OXFORD

Research Article

The influence and added value of a Standardized Assessment and Reporting System for functioning outcomes upon national rehabilitation quality reports

ROXANNE MARITZ[1,2], CRISTINA EHRMANN[2], BIRGIT PRODINGER[1,2,3], ALAN TENNANT[1], and GEROLD STUCKI[1,2]

[1]Department of Health Sciences and Medicine, University of Lucerne, 6002 Lucerne, Switzerland, [2]Swiss Paraplegic Research, 6207 Nottwil, Switzerland, and [3]Faculty of Applied Health and Social Sciences, Technical University of Applied Sciences Rosenheim, 83024 Rosenheim, Germany

Address reprint requests to: Roxanne Maritz, Swiss Paraplegic Research, Guido A. Zächstrasse 4, 6207 Nottwil, Switzerland Fax: +41 41 939 66 40; E-mail: roxanne.maritz@paraplegie.ch

Abstract

Objective: To demonstrate the influence and added value of a Standardized Assessment and Reporting System (StARS) upon the reporting of functioning outcomes for national rehabilitation quality reports. A StARS builds upon an ICF-based (International Classification of Functioning, Disability and Health) and interval-scaled common metric.

Design: Comparison of current ordinal-scaled Swiss national rehabilitation outcome reports including an expert-consensus-based transformation scale with StARS-based reports through descriptive statistical methods and content exploration of further development areas of the reports with relevant ICF Core Sets.

Setting: Swiss national public rehabilitation outcome quality reports on the clinic level.

Participants: A total of 29 Swiss rehabilitation clinics provided their quality report datasets including 18 047 patients.

Interventions: Neurological or musculoskeletal rehabilitation.

Main outcome measures: Functional Independence Measure™ or Extended Barthel Index.

Results: Outcomes reported with a StARS tended to be smaller but more precise than in the current ordinal-scaled reports, indicating an overestimation of achieved outcomes in the latter. The comparison of the common metric's content with ICF Core Sets suggests to include 'energy and drive functions' or 'maintaining a basic body position' to enhance the content of functioning as an indicator.

Conclusions: A StARS supports the comparison of outcomes assessed with different measures on the same interval-scaled ICF-based common metric. Careful consideration is needed whether an ordinal-scaled or interval-scaled reporting system is applied as the magnitude and precision of reported outcomes is influenced. The StARS' ICF basis brings an added value by informing further development of functioning as a relevant indicator for national outcome quality reports in rehabilitation.

Introduction

The measurement and monitoring of clinical performance are central for hospital quality improvement [1]. For the monitoring of institutional outcomes in national quality reports, the main health indicators of a health system need to be addressed, and for rehabilitation this indicator is functioning [2, 3]. Functioning is classified by the World Health Organization's International Classification of Functioning, Disability and Health (ICF), incorporating both biological health—the intrinsic health capacity described as body functions and structures, as well as lived health—the actual engagement of a person in activities and life situations in interaction with the environment [4, 5].

Currently, in rehabilitation, functioning outcomes are assessed with a variety of ordinal-scaled assessment tools, which makes it difficult to compare, aggregate [6, 7] and eventually learn from the related information for improvement processes. Therefore, standardization is essential for measurement of achieved outcomes within clinics, and critical for comparisons between clinics [1]. A concrete example comes from the Swiss public national rehabilitation outcome quality reports, in which musculoskeletal and neurological rehabilitation clinics can choose between two ordinal-scaled assessment tools assessing functioning outcomes in the domain of activities of daily living (ADL)—the Functional Independence Measure (FIM^{TM}) or the Extended Barthel Index (EBI) [8].

To standardize outcomes, two approaches exist: 1) define specific assessment tools, which have to be used by all stakeholders or 2) enable standardized reporting and thus comparability of routinely used assessment tools [6]. For the latter, different approaches, such as expert-consensus-based transformations [9] or Standardized Assessment and Reporting Systems (StARS) for functioning outcomes, based on a statistical, i.e. Rasch-based scale transformation approach [10] can be applied. Expert-consensus-based transformations have the advantage that experienced clinicians are involved in the process but the disadvantage that ordinal-scale properties of outcomes remain, thus restricting valid calculations of means or change scores [7, 11]. In contrast, a StARS for functioning outcomes includes a common metric as a core element, which has two main features: first, it is conceptually based on the ICF as the international standard for reporting functioning information and second, it is interval-scaled, thus, allowing for any parametric analyses in reporting and monitoring outcomes [6, 12].

National quality reports provide an excellent opportunity to examine the influence of different approaches toward standardized reporting of clinic outcomes. Therefore, the objective of the current study was to demonstrate the influence and added value of a StARS, with its interval-scaled ICF-based common metric, upon the reporting of functioning outcomes in national rehabilitation quality reports. Specific aims were related to the common metric's two main features:

1) To examine the influence of the common metric's interval-scaling feature in comparison to (i) functioning outcomes reported with ordinal-scaled assessment total scores and (ii) an ordinal-scaled expert-consensus-based transformation of these scores.

2) To outline the added value of the common metric's ICF basis for the identification of potential further functioning outcome indicators relevant for rehabilitation.

Switzerland was used as a case in point for this study, as both a currently applied ordinal-scaled expert-consensus-based system and a newly developed interval-scaled and ICF-based StARS exist for musculoskeletal and neurological rehabilitation national quality outcome reports.

Methods

Setting, participants and interventions

Secondary analysis of Swiss outcome quality reports in musculoskeletal and neurological rehabilitation, which are coordinated and published by the National Association for Quality Development in Hospitals and Clinics (ANQ) [8], was conducted. From 64 Swiss rehabilitation clinics providing musculoskeletal or neurological rehabilitation in 2016, 29 clinics agreed to provide their ANQ datasets for our study. Ethical approval was received from the Swiss Ethics Committees.

Outcome measures

Sociodemographic, treatment, health status and functioning-related data are routinely assessed for the ANQ quality reports. To assess functioning outcomes in musculoskeletal and neurological rehabilitation, clinics can choose between—FIM^{TM} and EBI [13, 14].

