown to be sufficiently explanatory for self-perceived health in general and clinical populations. The selection of these categories was based on a psychometric approach using data from three sources: (1) the German National Health Interview and Examination Survey 1998, (2) the United States National Health and Nutrition Examination Survey 2007/2008 and (3) the ICF Core Set studies.[12] The ICF Generic Set is aimed as a response to the challenge of creating a common metric of functioning to ensure comparability of data across studies and populations.[12] Data corresponding to this minimum set of functioning information can be generated with two different approaches: (1) mapping existing assessment tools of functioning to the ICF[13 14] and identifying items operationalising the functioning domains of the ICF Generic Set or (2) using the categories of the ICF Generic Set in combination with a rating scale as items.[15] Regarding the first approach, Oberhauser *et al* showed that a psychometrically sound metric can be developed for tracking and comparing functioning in people living in private households in England.[16] The second approach was first tested within a Chinese initiative to build a national ICF-based data system for evaluating and monitoring health systems performance.[14 15] In a pilot study, the seven categories of the ICF Generic Set were used in routine clinical practice to collect functioning information by rehabilitation professionals, using the generic ICF qualifier (a five point ordinal scale: 0 no problem, 1 mild problem, 2 moderate problem, 3 severe problem, 4 complete problem) as a rating scale. Reinhardt *et al* demonstrated the feasibility of the use of the ICF Generic Set in clinical practice with about 6 min assessment time on average and the possibility to aggregate information across categories of the ICF Generic Set into a functioning score that was sensitive to change during inpatient rehabilitation treatment.[17] However, results from above study and feedback from clinical raters also revealed several limitations that future studies would need to address: (1) clinical raters reported difficulties in applying the generic ICF qualifier scale, in particular with regard to differentiation between scale points, (2) the descriptions of the ICF categories were not always consistently understood, (3) remunerative employment had to be removed from the scale as clinicians found themselves unable to appraise this category in the inpatient setting and (4) the study was confined to the rehabilitation setting.[17]

To address the above issues, a large multicentre study was conducted as a follow-up. (1) Instead of the generic ICF qualifier scale, where each qualifier is defined, a numeric rating scale from 0 (no problem) to 10 (complete problem), where only the extremes are defined, was used and (2) clinically meaningful descriptions of ICF categories developed in a consensus conference were employed.[14] (3) Information about remunerative employment (d850) was collected but not included in creating the sum score of functioning in the inpatient setting. (4) The study was conducted across various clinical departments.

The objective of this paper was to examine the psychometric properties of the ICF Generic Set when used in routine clinical practice to assess functioning. The specific aims were (1) to identify whether it is possible to aggregate information across categories contained in the ICF Generic 6 Set and assessed on a 11-point numeric rating scale into a metric functioning score, (2) to examine the ICF Generic 6 Set's sensitivity to change and (3) to investigate ceiling and floor effects affecting the detection of the change.

## METHODS

### Study design and setting

This was a prospective multicentre study conducted from 5 November 2014 to 28 February 2015. Patients admitted to the departments of Pulmonology, Cardiology, Neurology, Orthopaedics, Cerebral Surgery or Rehabilitation Medicine were included in this study. Inclusion criteria were: (1) adults aged 18 years and older; (2) with definite medical diagnosis and (3) with complete data at admission and study endpoint (discharge, death, transfer or end of study period). Participating centres comprised Grade II and grade III hospitals from 20 provinces of Mainland China. Grade II and grade III refer to size and available resources of the hospitals, with grade III being Province level hospitals meeting highest medical standards and Grade II being smaller but still well-equipped City level hospitals. The study was presented at the annual conference of the Chinese Nursing Association in Guangzhou and partners from participating hospitals were recruited there as well as through personal networks of the authors. The study protocol was available to the participating hospitals.

The study was performed according to the principles of the Helsinki Declaration and informed written or verbal (in case of illiteracy) consent was obtained from all study participants. We received ethical approval for the analysis and publication of the data for research purposes from of Shenzhen Southern Medical

University, Guangzhou, China where the study centre was located and the data was hosted on 20 September 2017 (No. NYSZYYEC20170013).

### Study population

Patients with different health conditions admitted to the participating hospitals and departments within above specified timeframe were recruited for this study. Based on their International Classification of Diseases (ICD)-10 diagnosis at admission, patients were assigned to six different health condition groups: (1) musculoskeletal health condition group including patients with limb dysfunctions or bone and joint diseases, (2) neurological health condition group including patients with stroke, traumatic brain injury or cerebral apoplexy, (3) cancer health condition group, for example, patients with lung cancer or bone tumours, (4) cardiovascular health condition group, for example, patients with hypertension or coronary heart disease, (5) respiratory health condition group, for example, patients with pneumonia or bronchiectasis disease and (6) group comprising other health conditions that could not be classified into one of the above.

### Measures and procedures

Six out of seven categories of the ICF Generic Set (excluding d850-remunerative employment) were used by clinical nurses to assess patients functioning on an 11-point numeric rating scale (0 (no problem) to 10 (complete problem)) at admission and discharge or study endpoint. Each ICF category was accompanied by a simple, clinical intuitive description.[14] For example, d230—Carrying out daily routine refers to 'actions of planning, managing and completing activities of daily living' as opposed to the original ICF description 'Carrying out simple or complex and coordinated actions in order to, plan, manage and complete the requirements of day-to-day procedures or duties, such as budgeting time and making plans for separate activities during the day'. The patients were not involved in rating their functioning. In rating each category, the assessors considered all previous data routinely collected in the hospital department in question: information from anamnesis, clinical examinations, single item scales like visual analogue scale for pain or standardised assessment tools such as the Barthel Index. Nurses received formal training on how to assess functioning with the ICF Generic Set by the authors (JR, XZ, WC, SL). The functioning of each patient was assessed by the same trained nurse at the admission and the study endpoint. Mean time between assessments was about 13.5 days (SD: 9.1, Minimum: 1, Maximum: 70). Mean assessment time was about 9.1 min at admission (SD: 5.3, Minimum: 1, Maximum: 36) and 7.1 min at discharge or study endpoint (SD: 4.2; Minimum: 1, Maximum: 30). Demographic (gender, age) and diagnostic data (ICD-10) were extracted from hospitals' patient journals by the authors (JR, XZ, WC, SL).

### Patient and public involvement

As clinicians were to rate patients in ICF categories based on available routinely collected information, patients were not involved in the design of the study or recruitment procedures and conduct. They were, however, informed about the purpose of the study and informed consent was obtained. After academic publication, patients and the public will be informed about the results in patient magazines and through social media.

### Statistical analysis

#### Descriptive statistics

Descriptive statistics were used to describe the study population and response distributions. The marginal homogeneity test was used to test change in response patterns for each ICF category between admission and study endpoint.

#### Rasch analysis

A one-parameter item response model, also known as Rasch model, was used to test if a valid interval score of functioning could be derived by aggregating responses across ICF Generic Set items, that is, ICF-categories combined with the 11-point numeric rating scale.[18 19] Although the Rasch model requires more assumptions than non-parametric item response models for measuring persons and items, it offers high stability of model parameter estimates and person ability estimation.[20] The RUMM2030 package was used for carrying out the Rasch analysis.[21] Based on item parameters estimated using a pairwise conditional method, RUMM2030 calculates person parameters using weighted maximum likelihood. In addition, item thresholds, that is, equal probability points between two adjacent response options, were estimated for each item. Thresholds should be ordered to be interpretable since they are supposed to reflect an increase on the functioning trait.

