

ORIGINAL REPORT

CAN THE ICF BE USED AS A REHABILITATION OUTCOME MEASURE? A STUDY LOOKING AT THE INTER- AND INTRA-RATER RELIABILITY OF ICF CATEGORIES DERIVED FROM AN ADL ASSESSMENT TOOL

Friedbert Kohler, MBBS, FAFRM^{1,2}, Carol Connolly, MBChB, FAFRM¹, Aroha Sakaria, Nurse, BAN¹, Kimberly Stendara, PT¹, Mark Buhagiar, B(ap)CPT¹ and Mohammad Mojaddidi, MBBS¹

From the ¹Department of Rehabilitation Medicine, Braeside Hospital, HammondCare and ²School of Public Medicine and Community Health, University of NSW, Sydney, Australia

Purpose: The categories of the International Classification of Functioning, Disability and Health (ICF) could potentially be used as components of outcome measures. Literature demonstrating the psychometric properties of ICF categories is limited.

Objective: Determine the agreement and reliability of ICF activities of daily living category scores and compare these to agreement and reliability of the Functional Independence Measure (FIM) item scores.

Method: Two investigators independently reviewed the clinical notes to score the ICF activities of daily living categories, of 100 patients using ICF qualifiers with additional scoring guidelines. The percentage agreement, interrater and intrarater reliability were compared with the matched FIM items scored by a separate set of two investigators using the same methodology. Kappa Statistic was calculated using Med Calc.

Results: ICF interrater reliability as indicated by Kappa values ranging from 0.42 to 0.81 was moderate or better for the eleven self care and mobility categories. The language ICF categories and problem solving generally have fair agreement, with Kappa values ranging from 0.21 for receiving verbal messages to 0.44 for basic social interactions. Absolute agreement was above 72% for all categories. Reliability and agreement of the FIM items was generally lower than the corresponding ICF categories.

Conclusion: The inter-rater and intra-rater reliability and agreement of the ICF activities of daily living categories were comparable or better than the corresponding FIM items. The results of this study provide an indication that the ICF categories could be used as components of rehabilitation outcome measures.

Key words: ICF; outcome assessment (health care); reproducibility of results; rehabilitation.

J Rehabil Med 2013; 45: 881–887

Correspondence address: Friedbert Kohler, Rehabilitation Medicine, Braeside hospital, 2164 Wetherill Park, Australia.
E-mail: F.Kohler@unsw.edu.au

Accepted Apr 15, 2013; Epub ahead of print Aug 8, 2013

INTRODUCTION

The International Classification of Functioning, Disability and Health (ICF) has been available for about 10 years. It was primarily developed as a comprehensive classification to describe human functioning. As the ICF is very comprehensive it has not been taken up in routine clinical practice. ICF core sets (1, 2) which contain only the ICF categories relevant to a particular health condition, have been developed for several health conditions to facilitate uptake of the ICF in clinical practice. Aligning outcome measures with the various categories contained in the ICF (3–5) and the development of ICF checklists (6) have also increased the utility of the ICF in routine practice. More recently, outcome measures based on the ICF categories have been suggested as feasible (7, 8, 9).

Rehabilitation medicine is the branch of medicine that deals with restoring or improving human functioning. The ICF could potentially be the framework for describing patients who require rehabilitation interventions. If relevant ICF categories could be used as parts of an assessment tool it would allow patient function to be assessed over a much broader spectrum than is now the case. Currently a large range number of assessment tools are used in rehabilitation, but the most commonly used activity of daily living (ADL) measures are the Functional Independence Measure (FIM) (10) and the Barthel ADL Index (11). The FIM is an 18 item instrument. Each item is assessed against a 7 point ordinal scale. The highest score of 7 signifies independence. The lowest score of 1 signifies total dependence.

The FIM and Barthel are particularly important in rehabilitation medicine. The main focus for rehabilitation inpatients is to improve independence and safety in basic ADLs and mobility. Over the last 20 years, ADL measures have been increasingly used to compare outcomes of different therapies and different rehabilitation units (12). More recently, they have been used as the basis for subacute patient classification and funding systems (13, 14). ADL measures which include more domains could increase the scope of assessment tools and classifications.