The FIM^{TM} includes 18 items: 13 items related to motor and five to cognitive skills. All items are scored from 1–7 resulting in an ordinal-scaled total score between 18 (total dependence) and 126 (complete independence) [13]. The EBI includes 16 items: 10 motor items based on the Barthel Index [15] and six cognitive items, of which five are derived from the cognitive FIM^{TM} items [14]. All items are scored 0–4, resulting in an ordinal-scaled total score between 0 (total dependence) and 64 (complete independence). While both EBI and FIM^{TM} are administered by health professionals, related training is only mandatory for FIM^{TM}. Recent studies in the context of quality reports showed that both tools can be reported as an interval-scaled total score when Rasch-based transformation is applied [16, 17].

Expert-consensus-based ANQ-ADL score and ICF-based interval-scaled common metric

To enable comparison of all rehabilitation clinics in national reports, irrespective of whether FIM^{TM} or EBI was assessed, two options exist:

(A) The ANQ-ADL score currently used in the ANQ reports, consists of an ordinal-scaled expert-consensus-based transformation algorithm between FIM^{TM} and EBI on item basis [9]. It ranges from 0 (complete dependence) to 60 (complete independence). It allows the comparison at item level but has its limitations: (1) its

Table 1 Main features of the ANQ-ADL score and the ICF-based interval-scaled common metric

	ANQ-ADL score	Common metric
Scale level	Ordinal-scaled	Interval-scaled
Included assessment tools	FIM^{TM} and EBI	FIM^{TM} and EBI
Scale range	0–60	0–100 (adaptable)
Development	Expert-consensus process, validation with a validation sample	Content equivalence assessed with ICF Linking Rules, score equivalence assessed with Rasch-based scale equating approach based on the same validation sample as the ANQ-ADL score
Strengths	– Involvement of experienced clinicians – Item-based approach	– Based on the international standard for reporting functioning outcomes (ICF) – Interval-scale allows for calculations such as means and change scores – Considers the operational range of the integrated assessment tools – Includes all items of both tools
Weaknesses	– Ordinal-scale does not allow for calculations – Does not consider the operational range of the included assessment tools – Does not include EBI Item 16 'Vision/Neglect', as there is no corresponding item in FIM^{TM}	– Total score-based approach – More specialized statistical resources (Rasch analysis) required for development

ANQ-ADL score = Swiss National Association for Quality development in hospitals and clinics Activities of Daily Living Score, ICF = International Classification of Functioning, Disability and Health, FIM^{TM} = Functional independence measure, EBI = Extended Barthel Index.

ordinal-scaling, (2) exclusion of EBI Item 16 'vision/neglect' and (3) automatic match of minimum and maximum scores of the two scales not considering their different operational ranges. The ANQ-ADL score was validated using a representative sample of 265 neurorehabilitation patients all being assessed with both FIM^{TM} and EBI [9].

(B) The newly developed ICF-based interval-scaled common metric [12]. It includes FIM^{TM} and EBI on total score level and was developed on the basis of the ANQ-ADL score validation sample, applying ICF Linking Rules [18] and Rasch methods for scale equating [10]. Its advantages include (1) its interval-scale, needed for calculations currently conducted in ANQ reports, (2) its consideration of the operational range of included tools [12] and (3) its ICF basis (see Appendix A1) allowing to compare the metric's content, e.g. with other tools. The common metric was designed to range from 0 to 100, which can be adjusted, as it is based on logit Rasch values (see Appendix A2).

The main features of these two options are summarized in Table 1.

Data analysis

The examination of the influence of the common metric's interval-scaling feature (specific aim 1) included three steps: (1) examination of the difference between reporting of functioning outcomes with ordinal-scaled FIM^{TM}, EBI and respectively ii) ANQ-ADL scores and the interval-scaled common metric; (2) examination of the difference between risk-adjusted funnel-plots of clinic performance based on the ordinal-scaled ANQ-ADL score and the interval-scaled common metric and (3) examination of floor and ceiling effects of ordinal-scaled FIM^{TM}, EBI and ANQ-ADL score as well as the interval-scaled common metric. Only those cases that could be clearly assigned to neurological (NEUR) or (MSK) rehabilitation and had complete data for admission and discharge of FIM^{TM} or EBI, as well as the risk-adjustment variables, were included.

In the fourth step (4), we compared the ICF categories covered by the common metric with relevant ICF Core Sets to outline the added value of the common metric's ICF basis (specific aim 2).

The analyses were conducted with RStudio (steps 1–3) and Microsoft Excel (step 4).

1) *Difference between reporting of ordinal-scaled assessment tools respectively ANQ-ADL scores and the interval-scaled common metric*

We created a descriptive table for the comparison of admission, discharge and change scores, i.e. discharge score minus admission score, separately for MSK and NEUR rehabilitation on the clinic level, including respective standard deviations. In order to compare the respective values of ordinal-scaled FIM^{TM}, EBI and ANQ-ADL score to the interval-scaled common metric, we adapted the range of the common metric according to the scale it was compared to, i.e. 18–126 for FIM^{TM}, 0–64 for EBI and 0–60 for the ANQ-ADL score, on the basis of its Rasch logits.

2) *Difference of risk-adjusted funnel-plots for clinic performance between the ordinal-scaled ANQ-ADL score and the interval-scaled common metric*

We reproduced the funnel-plots of clinic performance from the ANQ reports, once based on the ANQ-ADL scores and once on the common metric for both rehabilitation groups. We used the same risk-adjustment method as ANQ in 2016, i.e. simple linear regression including the discharge ANQ-ADL respectively common metric scores as dependent variable and the following independent variables: gender, age, nationality, residence before admission, residence after discharge, health insurance status and type, diagnosis group, Modified Cumulative Illness Rating Scale (CIRS), duration of rehabilitation and admission ANQ-ADL respectively common metric scores [19]. We then compared the two funnel plots within one rehabilitation group and analyzed which clinics changed in regard

to the three funnel-plot categories (significant upward deviation, no significant deviation and significant downward deviation from regression estimate).

3) *Floor and ceiling effects of ordinal-scaled FIM^{TM}, EBI and ANQ-ADL score and the interval-scaled common metric*

As floor and ceiling effects are important quality criteria of outcome measures in health [20], we assessed the percentage of people from each rehabilitation group attaining minimum and maximum scores in FIM^{TM}, EBI, ANQ-ADL score and common metric separately for admission and discharge. We defined an indication for floor respectively ceiling effect if > 5% and a clear floor respectively ceiling effect if > 15% reached minimum respectively maximum scores [20].