The three assumptions of the Rasch model, that is, local dependency, unidimensionality and invariance, were iteratively tested. First, the unidimensionality assumption of items being homogeneous in the sense of measuring a single latent trait of functioning was tested using the principal component analysis (PCA) method proposed by Smith.[22] For each patient, two separate abilities were estimated from the Rasch calibration of the set of items with positive loadings and the Rasch calibration of the set of items with negative loadings on the first residual component from the PCA followed by pairwise t-tests. The number of significant t-tests should be below 5% to indicate unidimensionality. Second, the local independence assumption implying no relations between pairs of items and unbiased parameter estimates was tested. The presence of local dependence among items after accounting for the trait (residual correlations of items) may be an indicator for additional dimensions, which again would violate the unidimensionality assumption.[23] To this end, Yen's $Q_3$ statistic representing correlations between item residuals of the Rasch analysis were used.[24] The critical value for Yen's $Q_3$ statistic, $Q_3$*, that is, the

difference between $Q_3$ and the average correlation, was calculated based on the parametric bootstrapping procedure implemented by Christensen *et al*.[25] Testlets, that is, super items combining individual items, were created for locally dependent items to absorb local dependency and improve model fit.[26] The iterative process in the testlet design is the same as in single item design, except that under the testlet design the thresholds ordering is not expected.[27] Third, to assess Differential Item Functioning (DIF) across age groups (above or below the age mean, ie, 58 years), gender, health conditions groups (musculoskeletal, neurological, cancer, cardiovascular, respiratory and others) and time of assessment (admission vs study endpoint), analysis of variance (ANOVA) tests based on an overall significance level of 0.05 with Bonferroni correction for the number of items were carried out.[28] A significant main effect of the respective group variable indicates that subgroups respond in a systematically different way (indicated by parallel item characteristic curves). Items demonstrating DIF were split into specific questions for each of the levels in the groups showing DIF.

The Partial Credit Model (PCM) was chosen after a likelihood ratio test was performed with the output of the initial analysis to identify which version of the polytomous Rasch model (Rating Scale or Partial-Credit) was appropriate.[29 30] While the item fit was examined with individual item $\chi^2$ probability values, the overall fit of the data to the Rasch model was checked based on the global $\chi^2$ of the items.[30 31]

The targeting of the functioning scale with regard to the sample was studied by comparing the distribution of person and item locations.[32] Reliability was studied with the person separation index (PSI) from the Rasch analysis with an adequate expectation of 0.70 or above at the group level.[31 33]

Two stratified random samples, called development sample and validation sample, were selected across admission and study endpoint so that each person was represented only once while ensuring equal representation of the two time points. Health condition (six groups), age (younger than 58 years vs 58 years and older) and gender were used as criterion for stratification when selecting patients for each random sample, with equal representation of each subgroup. The subgroup of females older than 58 years suffering from cancer had the smallest number of 32 patients and thus defined the size of the other subgroups. Stratified random samples for development and validation thus comprised 768 patients (32*6*2*2) each.

After obtaining the final logit score from the Rasch model, a user-friendly scale from 0 to 100 and a transformation table were created allowing deriving scores for the overall sample at both admission and discharge or study endpoint.

### Sensitivity to change

We assumed that on average, patients' functioning scores should improve during clinical treatment from admission to study endpoint. As we had to account for repeated measurements as well as for clustering of patients in evaluators in hospitals, we used mixed effects regression with maximum likelihood method to estimate 95% CIs for the average change in patients between admission and study endpoint. We compared all nested models with each other using likelihood ratio tests. The fully nested model employing random intercepts for patients, evaluators and hospitals showed superior fit. Cohen's $f^2$ was used as a measure for standardised effect size with values above 0.15 considered moderate and those above 0.35 considered large.[17 34] We, moreover, conducted stratified analysis by health condition group.

### Effect of ceiling and floor effects at baseline on detection of change

We used boxplots for studying whether ceiling and floor effects may prevent the detection of change for patients corresponding to each of the baseline quintiles of the ICF Generic Set raw score in comparison with the interval ICF Generic 6 Set.[35] In addition, for both the ICF Generic 6 Set raw score and the interval ICF Generic 6 Set, an F-test (based on ANOVA) followed by Tukey's Honest Significance Difference (HSD) posthoc tests, when significant, were used for determining if average change in functioning differed significantly across groups of patients corresponding to different quintiles of the ICF Generic 6 Set baseline score.[36] For testing how much the transformation of raw scores into interval scores was linked with the presence of ceiling and floor effect on detecting change, eta-squared measures, $\eta^2$, were calculated (as an indicator of the association between the total variability in the change of functioning and patients corresponding to different baseline quantiles of the ICF Generic 6 Set raw score).[37]

## RESULTS

### Baseline characteristics of study participants

Table 1 shows descriptive characteristics of the 4510 adults patients considered in this analysis after excluding children, 308 adults with no defined medical diagnosis and 58 adults with incomplete data at admission and study endpoint. From the 4510 adult patients, more than half were male and the mean age was about 58 years. While musculoskeletal and neurological conditions were the most common diagnoses, cancer was the least common. A total of 915 patients underwent surgery during inpatient treatment (510 from the musculoskeletal, 54 from the cancer, 217 from the cardiovascular, 13 from the respiratory, 83 from neurological health condition group and 38 from the group comprising other health conditions).

The sample of patients for each region was determined by the number and grade of hospitals. Table 2 shows the distribution of response options, mean and median for individual categories of the ICF Generic 6 Set at admission and study endpoint.

**Table 1** Descriptive information on sample demographics, diagnostic groups, departments and provinces at admission (n=4510)

| Variable | Values | Distribution |
|---|---|---|
| Gender | | |
| | Male, % (N) | 58.9 (2656) |
| | Female, % (N) | 41.1 (1854) |
| Age | | |
| | Mean (SD) | 58.16 (16.9) |
| | 18 to 29 years, % (N) | 7.1 (320) |
| | 30 to 49 years, % (N) | 22.5 (1017) |
| | 50 to 64 years, % (N) | 31.2 (1404) |
| | 65 years and older, % (N) | 39.2 (1769) |
| Diagnostic group | | |
| | Cancer, % (N) | 3.7 (170) |
| | Cardiovascular, % (N) | 17.2 (777) |
| | Musculoskeletal, % (N) | 25.8 (1165) |
| | Neurological, % (N) | 31.3 (1412) |
| | Respiratory, % (N) | 15.8 (711) |
| | Other, % (N) | 7.0 (275) |
| Discharge type | Planned-discharge, % (N) | 70.1 (3161) |
| | Self-discharge, % (N) | 9.5 (427) |
| | Transferred to other hospital/ department, % (N) | 2.5 (114) |
| | Died, % (N) | 0.4 (19) |
| | Remained hospitalised, % (N) | 10.9 (490) |
| | Other, % (N) | 6.6 (299) |
| Department | Rehabilitation, % (N) | 20.6 (929) |
| | Neurology, % (N) | 15.7 (705) |
| | Cerebral surgery, % (N) | 7.8 (348) |
| | Orthopaedics, % (N) | 19.5 (880) |
| | Pneumology, % (N) | 18.2 (821) |
| | Cardiology, % (N) | 18.2 (827) |
| Provinces | | |
| | Guangdong, % (N) | 29.1 (1314) |
| | Fujian, % (N) | 2.2 (98) |
| | Sichuan, % (N) | 8.3 (374) |
| | Xianxi, % (N) | 4.3 (195) |
| | Jiangsu, % (N) | 4.1 (187) |
| | Shandong, % (N) | 7.0 (316) |
| | Hainan, % (N) | 4.2 (192) |
| | Anhui, % (N) | 4.2 (192) |
| | Zhejiang, % (N) | 2.5 (114) |
| | Ningxia, % (N) | 4.1 (184) |
| | Jilin, % (N) | 2.1 (99) |
| | Xinjiang, % (N) | 2.1 (92) |
| | Chongqing, % (N) | 3.7 (170) |
| | Guizhou, % (N) | 4.2 (189) |
| | Jiangxi, % (N) | 2.2 (102) |
| | Heilongjiang, % (N) | 5.4 (244) |
| | Yunnan, % (N) | 2.0 (94) |
| | Hubei, % (N) | 2.3 (107) |
| | Qinghai, % (N) | 2.2 (99) |
| | Shanghai, % (N) | 3.2 (148) |