The ICF is an international classification with over 1,400 categories related to human functioning and is available in

multiple languages. As ICF categories are a comprehensive description of human functioning it is worthwhile exploring their suitability as basic elements for assessment tools. The ICF includes qualifiers to quantify the nature of any problems. The qualifiers range from 0, indicating no problem, to 4, indicating total problem. If the qualifiers can be shown to have acceptable psychometric properties, then this can potentially transform ICF categories into elements of ICF assessment tools. If minor modifications to the qualifier definitions are required to improve reliability of ratings, this should also be explored.

The ICF copyright is owned by the WHO, therefore the ICF is freely available and can be used by any country worldwide without charge. Any ICF based ADL measures would also be free of charge. This has significant advantages for non-industrialised and developing countries, ensuring that their resources are directed mainly to frontline healthcare rather than diverting health dollars and resources to owners of copyrighted ADL instruments. The WHO/World Bank World report on disability (15) suggests that more ICF based outcome instruments should be developed for this reason.

In order to develop a useful and validated outcome measure, the test-retest reliability, interrater and intra-rater reliability, internal consistency, criterion and construct validity, content validity, face validity, floor and ceiling effects and responsiveness (16) should be described. Responsiveness of ICF categories has been demonstrated in a small sample of patients with a lower limb amputation utilising an ICF based mobility measure (9).

On this background, comparing the agreement and reliability of ICF ADL categories with an established, widely used ADL outcome measure such as the FIM, would allow comparison and assist in the development or rebuttal of an argument that ICF categories and qualifiers could be used as component of more comprehensive outcome measures. Interrater and intra-rater reliability are important components of validity.

There are four published papers on the reliability of ICF qualifiers. One paper reviews the test-retest reliability (17), the second paper reviews the interrater as well as the intrarater reliability (18) and the other two papers review the interrater reliability only (19, 20). All studies conclude that the reliability is quite low. This corresponds to our experience in a pilot study in preparation for this study, in which we did not use further descriptors apart from the qualifiers as described in the WHO ICF publication (21).

The aim of this study is to determine the interrater and intrarater reliability of ICF categories corresponding to the ADL aspects addressed by the FIM items, and to compare that reliability to the FIM.

It was not intended to replicate the FIM as an ICF outcome tool but rather to compare the reliability of the FIM items to the corresponding ICF categories.

METHOD

The list of ICF categories was created by following established and published linkage rules to match properties in outcome measures to the ICF (3, 22). A conversion table was drawn up by listing the 18 FIM items and linking them to the most appropriate ICF codes. It was

not possible to directly replicate all the items included in the FIM. Thirteen of the 18 FIM items could be linked directly.

The FIM item of eating incorporates drinking, but in the ICF these are two separate items, so both ICF categories were included in the ICF list.

The 3 FIM items for transfer are best matched by category d410, Changing basic body position. This is defined as: "getting into and out of a body position and moving from one location to another such as getting out of a chair to lie down on bed, and getting into and out of position of kneeling or squatting". As the scores in the 3 transfer items in the FIM are closely related, combining the items is warranted. A separate analysis on the correlation and the agreement between the FIM transfer items was performed to support this.

The other category which includes transfers in the ICF is d420, Transferring while sitting, which is defined as: "moving from one surface to another, such as sliding along a bench or moving from bed to chair without changing body position". This item was initially included in our ICF based ADL tool for the pilot study, but was subsequently excluded as it only applies to a small number of our patients, such as those learning to transfer with a sliding board.

In the pilot study including 30 patients, the items Receiving nonverbal communication and Producing nonverbal communication were not relevant to the patients as they communicated verbally, so these items were not included in further analysis. These items would, however, be relevant for the subgroup of the population with expressive and receptive dysphasia.

The final ICF assessment tool which was developed contains 17 ICF categories. The categories are detailed in Table I with the relevant FIM items listed adjacent.

The agreement between two raters on the pilot project using the original ICF qualifier definitions was mainly poor to fair. Only 3 categories reached moderate agreement. Refinements of qualifier definitions were made in an attempt to increase the reliability of some of the ICF categories (Table II). Additional guidelines for scoring were developed based on the clinical knowledge of the authors and parameters used in established outcome measures. Details for item-specific scoring guidelines are included in Appendix I.