4) *Added value of the common metric's ICF basis*

The original linking of the items contained in the common metric to the ICF using ICF Linking Rules [12, 18] resulted in 26 covered ICF categories (see Appendix A1). These categories were contrasted to categories of relevant ICF Core Sets in order to define gaps and further development opportunities for the StARS common metric for the ANQ outcome quality reports. ICF Core Sets are purpose-tailored shortlists of ICF categories developed in a standardized multimethod scientific process [21]. There exist two generic ICF Sets, and diagnosis and rehabilitation group-specific sets [22], each with brief and comprehensive versions. We contrasted the common metric's ICF categories with the two generic ICF Sets (Generic-7, Generic-30) [23, 24] and the eight rehabilitation group-specific ICF Core Sets for MSK and NEUR, each in its acute and postacute respectively brief and comprehensive version [25].

Results

Sample characteristics

The overall sample included 18047 complete cases in musculoskeletal (MSK, *n* = 12160) and neurological (NEUR, *n* = 5887) rehabilitation form 26 clinics, of which 18 were located in the German-speaking, five in the French-speaking and three in the Italian-speaking part of Switzerland. Twelve clinics provided both MSK and NEUR rehabilitation, 11 provided only MSK and three only NEUR rehabilitation. Nineteen clinics were assessing FIM^{TM} (*n* = 11 636) and seven were assessing EBI (*n* = 6411). The gender distribution for MSK was 36.7% male (*n* = 4461) and 63.3% female (*n* = 7699), and for NEUR rehabilitation 52.5% male (*n* = 3091) and 47.5% female (*n* = 2796). The mean age of the MSK sample was 69.8 years ranging from 18 to 102. The mean age of the NEUR sample was 64.9 years ranging from 18 to 99. Average rehabilitation duration of MSK patients was 21 days (ranging from 7–182 days) and 37 days for NEUR patients (ranging from 7 to 351 days).

Difference between reporting of functioning outcomes with ordinal-scaled scores and the interval-scaled common metric

Table 2 shows the admission, discharge and change scores on clinic level, separately for MSK and NEUR rehabilitation. In 20 of the 23 MSK rehabilitation clinics, the change scores are higher when the ordinal scales of FIM^{TM}, EBI and ANQ-ADL score are used in comparison to the interval-scaled common metric. This was also the case for 14 of the 15 NEUR clinics, indicating a tendency to overestimate outcomes when reported with ordinal-scaled scores. For both rehabilitation groups, the total standard deviation of the

different values is smaller for the interval-scaled metric, indicating a greater degree of precision when the common metric is used.

Difference between the risk-adjusted funnel-plots of clinic performance

Figure 1 shows the four funnel-plots comparing risk-adjusted clinic performances when using the ANQ-ADL score and the ICF-based interval-scaled common metric for the two rehabilitation groups. In MSK rehabilitation, five clinics (22%) changed the funnel-plot categories. Two clinics changed from 'no significant deviation' to 'significant upward deviation'. The deviation refers to the regression estimate, which is based on the case-mix related risk-adjustment. So, an upward deviation indicates that clinics performed better than their case-mix-related mean estimation of their performance. One clinic changed from 'significant upward deviation' to 'no significant deviation' and two clinics changed from 'no significant deviation' to 'significant downward deviation'. In NEUR rehabilitation, one clinic (7%) changed from 'significant upward deviation' to 'no significant deviation'.

Floor and ceiling effects

Table 3 shows the results of the analysis for floor and ceiling effects. There was no indication of floor effects for all four scales and also for FIM^{TM} and the common metric no indication for ceiling effects. For EBI, there was a clear ceiling effect for MSK rehabilitation at admission (17.7%) and discharge (21.8%) and an indication of ceiling effect for NEUR rehabilitation at admission (5.1%) and discharge (12.2%). For the ANQ-ADL score, there was an indication for ceiling effect for MSK at admission (7.5%) and discharge for both MSK (12.8%) and NEUR rehabilitation (8.6%). This supports the results from the development of the common metric, which showed that FIM^{TM} has a larger operational range for patients in comparison to EBI (see Appendix A4).

Added value of the common metric's ICF basis

Table 4 shows the overview of the comparison of the common metric's ICF categories with relevant ICF Core Sets. The extensive comparison table on the level of the ICF categories can be found in Appendix A3. The ICF Core Sets are only covered by the common metric with a maximum of 40.0% (Generic-30 Set) and a minimum of 17.2% (ICF NEUR postacute Core Set comprehensive version). The most relevant ICF categories covered by 8 of the 10 analyzed ICF Sets, not present in the common metric were b130 'Energy and drive functions' and d415 'Maintaining basic body position'. The following relevant categories covered by eight or more of the analyzed Core Sets, which were already represented in the common metric, were b620 'Urination functions', d410 'Changing basic body position', d420 'Transferring oneself', d450 'Walking', d510 'Washing oneself', d520 'Caring for body parts', d530 'Toileting' and d550 'Eating', stressing the importance of these aspects.

Discussion

This study demonstrates the influence and added value of an ICF-based interval-scaled StARS for national quality reports, on two levels: (1) the statistical level contrasting the influence of the common metric's interval-scale in comparison to the ordinal-scaled instrument's raw score and an ordinal-scaled expert-consensus-based transformation and (2) the added value on the content level contrasting the

Table 2 Difference between reporting of functioning outcomes ordinal scores and the interval-scaled common metric: musculoskeletal rehabilitation (M) and neurological rehabilitation (N)