## Rasch analysis

For both development sample (Sample A) and validation sample (Sample B), Body Functions items loaded positively on the first residual component from the PCA, while the Activities and Participation items loaded negatively. The ICF Generic 6 Set showed unidimensionality in both samples, as less than 5% of pairwise t-tests were statistically significant (Sample A: 4.23%, (2.75–5.72); Sample B: 3.61% (2.24–4.97)). The local independence assumption was not met in any of the samples. According to the critical value of 0.12 for Yen's $Q_3$*, the following items showed local dependence 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). Moreover, for both samples A and B, all items showed DIF for health condition group and the item *Sensation of pain* (b280) for time of assessment. None of the items showed DIF by gender or age group. For both samples, two testlets (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)) were created. After examining the testlet design for DIF, the Body Functions testlet showed DIF for time of assessment and the Activities and Participation testlet for health condition group. For the Activities and Participation testlet, the item characteristic curves of musculoskeletal and neurological disorders group were parallel with the item characteristic curves of respiratory and cardiovascular disorders group and of cancer and other disorders group. Therefore, the DIF for the health conditions groups was accommodated by splitting this testlet in three specific items for the musculoskeletal and neurological disorders group, the respiratory and cardiovascular disorders group and the cancer and other disorders group. Splitting the Activities and Participation testlet had a good effect on the items fit, but the DIF for the Body Functions testlet for time of assessment was still present. We, however, did not adjust this testlet for time of assessment DIF since this DIF was found inconsistent due to the small differences of this testlet mean locations between the time of assessment for all class intervals (below 0.5 logits). Item locations and fit statistics and the targeting

Table 2 Distribution of response options from 0 (no problem) to 10 (complete problem) and average and median item scores at admission and discharge

| ICF item | Time* | 0 N (%) | 1 N (%) | 2 N (%) | 3 N (%) | 4 N (%) | 5 N (%) | 6 N (%) | 7 N (%) | 8 N (%) | 9 N (%) | 10 N (%) | p value† | Mean (Median) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| b130 Energy and drive functions | 1 | 853 (18.91) | 326 (7.22) | 594 (13.17) | 577 (12.79) | 383 (8.49) | 440 (9.75) | 364 (8.07) | 233 (5.16) | 401 (8.89) | 161 (3.56) | 178 (3.94) | <0.001 | 3.82 (3) |
| | 2 | 1533 (33.99) | 715 (15.85) | 761 (16.87) | 514 (11.39) | 260 (5.76) | 246 (5.45) | 150 (3.32) | 103 (2.28) | 107 (2.37) | 47 (1.04) | 74 (1.64) | | 2.14 (2) |
| b152 Emotional functions | 1 | 994 (22.03) | 410 (9.09) | 727 (16.12) | 591 (13.10) | 419 (9.29) | 394 (8.73) | 331 (7.34) | 194 (4.30) | 270 (5.98) | 85 (1.88) | 95 (2.10) | <0.001 | 3.21 (3) |
| | 2 | 1827 (40.51) | 736 (16.31) | 746 (16.54) | 436 (9.66) | 223 (4.94) | 196 (4.34) | 128 (2.83) | 70 (1.55) | 76 (1.68) | 33 (0.73) | 39 (0.86) | | 1.76 (1) |
| b280 Sensation of pain | 1 | 1320 (29.26) | 254 (5.63) | 514 (11.39) | 504 (11.17) | 441 (9.77) | 457 (10.13) | 329 (7.29) | 216 (4.78) | 251 (5.56) | 122 (2.70) | 102 (2.26) | <0.001 | 3.21 (3) |
| | 2 | 2125 (47.11) | 724 (16.05) | 671 (14.87) | 421 (9.33) | 205 (4.54) | 134 (2.97) | 81 (1.79) | 50 (1.10) | 58 (1.28) | 13 (0.29) | 28 (0.62) | | 1.44 (1) |
| d230 Carrying out daily routine | 1 | 843 (18.69) | 276 (6.11) | 440 (9.75) | 472 (10.46) | 346 (7.67) | 376 (8.33) | 332 (7.36) | 226 (5.01) | 451 (10) | 233 (5.16) | 515 (11.42) | <0.001 | 4.49 (4) |
| | 2 | 1285 (28.49) | 556 (12.32) | 605 (13.41) | 472 (10.46) | 299 (6.63) | 315 (6.98) | 247 (5.47) | 166 (3.68) | 201 (4.45) | 93 (2.06) | 271 (6.01) | | 3.05 (2) |
| d450 Walking | 1 | 1299 (28.80) | 271 (6.01) | 367 (8.13) | 394 (8.73) | 258 (5.72) | 257 (5.69) | 225 (4.98) | 155 (3.43) | 293 (6.49) | 149 (3.30) | 842 (18.67) | <0.001 | 4.22 (3) |
| | 2 | 1807 (40.06) | 455 (10.08) | 451 (10) | 337 (7.47) | 216 (4.79) | 218 (4.83) | 158 (3.50) | 108 (2.39) | 175 (3.88) | 78 (1.73) | 507 (11.24) | | 2.92 (1) |
| d455 Moving around | 1 | 930 (20.62) | 212 (4.70) | 296 (6.56) | 336 (7.45) | 279 (6.18) | 259 (5.74) | 185 (4.10) | 191 (4.23) | 348 (7.71) | 191 (4.23) | 1283 (28.45) | <0.001 | 5.32 (5) |
| | 2 | 1250 (27.71) | 380 (8.42) | 443 (9.82) | 335 (7.42) | 256 (5.67) | 243 (5.38) | 213 (4.72) | 148 (3.28) | 212 (4.70) | 122 (2.71) | 908 (20.13) | | 4.15 (3) |
| d850 Remunerative employment | 1 | 2487 (55.14) | 73 (1.62) | 116 (2.57) | 128 (2.83) | 130 (2.88) | 133 (2.94) | 151 (3.34) | 83 (1.84) | 170 (3.77) | 81 (1.79) | 954 (21.15) | <0.001 | 3.32 (0) |
| | 2 | 2610 (57.87) | 130 (2.88) | 188 (4.17) | 132 (2.92) | 131 (2.90) | 145 (3.21) | 173 (3.83) | 74 (1.64) | 109 (2.41) | 50 (1.10) | 764 (16.94) | 0.354 | 2.80 (0) |

*1=Admission; 2=Study endpoint.
†Marginal homogeneity test was used for indicating statistical significance between admission and discharge.

of the scale are shown in table 3. According to the PSI values, the reliability of the scale was just below 0.7 for both samples A and B (table 3).

Figure 1 illustrates the targeting of patients included in sample A and sample B as well as of the overall sample of 4510 patients at both admission and study endpoint. The functioning abilities for the overall sample were estimated using the item difficulties from the validation sample. When comparing the distribution of item thresholds with the persons' abilities, ICF Generic 6 Set items did not discriminate well between persons with a very low level of difficulties.

After fit to the Rasch model was achieved for the ICF Generic 6 Set, logit-scores were transformed into a user-friendly scale ranging from 0 (no problem) to 100 (complete problem) and a transformation table for total raw scores into an interval scale for use in parametric analyses was created (table 4).