Two investigators, who did not know the patients, independently reviewed and scored each of the 100 patients on each ICF category

Table I. *Functional Independence Measure (FIM) items and the corresponding International Classification of Functioning, Disability and Health (ICF) categories with ICF numeric code and name*

| FIM item | ICF code | ICF item |
|-----------------------|----------|-------------------------------------|
| Eating | d550 | Eating |
| | d560 | Drinking |
| Grooming | d520 | Caring for body parts |
| Bathing | d510 | Washing oneself |
| Dressing upper body | d540 | Dressing |
| Dressing lower body | as above | |
| Toileting | d530 | Toileting |
| Bladder management | b620 | Urination functions |
| Bowel management | b525 | Defecation functions |
| Bed/chair/wc transfer | d410 | Changing basic body position |
| Toilet transfer | as above | |
| Tub/shower transfer | as above | |
| Walking | d4500 | Walking |
| | d465 | Moving around using equipment |
| Stairs | d4551 | Climbing; |
| Comprehension | d310 | Communicate receive spoken messages |
| | | Speaking |
| Expression | d330 | Speaking |
| Social interactions | d710 | Basic interpersonal interactions |
| Problem solving | d175 | Solving problems |
| Memory | b144 | Memory functions |

Table II. Adaptation of International Classification of Functioning, Disability and Health (ICF) Qualifiers to be used as a general guide for scoring

| ICF qualifier score | ICF description | Description used to improve reliability |
|---------------------|---------------------|--|
| 0 | | Independent |
| 1 | Mildly impaired | One person is required for set up or supervision Stand by or minimum assistance |
| 2 | Moderate impaired | Moderate assistance |
| 3 | Severe | Maximal assistance or 2 people are required |
| 4 | Complete impairment | Total assistance |

on admission and discharge. The admission score was based on the entries in the patient’s clinical record during the first 3 days of the admission. In the week following discharge, the last 3 days of entries in the medical records were reviewed and each ICF item was again scored, providing the discharge score. If there was no reassessment in the days prior to discharge – it was assumed that the function was at the level last recorded. The admission scores were not available for review by the investigators at the time of rating the discharge performance. Scores were recorded in paper form and the data entered into a database by a non-clinical member of staff. The investigators were not involved in providing day to day treatment for the patients, so the scores were based solely on the clinical record. A second set of two investigators independently used the same process to score each patient on each item of the FIM. This set of investigators was FIM trained and credentialed. The raters did not have prior knowledge of the patient. In all cases, the score was assigned using the lowest function recorded in the 3 days after admission or the 3 days prior to discharge. Both the admission and the discharge scores were used in the analysis of reliability. All investigators had full access to the patient file and patient identifiers. However, as they did not know the patient this was unlikely to influence patient ratings. The process is shown graphically in Fig. 1.

One investigator from the ICF team and one investigator from the FIM team reviewed all the files after a period of about 3 months and

rescored the patients using the same method as described above to determine the intrarater reliability.

The study was carried out in a 36 bed general rehabilitation unit. The unit casemix consists of approximately 33% stroke patients, 33% orthogeriatric rehabilitation patients and approximately 33% other general rehabilitation patients, including deconditioning following acute illness and geriatric nonspecific physical problems. The mean age is around 70 years, with a mean length of stay of around 25 days.

One hundred consecutive patients admitted to the ward from March 1 2011 were included in the study. Patients who had a length of stay of less than 5 days, or who had an unplanned discharge back to acute care and who were not readmitted within 5 days for the completion of their rehabilitation programme were excluded from the study.

For all FIM items and ICF categories the individual ratings of the two raters were compared for interrater reliability. For intrarater reliability the two scores at different time points were compared. Two statistical tools were used to rate agreement. Raw agreement was determined as the percentage of the scores which were the same between, the two scores for intrarater agreement, and the two raters for interrater agreement. The unweighted Kappa Statistic and the confidence intervals (CIs) were used to quantify statistical agreement for both intrarater and interrater agreement.

To evaluate the validity of the concept of the 3 transfer items being combined, pairwise agreement and Spearman’s rho were calculated comparing the transfer items. The intraclass correlation coefficient (absolute agreement) for the 3 variables with CIs was also determined.

All statistical analyses were performed using MedCalc.

Ethics approval was obtained for this study. As no personal information was gathered and there was no risk involved in this study, waiver of patient consent was granted by the institutional ethics committee. This ensured a complete sample, reducing sampling bias.

RESULTS

No personal data are available regarding the cohort of patients included in this study.