Clinic Nr.	Outcome measure (score range)	Admission score ordinal (SD)	Discharge score ordinal (SD)	Change score ordinal (SD)	Admission score metric (SD)	Discharge score metric (SD)	Change score metric (SD)
1 M	ADL Score (0–60)	43.9 (8.6)	50.7 (7.5)	6.8 (5.1)	35.5 (4.8)	39.2 (4.3)	3.7 (2.4)
2M[a]	ADL Score (0–60)	50.4 (7.2)	55.3 (6.2)	4.9 (4.6)	43.1 (7.4)	48.1 (7.1)	5 (4.6)
3 M	ADL Score (0–60)	43.9 (8.4)	53 (6.9)	9 (5.8)	35.9 (5)	42.3 (5.1)	6.5 (3.6)
4 M	ADL Score (0–60)	46.8 (10.7)	50.6 (9.3)	3.8 (4.7)	38.6 (7)	41.8 (7)	3.3 (3.4)
5 M	ADL Score (0–60)	41.3 (10.9)	49.9 (10.8)	8.6 (5.9)	34.7 (6.1)	40.5 (7)	5.8 (4)
6 M	ADL Score (0–60)	55.3 (8.9)	58.5 (4.3)	3.2 (6.7)	45.9 (6.4)	47.6 (3.7)	1.7 (3.5)
7 M	ADL Score (0–60)	45.2 (10.3)	51 (9.8)	5.8 (5.8)	37.5 (6.9)	41.8 (7.2)	4.3 (3.9)
8 M	ADL Score (0–60)	41.8 (11.6)	51.3 (10.1)	9.5 (7.3)	35.5 (7.4)	41.8 (7.5)	6.3 (4.2)
9 M	ADL Score (0–60)	35.2 (10.9)	47.3 (10)	12.1 (8.7)	31.1 (5.1)	36.8 (4.7)	5.7 (3.9)
10M[a]	ADL Score (0–60)	51.8 (8.2)	54.9 (7.9)	3.1 (7)	45.1 (9.1)	48.9 (9.1)	3.8 (6.5)
11 M	ADL Score (0–60)	50.4 (8.2)	55.2 (5)	4.8 (5.7)	40.8 (54)	44.3 (3.8)	3.5 (3.5)
12 M	ADL Score (0–60)	44.7 (9.6)	53.1 (7.7)	8.4 (6.9)	37.6 (5.7)	43 (5)	5.4 (4.1)
13 M	ADL Score (0–60)	37.3 (13.2)	51 (9.2)	13.6 (9.6)	32.3 (6.2)	39.9 (6.1)	7.6 (4.4)
14 M	ADL Score (0–60)	51.2 (5.5)	57 (3.5)	5.8 (4.3)	40.9 (3.9)	45.5 (3.6)	4.6 (3)
15 M	ADL Score (0–60)	41.6 (9.8)	49.9 (7.8)	8.3 (7.5)	34.4 (5.4)	39.2 (5.3)	4.9 (4.6)
16 M[a]	ADL Score (0–60)	49.9 (4.1)	54.9 (3.1)	5.1 (2.7)	40.3 (3.9)	47.1 (5.4)	6.8 (4.3)
17 M	ADL Score (0–60)	50.7 (9.2)	54.6 (6.8)	4 (6.2)	41.2 (6)	43.9 (4.9)	2.6 (3.7)
18 M	ADL Score (0–60)	45.2 (10.5)	49.2 (9.3)	4 (5.1)	37.6 (6.3)	40.1 (5.8)	2.5 (3.1)
19 M	ADL Score (0–60)	41.9 (10)	52.4 (8.3)	10.5 (7.8)	34.7 (4.8)	40.6 (5.5)	5.9 (4.6)
20 M	ADL Score (0–60)	50.4 (7.9)	54.3 (5.8)	3.9 (4.7)	41 (5)	43.5 (4.1)	2.5 (2.7)
21 M	ADL Score (0–60)	45.1 (10.6)	53 (8.5)	8 (7)	36.7 (6.1)	42.4 (6.1)	5.8 (3.8)
22 M	ADL Score (0–60)	47.7 (7.2)	53.9 (5.9)	6.3 (5.3)	37.9 (4.3)	41.7 (3.9)	3.8 (3.1)
23 M	ADL Score (0–60)	42.3 (9.6)	50.5 (8.1)	8.2 (6.8)	35.2 (5.5)	40.4 (5.6)	5.2 (3.8)
Total	**ADL Score (0–60)**	**46.5 (10)**	**53.3 (7.6)**	**6.8 (6.5)**	**38.1 (6.4)**	**42.5 (5.7)**	**4.4 (3.8)**
1 M	FIM[TM] (18–126)	92.7 (14.1)	102.8 (11.5)	10.1 (7.4)	82 (8.6)	88.6 (7.8)	6.6 (4.3)
2M[a]	FIM[TM] (18–126)	108.8 (12.6)	116.7 (10.8)	8 (7.7)	95.6 (13.3)	104.5 (12.9)	9 (8.3)
3 M	FIM[TM] (18–126)	93.5 (14.9)	109.2 (11.6)	15.7 (9.7)	82.5 (9)	94.2 (9.2)	11.7 (6.4)