### Sensitivity to change

Patients from all health condition groups apart from cancer showed improvement from admission to study endpoint (figure 2). Across all health conditions, patients improved by 7.86 points of the Rasch transformed overall score (95% CI 7.53 to 8.19). Effect size in terms of Cohen's $f^2$ was 0.41 (large). Average improvement for the musculoskeletal and neurological health condition group was 6.75 (95% CI 5.89 to 7.61) with a Cohen's $f^2$ of 0.37 (large) and 10.88 for cardiovascular and pulmonary diagnoses (95% CI 9.15 to 12.62) with a Cohen's $f^2$ of 0.63 (large). With 4.19 points (95% CI 2.51 to 5.88), the cancer and other health conditions group showed the smallest improvement over time which was also reflected in the lowest standardised effect size of 0.12 (low).

### Effect of ceiling and floor effects at baseline on detection of change

Both the box-plots and non-significant Tukey's HSD test for the interval ICF Generic 6 Set score showed that a floor effect present in detecting change when using the ICF Generic 6 Set raw score was cancelled when using the interval score for patients with musculoskeletal and neurological health conditions (figure 3) as well as respiratory and cardiovascular health conditions. For patients with cancer and other health conditions, a floor effect was present in both raw and interval scales (figure 3 and significant differences between patients corresponding to the first baseline quintile of the ICF Generic 6 Set raw score and patients corresponding to all other baseline quintile groups). For all health conditions groups, there was a significant difference between patients corresponding to the fifth baseline quintile of the ICF Generic 6 Set raw score and patients corresponding to the other baseline quintile groups, with a larger decrease of scores for the fifth quintile. However, the transformation of raw scores into the interval score reduced this ceiling effect as indicated by the respective eta-squared measures (musculoskeletal and neurological health conditions: $\eta^2$—raw

Table 3 Individual item locations and fit statistics, including targeting, unidimensionality, reliability, local dependency and DIF for both samples A and B for final solution

| Part A: Individual item location and fit statistics | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Sample A | | | | Sample B | | | |
| | | Individual item fit statistic | | | | Individual item fit statistic | | | |
| Testlets | DIF strategy | Location | SE | FR | P values | Location | SE | FR | P Values |
| *Body Functions testlet* Energy and drive functions (b130),Emotional functions (b152), Sensations of pain (b280) | | 0.224 | 0.009 | 0.839 | 0.027 | 0.235 | 0.010 | 1.180 | 0.033 |
| *Activities and Participation testlet* Carrying out daily routine (d230), Walking (d450), Moving around (d455) | Musculoskeletal and neurological disorders | −0.185 | 0.011 | −0.428 | 0.023 | −0.211 | 0.011 | −0.713 | 0.050 |
| | Respiratory and cardiovascular disorders | −0.056 | 0.015 | −1.078 | 0.026 | −0.017 | 0.015 | −1.358 | 0.033 |
| | Cancer and other disorders | 0.017 | 0.012 | −0.580 | 0.024 | −0.008 | 0.012 | −0.741 | 0.083 |
| **Part B: Targeting, unidimensionality and overall fit statistic** | | | | | | | | | |
| Item-trait interaction—$\chi^2$ | Value | 75.861 | | | | 68.4677 | | | |
| | df | 36 | | | | 36 | | | |
| | p-value | 0.02* | | | | 0.40* | | | |
| Reliability—PSI | WITH extremes | 0.67 | | | | 0.63 | | | |
| Items | Mean (SD) | 0.000 (0.171) | | | | 0.000 (0.182) | | | |
| Fit Residual | Mean (SD) | −0.311 (0.815) | | | | −0.408 (1.099) | | | |
| Persons | Mean (SD) | −0.269 (0.537) | | | | −0.254 (0.436) | | | |
| Unidimensionality | Percentage of significant t-tests (95% CI) | 4.24 (2.75 to 5.72) | | | | 3.61 (2.24 to 4.97) | | | |

*Conditional test of fit.
DIF, Differential Item Functioning; FR, fit residual; PSI, person separation index; SE, standard error of measurement.

scores=0.20, $\eta^2$—interval scores=0.08; respiratory and cardiovascular health conditions: $\eta^2$—raw scores=0.40, $\eta^2$—interval scores=0.02; cancer and other health conditions: $\eta^2$—raw scores=0.23, $\eta^2$—interval scores=0.04).

## DISCUSSION

This nationwide validation study demonstrated that the ICF Generic 6 Set in combination with an 11-point numeric rating scale can be used for the assessment of functioning in routine clinical practice and across a variety of hospital departments and health conditions. After accounting for local dependence of items by creating a body function and a activity and participation testlet and DIF across health condition groups unidimensional interval scores for three different health condition groups could be established for the ICF Generic 6 Set. The interval ICF Generic 6 Set score was sensitive to change with large standardised effect sizes (with the exception of the cancer and other health conditions group). We could also show that ceiling and floor effects in the detection of change were reduced or cancelled when transforming the raw scores into Rasch-based interval scores.

In the application of the PCM, the unidimensionality, local dependency and DIF assumptions were tested. Our results confirmed previous findings from our pilot study.[17] Irrespective of the type of rating scale used for the ICF Generic 6 Set, local dependency among Body Functions items and Activities and Participation items was present. The dependent sets of items identified based on the critical value for the Yen's Q3* statistic are also content-dependent items when following ICF. Thus, the fitting statistics from the PCM were better than those of the individual items.[38] In both studies, DIF for health conditions groups was found for all items of the ICF Generic 6 Set. After items were combined in two testlets, the DIF at the level of body functioning testlet disappeared. This could be explained by the heterogeneity of functioning of individual items nested within the testlet. Moreover, due to the existence of the testlet effect, the DIF for time of assessment amplified for this testlet.[39] The DIF between