Seven patients were excluded from the study as they were unplanned transfers out of the unit. No clinical data is available for comparison, but it would be expected that these patients were no different from the other patients.

Detailed results of interrater and intrarater agreement and reliability for ICF categories and FIM items using unweighted Kappa values are presented in Tables III and IV.

When interpreting Kappa statistics, we used the published definitions with values less than 0 representing poor agreement, 0.00–0.20 = slight, 0.21–0.40 = fair, 0.41–0.60 = moderate, 0.61–0.80 = substantial, and 0.81–1.00 = almost perfect agreement (23).

Generally, the interrater reliability of both the ICF and the FIM are in the moderate to substantial agreement range.

Interrater agreement ranged from 68% to 93% for the ICF categories. For the FIM items interrater agreement ranged from 48% to 89%. Intrarater agreement ranged from 73% to 93% for ICF categories and 31% to 97% for the FIM items.

The ICF self care and mobility categories have a moderate or higher level of interrater agreement with Kappa values ranging from 0.42 for urination functions to 0.81 for climbing stairs. The ICF categories stair climbing has almost perfect agreement. The categories for eating, caring for body parts, washing, dressing, toileting, changing basic body position and walking have substantial interrater agreement.

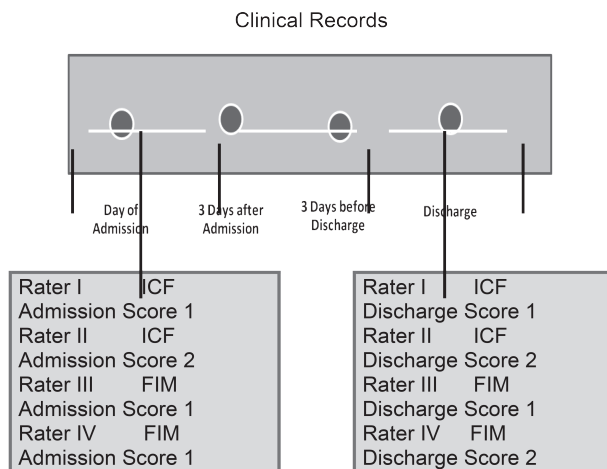


Fig. 1. Flowchart demonstrating time points for International Classification of Functioning, Disability and Health (ICF) and Functional Independence Measure (FIM) scoring.

Table III. Agreement, Kappa values and 95% confidence intervals (CIs) for interrater and intrarater reliability for the International Classification of Functioning, Disability and Health (ICF) categories

| ICF Item | ICF interrater % agreement | ICF interrater reliability Kappa | ICF interrater 95% CI | ICF Intrarater % agreement | ICF intrarater reliability Kappa | ICF intrarater reliability confidence interval |
|----------------------------------|----------------------------|----------------------------------|-----------------------|----------------------------|----------------------------------|--|
| Eating | 90 | 0.73 | 0.61 to 0.84 | 93 | 0.80 | 0.70 to 0.90 |
| Drinking | 85 | 0.47 | 0.30 to 0.63 | 92 | 0.72 | 0.60 to 0.84 |
| Caring for body parts | 84 | 0.71 | 0.62 to 0.80 | 83 | 0.72 | 0.63 to 0.80 |
| Washing oneself | 82 | 0.71 | 0.63 to 0.80 | 84 | 0.70 | 0.60 to 0.80 |
| Dressing | 80 | 0.72 | 0.64 to 0.80 | 87 | 0.81 | 0.75 to 0.88 |
| Toileting | 81 | 0.68 | 0.60 to 0.77 | 80 | 0.65 | 0.56 to 0.74 |
| Urination functions | 84 | 0.42 | 0.28 to 0.57 | 85 | 0.42 | 0.27 to 0.57 |
| Defecation functions | 92 | 0.53 | 0.38 to 0.68 | 93 | 0.57 | 0.39 to 0.76 |
| Changing basic body position | 76 | 0.63 | 0.54 to 0.72 | 84 | 0.76 | 0.69 to 0.84 |
| Walking | 73 | 0.60 | 0.50 to 0.69 | 83 | 0.75 | 0.68 to 0.83 |
| Moving around using equipment | 68 | 0.53 | 0.43 to 0.63 | 78 | 0.68 | 0.60 to 0.77 |
| Climbing | 88 | 0.81 | 0.73 to 0.87 | 88 | 0.81 | 0.73 to 0.90 |
| Receiving messages | 93 | 0.21 | 0.00 to 0.40 | 95 | 0.54 | 0.30 to 0.78 |
| Speaking | 92 | 0.30 | 0.08 to 0.52 | 92 | 0.37 | 0.15 to 0.60 |
| Basic interpersonal interactions | 86 | 0.44 | 0.28 to 0.60 | 82 | 0.38 | 0.26 to 0.50 |
| Solving problems | 73 | 0.36 | 0.24 to 0.50 | 73 | 0.45 | 0.34 to 0.56 |
| Memory functions | 85 | 0.42 | 0.27 to 0.58 | 85 | 0.57 | 0.42 to 0.71 |