4 M	FIM[TM] (18–126)	99.3 (18.4)	106.7 (16.1)	7.3 (8.2)	87.4 (12.6)	93.3 (12.7)	5.9 (6.1)
5 M	FIM[TM] (18–126)	89.9 (18.8)	104.1 (177)	14.2 (9.4)	80.5 (11)	91 (12.5)	10.5 (7.1)
7 M	FIM[TM] (18–126)	96.7 (18.6)	106.5 (16.9)	9.8 (10.1)	85.4 (12.4)	93.2 (12.9)	7.8 (6.9)
8 M	FIM[TM] (18–126)	90.8 (20.1)	106.3 (16.8)	15.5 (11.4)	81.9 (13.3)	93.3 (13.5)	11.4 (7.6)
9 M	FIM[TM] (18–126)	78.7 (16.8)	96.5 (14)	17.8 (12.9)	74 (9.2)	84.2 (8.4)	10.2 (7.1)
10M[a]	FIM[TM] (18–126)	110.3 (15.3)	115.7 (14.6)	5.3 (1 1.5)	99.1 (16.4)	106 (16.4)	6.9 (11.8)
13 M	FIM[TM] (18–126)	82.2 (20)	103.3 (14.1)	21.1 (13.6)	76.1 (11.1)	89.8 (10.9)	13.7 (7.9)
14 M	FIM[TM] (18–126)	107 (9.3)	11 5.7 (6.3)	8.7 (6.6)	91.7 (7.1)	99.9 (6.5)	8.2 (5.4)
15 M	FIM[TM] (18–126)	88.8 (16.1)	102.1 (12.5)	13.4 (12.1)	79.8 (9.6)	88.6 (9.5)	8.7 (8.2)
16 M[a]	FIM[TM] (18–126)	105.7 (8.3)	117.1 (6.8)	113 (6.7)	90.6 (7)	102.8 (97)	12.2 (7.7)
18 M	FIM[TM] (18–126)	97.5 (17.5)	103.9 (14.6)	6.5 (8.4)	85.7 (11.3)	90.1 (10.4)	4.5 (5.5)
19 M	FIM[TM] (18–126)	90.3 (14.6)	105.4 (12.7)	15.1 (11)	80.5 (8.6)	91.2 (9.9)	10.6 (8.2)
21 M	FIM[TM] (18–126)	95.2 (17.7)	108.8 (14.2)	13.5 (10.6)	84 (10.9)	94.4 (11)	10.4 (6.8)
23 M	FIM[TM] (18–126)	91.6 (15.8)	104.9 (135)	13.3 (10.4)	813 (10)	90.7 (10.1)	9.3 (6.9)
Total	**FIM[TM] (18–126)**	**95.5 (17.2)**	**107.7 (13.7)**	**12.1 (10.2)**	**84.4 (1 1.3)**	**93.5 (11)**	**9.1 (7)**
6 M	EBI (0–64)	59.2 (9.1)	62.3 (4.5)	3 (6.3)	59.4 (8.4)	61.6 (4.9)	2.2 (4.6)
11 M	EBI (0–64)	53.4 (8.9)	58.8 (5.4)	5.4 (6.1)	52.7 (7)	57.3 (5)	4.6 (4.5)
12 M	EBI (0–64)	48.2 (9.7)	56.9 (7.9)	8.7 (7)	48.5 (7.4)	55.6 (6.6)	7.1 (5.4)
17 M	EBI (0–64)	54 (9.4)	58 (7.2)	4 (5.9)	53.2 (7.9)	56.7 (6.5)	3.5 (4.8)
20 M	EBI (0–64)	54.1 (7.8)	58 (5.8)	4 (4.6)	52.9 (6.6)	56.2 (5.4)	3.3 (3.5)
22 M	EBI (0–64)	49.6 (7.4)	55.8 (6.2)	6.2 (5.1)	48.9 (5.6)	53.8 (5.1)	4.9 (4)
Total	**EBI (0–64)**	**52.8 (9.1)**	**58 (6.4)**	**5.2 (5.8)**	**52.2 (7.8)**	**56.4 (5.9)**	**4.2 (4.4)**
1 N	ADL Score (0–60)	46.2 (13)	49.2 (11.7)	3 (5.9)	37.3 (7.7)	39.5 (7.4)	2.2 (3.3)
2 N	ADL Score (0–60)	41.4 (16.7)	45.6 (15.4)	4.3 (7.6)	35.4 (11.6)	38 (10.4)	2.6 (6.3)
3 N	ADL Score (0–60)	31.8 (16)	42.2 (15.9)	10.3 (9.6)	29.5 (9.8)	35.7 (9.9)	6.2 (5.5)
4 N	ADL Score (0–60)	45.7 (14.5)	51.4 (11.2)	5.7 (9.3)	37.7 (9.3)	41.3 (7.4)	3.5 (6)
5 N	ADL Score (0–60)	31.5 (17.4)	42.4 (16.9)	10.9 (9.2)	28.6 (1 0.8)	34.9 (9.9)	6.3 (5.1)
[a]6N	ADL Score (0–60)	44.9 (15.5)	47.5 (16)	2.6 (6)	39.7 (11.8)	43.5 (14.6)	3.8 (6.6)
7 N	ADL Score (0–60)	41.2 (14.9)	47.9 (13.1)	6.6 (8.6)	35.4 (9.2)	39.8 (8.3)	4.4 (5.3)