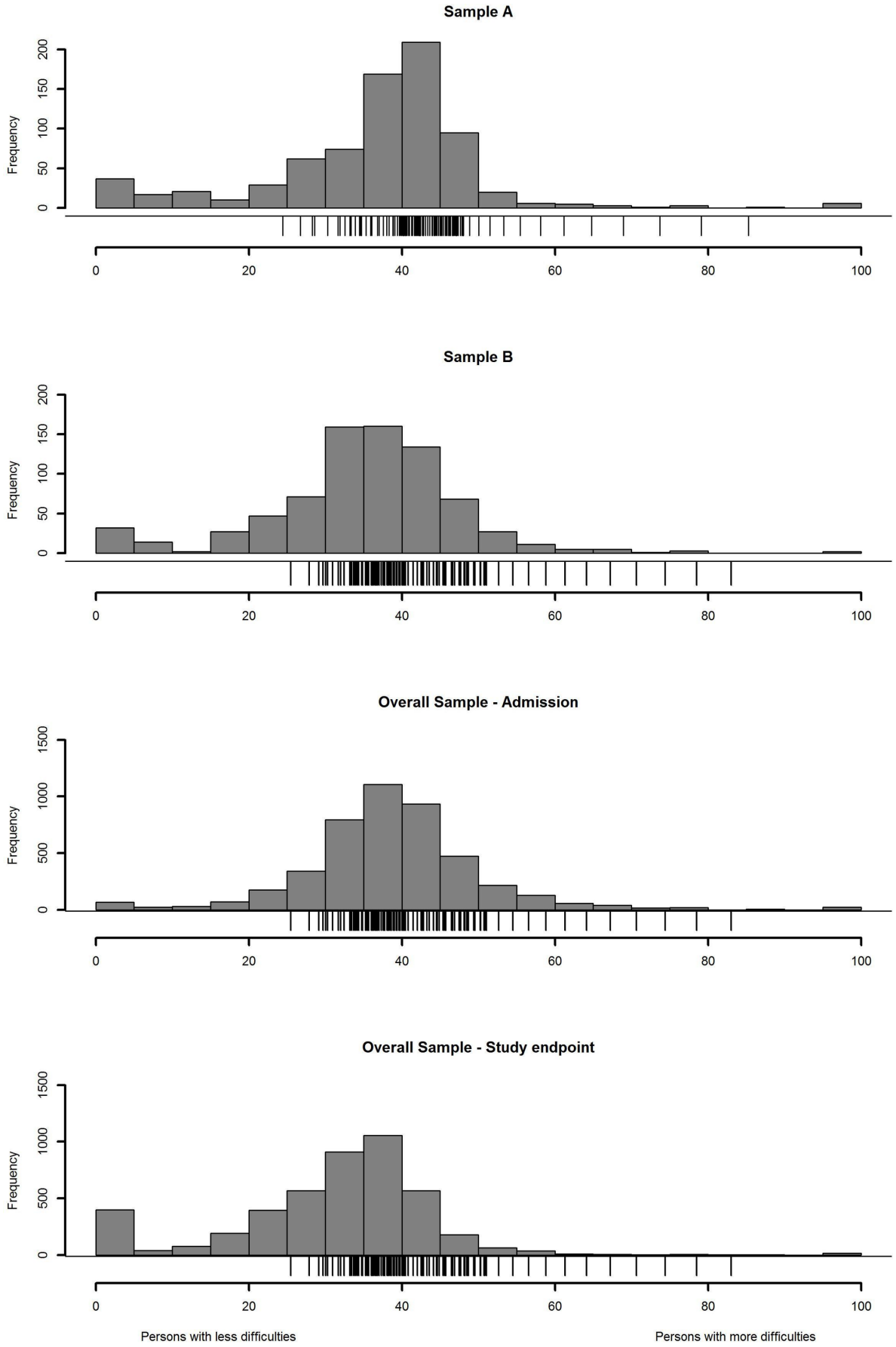


**Figure 1** Histogram of people's functioning (grey columns) and item thresholds (small vertical black lines) for both samples A and B and for overall sample at admission and study endpoint.

**Table 4** Transformation of the ICF Generic 6 Set score raw scores into interval ICF Generic 6 Set score ranging from 0 (no problem) to 100 (complete problem) by health condition groups

| Overall raw scores | Musculoskeletal and neurological disorders | Respiratory and cardiovascular disorders | Cancer and other disorders |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 1 | 11 | 15 | 17 |
| 2 | 18 | 21 | 23 |
| 3 | 22 | 24 | 26 |
| 4 | 24 | 26 | 29 |
| 5 | 26 | 28 | 30 |
| 6 | 28 | 29 | 31 |
| 7 | 29 | 30 | 32 |
| 8 | 30 | 31 | 33 |
| 9 | 31 | 32 | 34 |
| 10 | 32 | 32 | 35 |
| 11 | 32 | 33 | 35 |
| 12 | 33 | 34 | 36 |
| 13 | 34 | 34 | 36 |
| 14 | 34 | 35 | 37 |
| 15 | 35 | 35 | 37 |
| 16 | 35 | 36 | 38 |
| 17 | 36 | 36 | 38 |
| 18 | 36 | 37 | 39 |
| 19 | 37 | 38 | 39 |
| 20 | 37 | 38 | 40 |
| 21 | 37 | 39 | 40 |
| 22 | 38 | 39 | 40 |
| 23 | 38 | 40 | 41 |
| 24 | 38 | 40 | 41 |
| 25 | 38 | 41 | 42 |
| 26 | 39 | 41 | 42 |
| 27 | 39 | 42 | 43 |
| 28 | 39 | 43 | 43 |
| 29 | 40 | 43 | 44 |
| 30 | 40 | 44 | 44 |
| 31 | 40 | 44 | 44 |
| 32 | 41 | 45 | 45 |
| 33 | 41 | 45 | 45 |
| 34 | 41 | 46 | 46 |
| 35 | 42 | 47 | 46 |
| 36 | 42 | 47 | 46 |
| 37 | 42 | 48 | 47 |
| 38 | 43 | 48 | 47 |
| 39 | 43 | 49 | 48 |
| 40 | 44 | 50 | 48 |
| 41 | 45 | 50 | 49 |
| 42 | 45 | 51 | 49 |
| 43 | 46 | 52 | 50 |
| 44 | 47 | 52 | 50 |
| 45 | 48 | 53 | 51 |
| 46 | 49 | 54 | 52 |
| 47 | 51 | 55 | 52 |
| 48 | 52 | 56 | 53 |
| 49 | 54 | 57 | 55 |
| 50 | 55 | 58 | 56 |
| 51 | 57 | 60 | 58 |
| 52 | 59 | 61 | 60 |
| 53 | 61 | 63 | 62 |
| 54 | 64 | 66 | 64 |
| 55 | 67 | 68 | 67 |
| 56 | 70 | 72 | 70 |
| 57 | 74 | 75 | 74 |
| 58 | 79 | 80 | 79 |
| 59 | 86 | 88 | 87 |
| 60 | 100 | 100 | 100 |

health condition groups reflects the complexity of each health condition. This did not cause a major problem as we could statistically adjust for it and accounted for DIF by providing different transformation tables for three health condition groups.

With respect to the distribution of persons and items along the continuum of functioning of the ICF Generic 6 Set, the items did not completely match the expected patients' abilities at the lower end of the continuum. This finding is also reflected by the PSI value for both samples A and B. Although we used a heterogeneous sample in this study, the reliability was slightly better than in the pilot study.[17] Further research is needed as to whether this result is due to the use of the 11-point numeric rating scale in contrast to the ICF qualifiers.

While in the pilot study we split the Body Functions testlet for time of assessment group, in this study, we ignored this issue since the testlet showed good fit. In contrast to the pilot study, *Sensation of pain (b280)* fitted the Rasch model better. The DIF for time of assessment groups could be neglected as the testlet that included this item showed an overall good fit. However, *Sensation of pain* was clinician administered, and further research is necessary, also in other countries, to clarify its fit to the metric of functioning.[17] The patient's self-reported pain may be used since pain is subjective. While listed as a body function in WHO's ICF and ICD-11, it may be debated if