CI: confidence interval.

The language ICF categories and problem solving generally have fair agreement, with Kappa values ranging from 0.21 for receiving verbal messages to 0.44 for basic social interactions. This is in line with published findings about several outcome tools including the FIM (24), where language and cognitive items were found to be less reliable than self care and mobility items.

The intrarater reliability demonstrated near perfect agreement for eating, dressing and climbing stairs and substantial agreement for: drinking, caring for body parts, washing, toilet-

ing, changing basic body position, walking and moving around using equipment. As with interrater agreement, the urination functions had less agreement. Language and cognitive ICF categories generally had a similar or higher level of intrarater agreement compared to interrater agreement. All were at the level of fair agreement. Basic personal interactions and memory had moderate agreement.

For all ICF categories the percentage of raw agreement was higher than 80% except for changing basic body position,

Table IV. Agreement, Kappa values and 95% confidence intervals (CIs) for interrater and intrarater reliability for the Functional Independence Measure (FIM) items

| ADL activity | FIM interrater % agreement | FIM interrater reliability Kappa | FIM interrater 95% CI | FIM intrarater percentage agreement | FIM intrarater reliability Kappas | FIM intrarater 95% CI |
|-------------------------------|----------------------------|----------------------------------|-----------------------|-------------------------------------|-----------------------------------|-----------------------|
| Eating | 85 | 0.68 | 0.54 to 0.82 | 97 | 0.91 | 0.82 to 1.01 |
| Caring for body parts | 80 | 0.70 | 0.58 to 0.80 | 97 | 0.95 | 0.89 to 1.01 |
| Washing oneself | 81 | 0.76 | 0.67 to 0.85 | 84 | 0.94 | 0.88 to 1.00 |
| Dressing upper body | 50 | 0.39 | 0.28 to 0.50 | 86 | 0.95 | 0.90 to 1.01 |
| Dressing lower body | 77 | 0.72 | 0.63 to 0.81 | 84 | 0.93 | 0.87 to 0.99 |
| Toileting | 86 | 0.60 | 0.50 to 0.71 | 89 | 0.85 | 0.77 to 0.94 |
| Urination functions | 55 | 0.33 | 0.20 to 0.46 | 31 | 0.10 | -0.04 to 0.24 |
| Defecation functions | 48 | 0.20 | 0.07 to 0.39 | 90 | 0.85 | 0.75 to 0.94 |
| Bath shower transfer | 89 | 0.63 | 0.53 to 0.73 | 93 | 0.90 | 0.84 to 0.98 |
| Bed/Chair/Wheelchair transfer | 75 | 0.69 | 0.60 to 0.79 | 95 | 0.94 | 0.89 to 1.00 |
| Toilet transfer | 67 | 0.56 | 0.45 to 0.67 | 93 | 0.91 | 0.84 to 0.98 |
| Walking short distances | 77 | 0.71 | 0.62 to 0.80 | 93 | 0.92 | 0.85 to 0.98 |
| Climbing | 75 | 0.65 | 0.54 to 0.76 | 94 | 0.92 | 0.85 to 0.99 |
| Comprehension | 63 | 0.37 | 0.23 to 0.52 | 90 | 0.81 | 0.70 to 0.93 |
| Speaking | 77 | 0.12 | -0.16 to 0.40 | 94 | 0.84 | 0.70 to 0.98 |
| Basic personal interactions | 65 | 0.17 | -0.03 to 0.37 | 94 | 0.84 | 0.70 to 0.98 |
| Solving problems | 48 | 0.09 | -0.06 to 0.24 | 94 | 0.82 | 0.67 to 0.98 |
| Memory | 73 | 0.22 | -0.00 to 0.45 | 95 | 0.87 | 0.75 to 1.00 |

ADL: activities of daily living.