Table 2 Continued

Clinic Nr.	Outcome measure (score range)	Admission score ordinal (SD)	Discharge score ordinal (SD)	Change score ordinal (SD)	Admission score metric (SD)	Discharge score metric (SD)	Change score metric (SD)
8 N	ADL Score (0–60)	33.2 (16.1)	46.2 (15.1)	13 (11.6)	30.1 (8.9)	37.8 (9.6)	7.7 (6.2)
9 N	ADL Score (0–60)	30.7 (15.2)	40.6 (16.3)	9.9 (10.2)	28.6 (8.7)	34.3 (9.6)	5.7 (6.4)
10 N	ADL Score (0–60)	42 (14)	44.6 (14.1)	2.6 (5.6)	35.3 (8.5)	37.1 (8.5)	1.9 (3.2)
11 N	ADL Score (0–60)	24.1 (19.2)	33.3 (20.2)	9.2 (13.6)	23.1 (1 3.9)	29.8 (13.8)	6.7 (9.1)
12 N	ADL Score (0–60)	40.9 (15.6)	48.3 (13.4)	7.5 (9.7)	35.5 (9.2)	40.2 (8.2)	4.7 (5.5)
13 N	ADL Score (0–60)	30.4 (14.5)	40.7 (12.2)	10.2 (7.2)	28.8 (9)	34.2 (5.6)	5.4 (5)
14 N	ADL Score (0–60)	31.4 (14.7)	43.8 (15.2)	12.4 (1 0.4)	29.3 (8.4)	36.2 (8.9)	6.8 (5.5)
15 N	ADL Score (0–60)	37.3 (16.3)	44.6 (15.2)	7.2 (10.9)	32.3 (9)	36.9 (8.7)	4.6 (6.5)
Total	**ADL Score (0–60)**	**37.7 (1 6.4)**	**45.6 (14.9)**	**7.9 (10)**	**32.9 (9.9)**	**37.8 (9.3)**	**4.9 (5.9)**
1 N	FIM (18–126)	96.2 (21.6)	101.5 (19.4)	5.3 (9)	85.2 (13.8)	89.1 (13.3)	3.9 (6)
3 N	FIM (18–126)	73.9 (27.3)	90.9 (26.5)	17 (15.3)	71.1 (17.6)	82.2 (17.9)	11.2 (9.9)
5 N	FIM (18–126)	73 (28.5)	89.9 (26.2)	16.9 (14.4)	69.6 (19.5)	80.8 (17.8)	11.3 (9.2)
[a]6N	*FIM (18–126)*	*97.5 (28.8)*	*102.9 (29.1)*	*5.4 (9.6)*	*89.4 (21.2)*	*96.3 (26.2)*	*6.8 (11.9)*
8 N	FIM (18–126)	75.8 (25.4)	96.7 (24.2)	20.8 (17.4)	72.1 (16)	86 (17.3)	13.9 (11.2)
9 N	FIM (18–126)	71.4 (25)	87.3 (26.2)	15.8 (16.8)	69.5 (15.7)	79.7 (17.4)	10.2 (11.5)
10 N	FIM (18–126)	91 (23.6)	95.6 (23.1)	4.6 (8.9)	81.5 (15.3)	84.8 (15.4)	3.4 (5.8)
11 N	FIM (18–126)	59.7 (33.2)	75.9 (34.8)	16.2 (23.1)	59.6 (25.1)	71.7 (24.9)	12.1 (16.4)
13 N	FIM (18–126)	72.6 (24.8)	88.2 (17.9)	15.6 (123)	69.8 (16.2)	79.5 (10)	9.7 (9)
14 N	FIM (18–126)	73.5 (24.3)	92.7 (24.6)	19.2 (15.8)	708 (15.1)	83.1 (16)	12.3 (9.9)
15 N	FIM (18–126)	82.5 (25.5)	94.4 (23.3)	11.9 (17.3)	76.1 (16.2)	84.5 (15.7)	8.4 (11.7)
Total	**FIM (18–126)**	**76.4 (27.3)**	**92 (25.9)**	**15.5 (17.1)**	**72.3 (17.9)**	**82.8 (17.8)**	**10.5 (11.4)**
2 N	EBI (0–64)	44.4 (17.5)	48.6 (16.1)	4.3 (8.2)	45.6 (15.3)	49.1 (13.6)	3.4 (8.2)
4 N	EBI (0–64)	48 (15.4)	53.7 (12.1)	5.7 (10.2)	48.7 (12.2)	53.3 (9.7)	4.6 (7.9)
7 N	EBI (0–64)	44.2 (15.4)	51.4 (13.7)	7.2 (8.9)	45.5 (12.1)	51.3 (10.9)	5.8 (6.9)
12 N	EBI (0–64)	44.7 (15.3)	52.2 (13.2)	7.5 (9.4)	45.7 (12)	51.9 (10.7)	6.1 (7.2)
Total	**EBI (0–64)**	**44.6 (15.4)**	**51.8 (13.5)**	**7.2 (9.2)**	**45.8 (12.2)**	**51.6 (10.8)**	**5.8 (7.2)**

ADL Score = Activities of Daily Living Score, FIM[TM] = Functional Independence Measure, EBI = Extended Barthel Index, N = N.
[a]Clinics represented in italics show a higher change score when the interval-scaled metric is applied in comparison the majority of clinics, which show a smaller change score when the interval-scaled metric is applied.

Table 3 Examination of floor and ceiling effects

Scale and rehabilitation group	% of people reaching minimum score at admission (N)	% of people reaching maximum score at admission (N)	% of people reaching minimum score at discharge (N)	% of people reaching maximum score at discharge (N)
FIM[TM] MSK	0.0 (1)	0.4 (35)	0.0 (0)	1.3 (103)
FIM[TM] NEUR	2.1 (76)	0.2 (8)	0.8 (31)	1.1 (39)
EBI MSK	0.0 (1)	17[b] (741)	0.0 (0)	21.8[b] (911)
EBI NEUR	0.1 (2)	5.1[a] (114)	0.0 (1)	12.2[a] (273)
ADL Score MSK	0.0 (2)	7.5[a] (917)	0.0 (0)	12.8[a] (1553)
ADL Score NEUR	1.6 (92)	3.1 (182)	0.6 (35)	8.6[a] (508)
Common metric MSK	0.0 (1)	0.3 (35)	0.0 (0)	0.8 (103)
Common metric NEUR	1.3 (76)	0.1 (6)	0.5 (31)	0.6 (34)

MSK = Musculoskeletal rehabilitation, NEUR = Neurological rehabilitation, ADL Score = Activities of Daily Living Score, FIM[TM] = Functional Independence Measure, EBI = Extended Barthel Index.
[a]Indication of ceiling effect (>5%).
[b]Clear ceiling effect (>15%).

common metric's functioning categories with the content of relevant ICF Core Sets.

When the interval-scaled common metric is applied and contrasted to the currently used ordinal-scaled functioning outcomes, change scores on the clinic level tended to be smaller on the common metric but more precisely estimated. The main reason for this is that the units in the ordinal scale are not equal and tend to be smaller in the center of a scale than at the margins [7]. Consequently, when