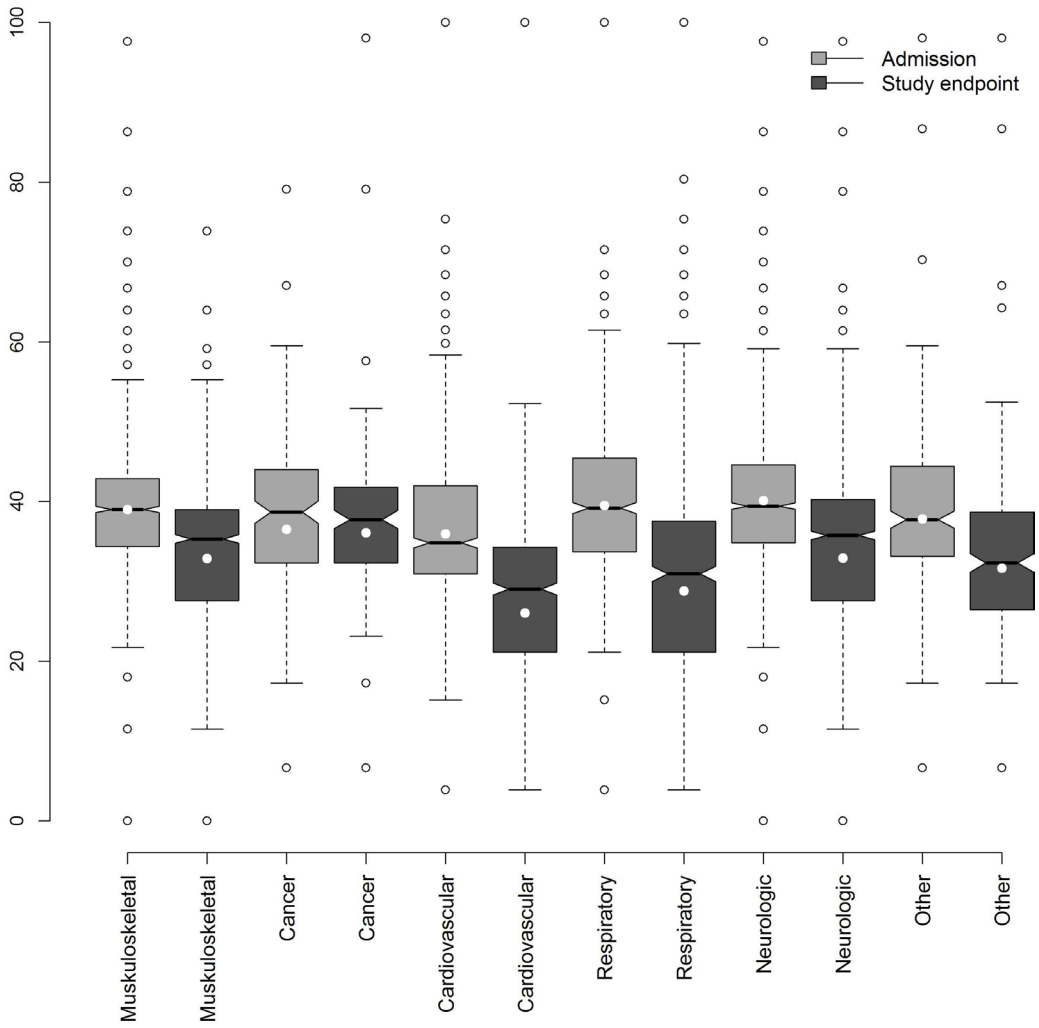


**Figure 2** Distribution of interval-scale ICF Generic 6 Set scores (0–100 scale) at admission and study endpoint for each health condition group. The central rectangle spans the first quartile to the third quartile, with the central segment showing the median and the whiskers above and below the box extend until each reaches no more than 1.5 times the height of the box (third quartile–first quartile). The points above and below the end of whiskers are the outliers. The white points indicate the mean.

pain is a function or a symptom.[40] Furthermore, there is a difference between the actual sensation of pain and cases where this sensation is impaired, that is, patients are not able to feel pain in certain body parts in spite of tissue damage. Moreover, there are the issues of neuropathic pain and phantom pain. Future research is needed on how this category is actually understood by the raters as well as patients in different situations.

*Remunerative employment (d850)* was not included in the Rasch model since the category was difficult to assess in the clinical setting. However, since previous research claimed that this ICF category is relevant to community follow-up, *Remunerative employment (d850)* should be assessed and reported alongside the interval ICF Generic 6 Set score.[41]

As in our pilot study, the interval ICF Generic 6 Set score was sensitive to improvement of patients' functioning during inpatient treatment. With regard to the new COSMIN guidelines for testing sensitivity to change (responsiveness) of a measure, we could not assess proportions of correlation between the change in the interval ICF Generic 6 Set score and change in another functioning measure.[42] In line with the results of the pilot study, we would expect moderate to large treatment effects. In contrast to other health conditions groups, the standardised effect size was, however, small for cancer and other unclassified diagnoses. This may be owed to progression of disease counteracting treatment effects. In addition, the heterogeneity of this group may have posed an issue. Furthermore, beyond standardised effect sizes, minimal clinically important differences in scores need to be determined in future research. In contrast to untransformed raw scores of the ICF Generic 6 Set score, the Rasch-transformed interval score could largely reduce or cancel ceiling or floor effects in identifying change for patients with very low or very high baseline scores. This finding is indeed promising in that it shows a wide applicability of the interval-scale ICF Generic 6 Set across the patient spectrum.

We note several limitations to our study. First, most of the patients had neurological and musculoskeletal conditions, which may limit the generalisability to other

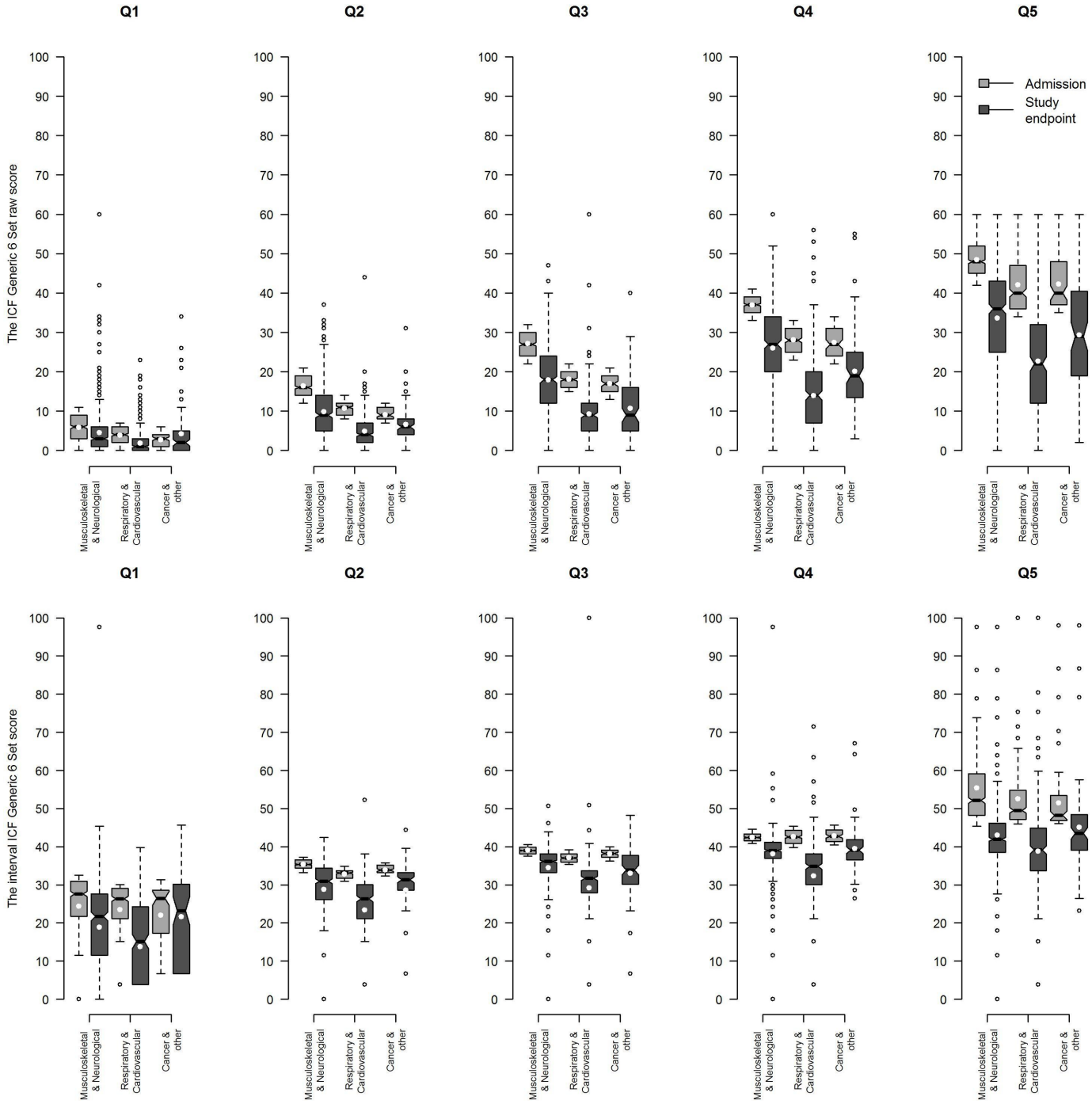


**Figure 3** Distribution of ICF Generic 6 Set raw scores (0–60 scale; upper part of the figure) and interval-scale ICF Generic 6 Set scores (0–100 scale; lower part of the figure) stratified at baseline quintiles (Q1=the best initial functioning, Q5=the worst initial functioning) for each health condition group. The central rectangle spans the first quartile to the third quartile, with the central segment showing the median and the whiskers above and below the box extend until each reaches no more than 1.5 times the height of the box (third quartile–first quartile). The points above and below the end of whiskers are the outliers. The white points indicate the mean.