Table V. Comparison of scores in the 3 Functional Independence Measure transfer items

| Comparison | Spearman's Rho | Agreement % | Positive difference % | Negative difference % |
|---|----------------|-------------|-----------------------|-----------------------|
| Bed/Chair/WC transfer vs Toilet transfer | 0.829 | 72 | 20 | 8 |
| Bath/Shower transfer vs Toilet transfer | 0.950 | 90 | 7 | 3 |
| Bath/Shower transfer vs Bed Chair/WC transfer | 0.816 | 71 | 10 | 19 |

walking, and moving around using equipment, although these categories showed moderate or substantial agreement by the Kappa statistic. Solving problems also had poor agreement on the Kappa statistic (Table V).

Pairwise comparison of the transfer items showed correlation as determined by Spearman's rho of between 0.816 and 0.950. Absolute agreement between the items ranged from 71% to 90%. The 3-way intraclass correlation coefficient on single measures was 0.89 (95% CI, 0.86 to 0.91). The 3 way correlation coefficient on mean measures was 0.96 (95% CI, 0.95 to 0.97).

DISCUSSION

Using the qualifiers as published in the ICF without any further modification generally resulted in fair to poor inter-rater and intrarater reliability in our pilot study and previously published studies (17–20). This was explicable by the difficulty of translating percentages or broad terminology into exact terms used for rating the patients. For the purpose of this study we chose to give some clearer definitions to some of the qualifiers. The scoring guidelines we used were very simple and easy to follow and would be easy to translate into other languages. We accept that this is a modification to the ICF as published, but feel this is one way which could be considered to improve agreement of ICF category scores. Other ways may be to reduce the qualifiers to fewer response options to potentially 2 or 3 only, as has also been suggested by others (17), and further studies with this approach would be worthwhile.

The level of inter-rater agreement suggests that most self care and mobility ICF categories are sufficiently reliable to be used in an outcome tool. The interrater agreement in our study was considerably higher than the 50% interrater agreement found in a previous study on stroke patients (19) and the 44% found in a study on patients with low back pain (20). The Kappa values in our study are also higher than the median Kappa values of 0.41 and 0.22 in the studies quoted above. The unweighted Kappa values in this study are generally higher than the weighted Kappa values published in a study on patients in geriatric care (17).

Urination functions and defecation functions have only moderate agreement according to the Kappa values but high percentage agreement. The Kappa values may be low because of the Kappa Paradox. It was generally felt that reasonably robust definitions had been developed for these items. In comparison, 3 studies also showed lower Kappas for the body function items compared to the activity and participation items (17, 19, 20).

Lack of clarity of definitions for scores may have contributed to low agreement in some ICF categories. This is particularly likely in those ICF categories where both interrater and intrarater reliability is low, as in speaking. Review and possibly further refinement of the definitions for each score of the ICF would be warranted.

As all the scores were based on the entries in the patients' files, it is possible that in some instances the medical record does not contain sufficient information to allow the patient to be allocated an ICF category or a FIM score with certainty, and that this has resulted in a lack of agreement in some cases. It is particularly likely that this is the case where the reliability is low in both interrater and intrarater reliability of the FIM item and ICF category scores as is the case for urination functions. Review of our clinical documentation may assist in further improvement of these categories.

Comparing the interrater reliability of individual categories between the ICF ADL categories and the FIM items shows that in most cases they are comparable, with both the percentage of agreement and the Kappa values generally being higher in the ICF categories. Notably, defecation function, speaking and the cognitive items had lower interrater reliability in the FIM.

The intrarater reliability of the FIM was near perfect, with Kappa values greater than 0.80 for all items except urination functions. This supports an underlying robustness of the FIM as an outcome measure in standardised, highly skilled and experienced raters. The FIM rater used for this study was an experienced clinician who is our facility FIM trainer. Like many experienced therapists, however, part of their role is administration, so generally this person would not assess many of the FIM scores in our clinical day to day practice. To have such experienced people available to perform the scoring for the majority of patients is not a realistic option.