contrasted with the common metric, patients passing over the center of the ordinal scale pick up raw score points quickly, whereas the opposite is true for those moving across the margins. The metric removes this bias and provides a more accurate estimation of the actually achieved change. Even though it is known that ordinal-level scales lead to over- or underestimation of health-related outcomes [6, 7, 11, 26], many comparable outcome reports do use ordinal-scaled data without considering this fallacy [27, 28]. The results of
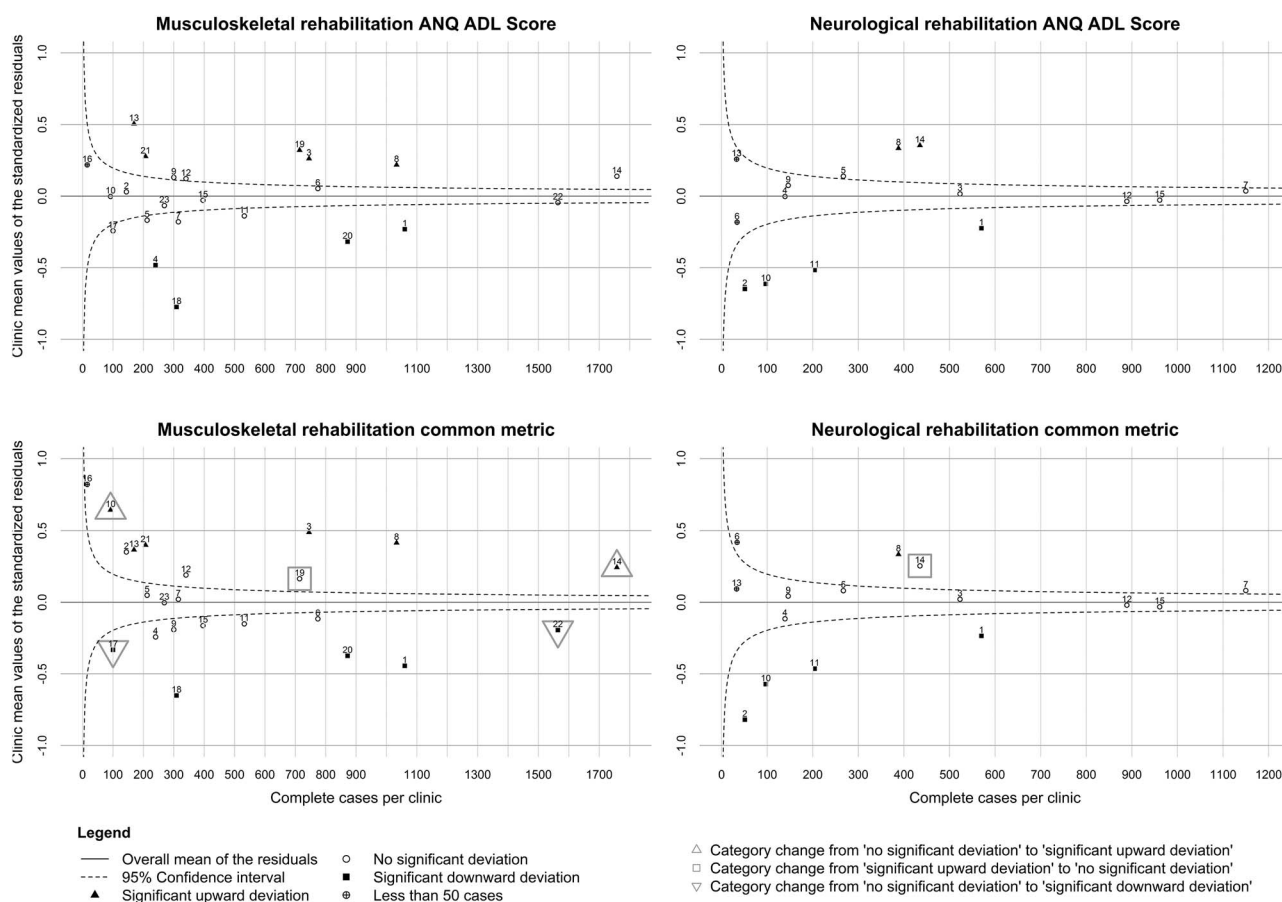
Figure 1 Funnel-plots comparing risk-adjusted clinic performance when using the ANQ-ADL score or the ICF-based interval-scaled common metric. ANQ = Swiss National Association for Quality Development in Hospitals and Clinics, ADL Score = Activities of Daily Living Score, ICF = International Classification of Functioning, Disability and Health.

the present study suggest that the biased ordinal-scaled reporting potentially leads to erroneous clinical decision-making and unfair benchmarking of clinic performance. No statistical hypothesis involving ordinal-scaled data should be tested before the ordinal-scaled data are transformed onto interval-scale level [29].

In the current study, MSK rehabilitation was affected more than NEUR rehabilitation by the difference between ordinal- and interval-scaled reporting approaches, no matter if risk-adjustment was conducted or not. The MSK clinics had baseline scores closer to the scales' upper limits in comparison to the NEUR clinics with scores located more around the center of the scales, indicating the importance that not only change scores on its own, but mean admission and discharge scores should be reported [11]. Furthermore, the MSK sample also showed stronger ceiling effects, especially with EBI, reflecting that FIM^{TM} and EBI are discriminating for the population they were developed for, i.e. the FIM^{TM} [13] for generic rehabilitation and EBI for NEUR rehabilitation [14]. The information of floor and ceiling effects can inform the clinics' decision for a tool most suitable for their specific patient population.

The common metric's ICF basis allowed the comparison with relevant ICF Core Sets, showing potential development opportunities in the functioning outcome indicator included in the current reports in NEUR and MSK rehabilitation such as 'energy and drive functions' but also confirmed relevant functioning outcome aspects that are already represented.

A StARS, with the common metric as the core element, can also be applied for other contexts outside of quality reports such as the comparison of outcome measures in meta-analyses [29]. In any case, a StARS has to be developed for its purpose [30], and it makes sense to consider its influence and added value before its actual implementation.

A limitation of the current study is that the analysis of the influence and added value is at the level of rehabilitation groups. As such, it would be interesting to consider the influence on a more detailed level such as diagnosis-related groups, for example stroke in neurological rehabilitation. A further limitation is that the study is based on a descriptive approach, which helps to describe the differences between the two reporting approaches but does not allow to make statements whether the discovered difference of the common metric is significant or not.

Conclusions

This study shows that it matters if functioning outcomes are reported on ordinal- or interval-scale level. A StARS can help to incorporate several conceptually similar assessment tools into one interval-scaled reporting system, thus enabling the comparison across clinics using different tools, as well as the calculations of means and change scores. Furthermore, the ICF basis of the common metric serves as an opportunity to inform further development of internationally

Table 4 Overview of the comparison of the common metric's ICF categories to relevant ICF Core Sets

ICF Core Set	Generic-7	Generic-30	MSK acute brief	MSK acute comp.	MSK post acute brief	MSK post acute comp.	NEUR acute brief	NEUR acute comp.	NEUR post acute brief	NEUR post acute comp.
Number of ICF categories in Set	7	30	27	48	31	70	33	85	38	116
Coverage of ICF categories with common metric	2 of 7	12 of 30	9 of 27	9 of 48	9 of 31	13 of 70	10 of 33	16 of 85	12 of 38	20 of 116
% of coverage with common metric	28.60%	40.0%	33.3%	18.8%	29.0%	18.6%	30.3%	18.8%	31.6%	17.2%

ICF = International Classification of Functioning, Disability and Health, MSK = Musculoskeletal, NEUR = Neurological, comp. = Comprehensive.

relevant functioning outcome indicators in rehabilitation quality reports.