diagnostic groups such as cancer. We, however, accounted for this in determining samples for the Rasch analysis by equal representation of all health conditions, genders and age groups. Second, patients were from grade III and II hospitals that were recruited at a nursing conference or through authors' personal networks. We thus cannot exclude selection bias further limiting generalisability of our study. It should, however, be noted that there are usually no more than 2–3 grade III hospitals per province so that our sample should be at least fairly representative for the 20 included Provinces. Third, the assessment of functioning at admission and discharge by the same nurse might affect applicability of our results to clinical practice. We, however, collected additional data on 703

patients with functioning rated by two independent nurses at each time point for investigating interrater reliability. A follow-up study will report the results of the respective analysis. Fourth, although most floor effects in the detection of change were no longer present after transformation of the ordinal raw score to an interval score based on Rasch-abilities, reduced floor effects remained for the cancer and other health conditions groups. Clinical studies using the ICF Generic 6 Set as an outcome measure could deal with this problem by employing Tobit models.[43] Fifth, although it is possible to assess how a person manages daily routine along a wide range of activities of daily living (ADL) in hospital environments, these environments differ from those which patients face when discharged to their homes and communities. Performance in managing daily routine (d230), but also walking (d450) or moving around (d455) once discharged to their homes and communities can thus only be inferred from what patients are able to do in the hospital. How good this inference is must be evaluated in future studies examining the use of the ICF Generic 6 in community follow-up.

This study marks the first attempt to apply the ICF nationwide and generate a reliable, interval score of functioning. Further research studying the interrater reliability, convergent validity, known group validity and predictive validity of the ICF Generic 6 Set is underway. In line with the actual efforts to validate the ICF Core Sets across the six WHO regions, similar attempts are needed and have in fact been initiated in other countries to apply the ICF Generic Set. For instance, under the leadership of the European Union of Medical Specialists, Board of Physical and Rehabilitation Medicine an initiative was launched towards developing ICF-based clinical data collection tools following the approach described in this study.[44 45] Collaborations across countries will allow developing a universal scoring algorithm of functioning which will ultimately allow comparison of functioning outcomes across health conditions and clinics as shown in this study and across countries.

## CONCLUSION

In conclusion, the ICF Generic 6 Set in combination with an 11-point numeric rating scale can be used for creating an interval score of functioning that is sensitive to change in clinical practice and across a wide range of health conditions. We recommend the use of the ICF Generic 6 Set on a 11-point numeric rating scale in clinical practice and research within Mainland China. However, the reliability of the ICF Generic 6 Set in terms of PSI was only moderate. Our finding also revealed that some items, for example, *Sensation of pain (b280)*, require specific attention. Based on the evidence gained in this study, future studies are needed to test the ICF Generic Set as a standard for the reporting of functioning information in different healthcare systems and countries.

**Author affiliations**
[1]Swiss Paraplegic Research, Nottwil, Switzerland
[2]Department of Health Sciences and Health Policy, University of Lucerne, Luzern, Switzerland
[3]ICF Research Branch, a cooperation partner within the WHO Collaborating Centre for the Family of International Classifications in Germany (at DIMDI), Nottwil, Switzerland
[4]Shenzhen Hospital of the Southern Medical University, Guangzhou, China
[5]Department of Rehabilitation Medicine, Third Affiliated Hospital of Peking University, Beijing, China
[6]Department of Rehabilitation Medicine, First Affiliated Hospital of Nanjing Medical University, Nanjing, China
[7]Department of Health Sciences, Institute for Disaster Management and Reconstruction, Sichuan University and Hong Kong Polytechnic University, Chengdu, China

**Acknowledgements** The authors of this study would like to acknowledge the support of the National Health and Family Planning Commission of the People's Republic of China, the Chinese Association of Rehabilitation Medicine and the Rehabilitation Nursing Committee of the Guangdong Nursing Association. We also express our deep gratitude to all of participating hospitals and clinical raters. The participating hospitals were: Affiliated Hospital of Guangdong Medical College, Guangzhou Panyu Central Hospital, Zhujiang Hospital, First Affiliated Hospital of Guangzhou Traditional Chinese Medical University, Guangzhou Huiai Hospital (Guangzhou Brain Hospital), People's Hospital of Baoan Shenzhen, People's Hospital of New District Longhua Shenzhen, Shenzhen Baoan Hospital of Traditional Chinese Medicine, Shantou Central Hospital, Haojiang Hospital (First Affiliated Hospital of Shantou Medical College), Beijiao Hospital of Southern Medical University, People's Hospital of Nanhai District of Foshan, Guangdong Tongjiang Hospital, Third People's Hospital of Huizhou, Zhaoqing Gaoyao People's Hospital, Dongguan Kanghua Hospital, Fuzhou General Hospital of Nanjing Military, West China Hospital of Sichuan University, Rehabilitation Hospital of Sichuan Province, Nanchong Central Hospital, Sichuan Provincial People's Hospital, First Affiliated Hospital of Shanxi Medical University, First Affiliated Hospital of the Fourth Military Medical University, Jiangsu Provincial Hospital, Jiangsu Provincial Hospital of Traditional Chinese Medicine, Qingdao Municipal Hospital, Second Affiliated Hospital of Shandong University of Traditional Chinese Medicine, The Affiliated Hospital of Qingdao University, Hai Nan General Hospital, Affiliated Hospital of Hainan Medical University, People's Hospital of Fuyang, First people's Hospital of Hefei City (Third Affiliated Hospital of Anhui Medical University), Second Affiliated Hospital of Wenzhou Medical College, General Hospital of Ningxia Medical University, Cardiovascular Disease Hospital of Ningxia Medical University, Jilin University Sino-Japanese Friendship Hospital, People's Hospital of Xinjiang Uygur Autonomous Region, First Affiliated Hospital of Chongqing Medical University, Second Affiliated Hospital of Chongqing Medical University, Guizhou Provincial People's Hospital, Affiliated Hospital of Guiyang Medical College, People's Hospital of Xinyu, Rehabilitation Hospital of Heilongjiang Provincial Seafarers General Hospital, First Affiliated Hospital of Jiamusi University, Third Affiliated Hospital of Jiamusi University, Second Affiliated Hospital of Kunming Medical University, General Hospital of the Yangtze River Shipping (Wuhan Brain Hospital), Affiliated Hospital of Qinghai University, Elderly Hospital of Shanghai Jingan, Shanghai Ledu Hospital.

**Contributors** CE contributed to the conception and design of the study, performed and interpreted all the analyses of the data and drafted the article. JDR, GS and BP contributed to conception of the design of the study and interpretation of the data and provided supervision. WC, XZ, SL, SL and JL contributed to the acquisition of the data, conception and the design of the study. All authors contributed to critically revising the paper for important intellectual content and all authors read and approved the final article.

**Funding** The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

**Competing interests** None declared.

**Patient consent** Obtained.

**Ethics approval** This study has been approved by the Chinese Association of Rehabilitation Medicine and exempt from hospital ethics approval since it involved non-invasive clinician based assessment of patients based on routinely collected clinical data.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data sharing statement** The materials and datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.