That the intrarater reliability in the FIM items is higher than the ICF categories is not surprising as experience with an outcome measure should result in better reliability. As ICF categories have not been previously used as scoring tools, and the clinicians involved in the ICF scoring had not used the ICF prior to involvement in this study, it is likely that intra-rater reliability (and probably interrater reliability) would increase over time with practice. Okochi et al. (17) demonstrated that the experience of the clinicians also affected the interrater reliability, with clinicians with more than 8 years of experience having achieving better reliability than those of lesser experience. In this study both the raters collecting data for the interrater reliability part of the study had more than 8 years clinical experience.

Potentially, the results would be different if the scoring was done by direct patient observation rather than scoring from the

patients' files. In particular we note that this deviates from the normal recommended procedure for scoring the patient on FIM items. Scoring from the clinical notes introduces potential confounders, such as the thoroughness of the clinical file review, the interpretation of ambiguously worded descriptions and the problem of less than complete documentation. A further step in determining the reliability of ICF categories would be to repeat a similar study by direct patient observation. However, we feel that for the purpose of this study this is balanced by eliminating clinical performance based variations, particularly for the intrarater reliability component of the study.

While the FIM scores in this study were not obtained in the manner recommended in the FIM manual, we feel that by using a standardised format basing scores on the medical records would not significantly affect the scores between the different raters or the two scores obtained by the individual rater. This rationale is supported by the high percentage of absolute agreement in both the interrater and the intrarater assessments.

Intrarater reliability can be increased by individuals who remember previous scoring. This problem is difficult to eliminate, but we consider that having 3 months between scores and performing admission and discharge scores independently of each other for each of 100 patients would minimise the contribution of remembered scores.

The combination of 3 FIM transfer items was supported by the high correlation between the 3 items on pairwise comparison as well as the intraclass correlation coefficient of 0.96.

Not being able to link some other FIM items, or having the two ICF categories of eating and drinking in the ICF instead of the one FIM item for eating would not have any effect on the reliability of the individual categories items tested.

In conclusion, the interrater and intrarater reliability and agreement of the ICF ADL categories measured were comparable or better than the comparable FIM items.

The results of this study provide an indication that the ICF categories could be used as components of rehabilitation outcome measures. The benefit of this approach is that as the ICF is a very broad classification, there is potential to use its metric properties for the development of significantly more encompassing outcome instruments than we currently use.

While further work is required on other psychometric properties of ICF qualifiers and ICF based measures, based on the demonstrated reliability in this study, there is a potential to develop clinical assessment tools using ICF categories.

ACKNOWLEDGEMENTS

The authors declare no conflicts of interest. All authors are members of the Braeside Rehabilitation Research Group.