Supplementary material

Supplementary material is available at *INTQHC Journal* online.

Declaration of no objection

The authors have no conflicts of interests to declare.

References

1. Shaw C, *WHO Regional Office for Europe* (Health Evidence Network), Copenhagen, Denmark. http://www.euro.who.int/document/e82975.pdf (28 February 2020, date last accessed)
2. Stucki G, Bickenbach J. Functioning: the third health indicator in the health system and the key indicator for rehabilitation. *Eur J Phys Rehabil Med* 2017;**53**:134–8.
3. National Academies of Sciences - Engineering - Medicine. *Crossing the global quality chasm: Improving health care worldwide*. Washington (DC): National Academies Press, 2018.
4. World Health Organization. *International Classification of Functioning, Disability and Health (ICF)*. Geneva: WHO, 2001.
5. Stucki G, Bickenbach J, Melvin J. Strengthening rehabilitation in health systems worldwide by integrating information on functioning in National Health Information Systems. *Am J Phys Med Rehabil* 2017;**96**:677–81.
6. Prodinger B, Tennant A, Stucki G. Standardized reporting of functioning information on ICF-based common metrics. *Eur J Phys Rehabil Med* 2018;**54**:110–7.
7. Andrich D. Rating scales and Rasch measurement. *Expert Rev Pharmacoecon Outcomes Res* 2011;**11**:571–85.
8. *ANQ Nationaler Verein für Qualitätsentwicklung in Spitälern und Kliniken*, Bern, Switzerland: https://www.anq.ch/en/departments/rehabilitation/ (28 February 2020, date last accessed)
9. *ANQ Nationaler Verein für Qualitätsentwicklung in Spitälern und Kliniken*, Bern, Switzerland: https://www.anq.ch/de/rehabilitation-neuer-adl-score-macht-fim-und-ebi-resultate-vergleichbar/ (in German) (28 February 2020, date last accessed)

10. Andrich D. The polytomous rasch model and the equating of two instruments. In: KBK C, S. Mesbah M (eds.). *Rasch Models in Health*. Hoboken: ISTE Ltd and John Wiley and Sons, Icn, 2013, 163–96.

11. Stucki G, Daltroy L, Katz JN *et al.* Interpretation of change scores in ordinal clinical scales and health status measures: the whole may not equal the sum of the parts. *J Clin Epidemiol* 1996;**49**:711–7.

12. Maritz R, Tennant A, Fellinghauer C *et al.* Creating a common metric for ADL scores as a basis for standardized reporting of functioning outcomes achieved during rehabilitation. *J Rehabil Med* Status: Submitted December 2019.

13. Keith RA, Granger CV, Hamilton BB *et al.* The functional independence measure: a new tool for rehabilitation. *Adv Clin Rehabil* 1987;**1**:6–18.

14. Prosiegel M, Böttger S, Schenk T *et al.* The extended Barthel index - a new scale for the assessment of disability in neurological patients [German]. *Neurol Rehabil* 1996;**1**:7–13.

15. Mahoney FI, Barthel DW. Functional evaluation: the barthel index. *Maryland State Med J* 1965;**14**:61–5.

16. Maritz R, Tennant A, Fellinghauer C *et al.* The functional independence measure 18-item version can be reported as a unidimensional interval-scaled metric: internal construct validity revisited. *J Rehabil Med* 2019;**51**:193–200.

17. Maritz R, Tennant A, Fellinghauer CS *et al.* The extended Barthel index (EBI) can be reported as a unidimensional interval-scaled metric – a psychometric study. *Phys Med Rehab Kuror* 2019;**29**:224–32.

18. 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* 2019;**41**:574–83.

19. *ANQ Nationaler Verein für Qualitätsentwicklung in Spitälern und Kliniken*, Bern, Switzerland: https://results.anq.ch/fileadmin/documents/anq/27_31/20180913_ANQ_Reha_Nationaler-Vergleichsbericht_Muskuloskelettale-Rehabilitation_2016.pdf https://results.anq.ch/fileadmin/documents/anq/27_31/20180913_ANQ_Reha_Nationaler-Vergleichsbericht_Neurologische-Rehabilitation_2016.pdf (in German) (28 February 2020, date last accessed)

20. Terwee CB, Bot SD, de Boer MR *et al.* Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol* 2007;**60**:34–42.

21. 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.

22. ICF Research Branch, ICF Research Branch, Nottwil, Switzerland: https://www.icf-research-branch.org/icf-core-sets (28 February 2020, date last accessed)

23. Prodinger B, Cieza A, Oberhauser C *et al.* Toward the international classification of functioning, disability and health (ICF) rehabilitation set: a minimal generic set of domains for rehabilitation as a health strategy. *Arch Phys Med Rehabil* 2016;**97**:875–84.

24. 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.

25. Grill E, Strobl R, Muller M *et al.* ICF Core sets for early post-acute rehabilitation facilities. *J Rehabil Med* 2011;**43**:131–8.

26. Ørnbjerg LM, Christensen KB, Tennant A *et al.* Validation and assessment of minimally clinically important difference of the unadjusted health assessment questionnaire in a Danish cohort: uncovering ordinal bias. *Scand J Rheumatol* 2020;**49**:1–7.

27. AROC, Australasian Rehabilitation Outcomes Centre, Wollongong, Australia: https://ahsri.uow.edu.au/content/groups/public/@web/@chsd/@aroc/documents/doc/uow256897.pdf 28 February 2020, date last accessed.

28. *Canadian Institute for Health Information*, Ottawa, Canada: https://www.cihi.ca/en/nrs_pia_en.pdf 28 February 2020, date last accessed.

29. Stucki G, Pollock A, Engkasan JP *et al.* How to use the international classification of functioning, disability and health as a reference system for comparative evaluation and standardized reporting of rehabilitation interventions. *Eur J Phys Rehabil Med* 2019;**55**:384–94.

30. 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.