## REFERENCES


1. WHO (ed). *International Classification of Functioning, Disability, and Health*. Geneva: World Health Organization, 2001.
2. Kostanjsek N, Rubinelli S, Escorpizo R, *et al*. Assessing the impact of health conditions using the ICF. *Disabil Rehabil* 2011;33(15-16):1475–82.
3. Rabinowitz J, Modai I, Inbar-Saban N. Understanding who improves after psychiatric hospitalization. *Acta Psychiatr Scand* 1994;89:152–8.
4. McCrone P, Phelan M. Diagnosis and length of psychiatric in-patient stay. *Psychol Med* 1994;24:1025–30.
5. Ormel J, Oldehinkel T, Brilman E, *et al*. Outcome of depression and anxiety in primary care. A three-wave 3 1/2-year study of psychopathology and disability. *Arch Gen Psychiatry* 1993;50:759–66.
6. Narain P, Rubenstein LZ, Wieland GD, *et al*. Predictors of immediate and 6-month outcomes in hospitalized elderly patients.The importance of functional status. *J Affect Disord Am Geriatr Soc* 1988;36:775–83.
7. Rock BD, Goldstein M, Harris M, *et al*. Research changes a health care delivery system: a biopsychosocial approach to predicting resource utilization in hospital care of the frail elderly. *Soc Work Health Care* 1996;22:21–37.
8. Baseman S, Fisher K, Ward L, *et al*. The relationship of physical function to social integration after stroke. *J Neurosci Nurs* 2010;42:237–44.
9. Hopfe M, Stucki G, Marshall R, *et al*. Capturing patients' needs in casemix: a systematic literature review on the value of adding functioning information in reimbursement systems. *BMC Health Serv Res* 2016;16:40.
10. Doganay Erdogan B, Leung YY, Pohl C, *et al*. Minimal clinically important difference as applied in rheumatology: An omeract rasch working group systematic review and critique. *J Rheumatol* 2016;43:194–202.
11. Selb M, Escorpizo R, Kostanjsek N, *et al*. A guide on how to develop an international classification of functioning, disability and health core set. *Eur J Phys Rehabil Med* 2015;51:105–17.
12. Cieza A, Oberhauser C, Bickenbach J, *et al*. Towards a minimal generic set of domains of functioning and health. *BMC Public Health* 2014;14:218.
13. Cieza A, Fayed N, Bickenbach J, *et al*. Refinements of the ICF Linking Rules to strengthen their potential for establishing comparability of health information. *Disabil Rehabil* 2016:1–10.
14. Prodinger B, Reinhardt JD, Selb M, *et al*. Towards system-wide implementation of the international classification of functioning, disability and health (ICF) in routine practice: Developing simple, intuitive descriptions of ICF categories in the ICF Generic and Rehabilitation Set. *J Rehabil Med* 2016;48:508–14.
15. Stucki G, Prodinger B, Bickenbach J. Four steps to follow when documenting functioning with the international classification of functioning, disability and health. *Eur J Phys Rehabil Med* 2017;53:144–9.
16. Oberhauser C, Chatterji S, Sabariego C, *et al*. Development of a metric for tracking and comparing population health based on the minimal generic set of domains of functioning and health. *Popul Health Metr* 2016;14:19.
17. Reinhardt JD, Zhang X, Prodinger B, *et al*. 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. *J Rehabil Med* 2016;48:515–21.
18. Fischer GH, Molenaar LW, eds. *Rasch models, Fundations, recent development, and applications*. New York, 1995.
19. Christensen KB, Kreiner S, Mesbah M, eds. *Rasch Models in Health*, 2013.
20. Zhou Y. *Comparing parametric item response theory and nonparametric item response theory: Application in psychological research using polytomous items: ETD Collection for Fordham University*, 2011.
21. Andrich D, Sheridan B, Luo G, eds. *Rasch models for measurement: RUMM2030*. Perth, Western Australia, 2010.
22. Smith EV. Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *J Appl Meas* 2002;3:205–31.
23. Wainer H, Thissen D. Estimating ability with the wrong model. *Journal of Educational Statistics* 1987;12:339–68.
24. Yen WM. Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Appl Psychol Meas* 1984;8:125–45.
25. Christensen KB, Makransky G, Horton M. Critical Values for Yen's $Q_3$: Identification of Local Dependence in the Rasch Model Using Residual Correlations. *Appl Psychol Meas* 2017;41:178–94.
26. Wainer H, Kiely GL. Item clusters and computerized adaptive testing: A case for testlets. *J Educ Meas* 1987;24:185–201.
27. Andrich D. Item discrimination and Rasch-Andrich thresholds revisited. *Rasch Meas Transact* 2006;20:1055–7.
28. Hambleton RK. Good practices for identifying differential item functioning. *Med Care* 2006;44(11 Suppl 3):S182–S188.
29. Masters GN. A rasch model for partial credit scoring. *Psychometrika* 1982;47:149–74.
30. Wright BD, Masters GN, eds. *Rating scale analysis*. Chicago: MESA Press, 1982.
31. Andrich D. *Rasch models for measurement. Sage University Paper Series on Quantitative Applications in the Social Sciences series no 07-068 Newbury Park*. California: Sage Publications, 1988.
32. Bond T, Fox C, eds. *Applying the Rasch model: Fundamental measurement in the human sciences*. 2nd ed. NJ: LEA, 2007.
33. Andrich D. An index of person separation in latent trait theory, the traditional KR. 20 index, and the Guttman scale response pattern. *Educational Research and Perspectives* 1982;9:95–104.
34. Cohen JE, ed. *Statistical Power Analysis for the Behavioral Sciences Hillsdale*. NJ: Lawrence Erlbaum Associates, Inc, 1988.
35. Stucki G, Michael BA. How to measure improvement: rules and fallacies. *Rheumatology in Europe* 1995;24:107–11.
36. Tukey JW. Comparing individual means in the analysis of variance. *Biometrics* 1949;5:99–114.
37. Tabachnick BG, Fidell LS, eds. *Using Multivariate Statistics*. 5th ed. Upper Saddle River,NJ: Pearson Allyn & Bacon, 2001.
38. Yan J, ed. *Examining local item dependence effects in a large-scale science assessment by a Rasch partial credit model*, 1997.
39. Bao H, Dayton CM, Hendrickson AB. Differential item functioning amplification and cancellation in a reading test. *Practical Assessment, Research & Evaluation* 2009;14.
40. WHO. Eleventh revision of the international classification of diseases (icd-11) for mortality and morbidity statistics. https://icd.who.int/browse11/l-m/en
41. Li J, Prodinger B, Reinhardt JD, *et al*. TOwards the system-wide implementation of the international classification of functioning, disability and health in routine practice: Lessons from a pilot study in China. *J Rehabil Med* 2016;48:502–7.
42. COSMIN-manual. http://www.cosmin.nl/images/upload/File/COSMIN%20checklist%20manual%20v6.pdf
43. Twisk J, Rijmen F. Longitudinal tobit regression: a new approach to analyze outcome variables with floor or ceiling effects. *J Clin Epidemiol* 2009;62:953–8.
44. Prodinger B, Scheel-Sailer A, Escorpizo R, *et al*. European initiative for the application of the international classification of functioning, disability and health: Development of clinical assessment schedules for specified rehabilitation services. *Eur J Phys Rehabil Med* 2017;53.
45. Selb M, Gimigliano F, Prodinger B, *et al*. Toward an international classification of functioning, disability and health clinical data collection tool: the italian experience of developing simple, intuitive descriptions of the rehabilitation set categories. *Eur J Phys Rehabil Med* 2017;53.