REFERENCES

1. Stucki G, Cieza A, Ewert T, Kostanjsek N, Chatterji S, Üstün TB. Application of the International Classification of Functioning, Disability and Health (ICF) in clinical practice. *Disabil Rehabil* 2002; 24: 281–282.
2. Cerniauskaite M, Quintas R, Boldt C, Raggi A, Cieza A, Bickenbach JE, et al. Systematic literature review on ICF from 2001 to 2009: its use, implementation and operationalisation. *Disabil Rehabil* 2011; 33: 281–309.
3. Cieza A, Brockow T, Ewert T, Amman E, Kollerits B, Chatterji S. Linking health-status measurements to the international classification of functioning, disability and health. *J Rehabil Med* 2002; 34: 205–210.
4. Cieza A, Geyh S, Chatterji S, Kostanjsek N, Ustun B, Stucki G. ICF linking rules: an update based on lessons learned. *J Rehabil Med* 2005; 37: 212–218.
5. Scheuringer M, Grill E, Boldt C, Mittrach R, Müllner P, Stucki G. Systematic review of measures and their concepts used in published studies focusing on rehabilitation in the acute hospital and in early post-acute rehabilitation facilities. *Disabil Rehabil* 2005; 27: 419–429.
6. WHO. The ICF checklist: development and application. [document WHO/HFS/CAS/C/02.95] Geneva: World Health Organization; 2002.
7. Grill E, Stucki G. Scales could be developed based on simple clinical ratings of the International Classification of Functioning, Disability and Health Core Set categories. *J Clin Epidemiol* 2009; 62: 891–898.
8. Huber EO, Tobler A, Gloor-Juzi T, Grill E, Gubler-Gut B. The ICF as a way to specify goals and to assess the outcome of physiotherapeutic interventions in the acute hospital. *J Rehabil Med* 2011; 43: 174–177.
9. Kohler F, Xu J, Siva-Withmory C, Arockiam J. Feasibility of using a checklist based on the ICF as an outcome measure in individuals following lower limb amputation. *Prosthet Orthot Int* 2011; 35: 294–301.
10. Hamilton B, Granger C, Sherwin F, Zielezny M, Tashman J. A uniform national data system for medical rehabilitation. In: Fuhrere M, editor. *Rehabilitation outcomes: analysis and measurement*. Baltimore: Brooks; 1987.
11. Mahoney F, Barthel D. Functional evaluation: the Barthel Index. *Md State Med J* 1965; 14: 61–65.
12. Stineman MG, Goin JE, Hamilton BB, Granger CV. Efficiency Pattern Analysis for Medical Rehabilitation. *Am J Med Qual* 1995; 10: 190–198.
13. Stineman MG, Escarce JJ, Goin JE, Hamilton BB, Granger CV, Williams SV. A casemix classification for medical rehabilitation. *Med Care* 1994; 32: 366–379.
14. Eagar K, Gordon R, Hodkinson A, Green J, Eagar L, Erven J, et al. The Australian National Sub-Acute and Non-Acute Patient Classification (AN-SNAP): report of the National Sub-Acute and Non-Acute Casemix Classification Study. Centre for Health Service Development, University of Wollongong; 1997.
15. World Health Organisation. 2011. World report on disability. Geneva: WHO. Available from: http://www.who.int/disabilities/world_report/2011/report/en/.
16. Streiner D, Norman G. *Health measurement scales*. Oxford: Oxford University Press; 1989.
17. Okochi J, Utsunomiya S, Takahashi T. Health measurement using the ICF: Test-retest reliability study of ICF codes and qualifiers in geriatric care. *Health Qual Life Outcomes* 2005; 3: 46.
18. Uhlig T, Lillemo S, Moe RH, Stamm T, Cieza A, Boonen A, et al. Reliability of the ICF Core Set for rheumatoid arthritis. *Ann Rheum Dis* 2007; 66: 1078–1084.
19. Starrost K, Geyh S, Trautwein A, Grunow J, Ceballos-Baumann A, Prosiemel, M, et al. Interrater reliability of the extended ICF Core Set for stroke applied by physiotherapists. *Phys Ther* 2008; 88: 841–851.
20. Hilfiker R, Orbist S, Christen G, Lorenz T, Cieza A. The use of the comprehensive International Classification of Functioning, Dis-

- ability and Health Core Set for low back pain in clinical practice: a reliability study. *Physiother Res Int* 2009; 14: 147–166.
21. World Health Organisation. International Classification of Functioning, Disability and Health. Geneva: WHO; 2001.
 22. Cieza A, Geyh S, Chatterji S, Kostanjsek N, Ustun B, Stucki G. ICF linking rules: an update based on lessons learned. *J Rehabil Med* 2005; 37: 212–218.
 23. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33: 159–174.
 24. Segal ME, Ditunno JF, Staas WE. Inter-institutional agreement of individual Functional Independence Measure (FIM) items at two sites on one sample of SCI patients. *Paraplegia* 1993; 31: 622–631.

APPENDIX I

D 540 Dressing

- If UBD and LBD differ – consider the limitation of whole body e.g. UBD=0, LBD=3 score 2 (give more weight to the greater impairment)

D620 Urination (continence)

- Continent or continent with use of pads/device independently=0
- Needs supervision with pads/device or timed toileting=1
- Assist with pads/device (continent)=2,
- Assist but incontinent (less often than daily)=3
- Assist but incontinent (daily)=4

D525 Defaecation functions

- Continent or continent with independent use of pads=0
- Continent with aperients or timed toileting=1
- Assist with pads/enemas=2
- Assist but incontinent (less than daily)=3
- Assist but incontinent (daily)=4

D420 Transfer while in sitting

- Only applicable if the goal is sliding transfers, e.g. Whilst wheel chair dependent

D4500 Walking (distance plus level of assistance)

- Independent 50+m +/- aid=0
- Walk 15–50m +/- aid=1
- Independent 15+m +/- aid=2
- Supervision/Min assistance=1
- Mod assistance=2
- Two assistance=3
- More than 2 assistance=4

D465 Moving with equipment

- As above
-