**PEER REVIEW HISTORY**

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form [http://bmjopen.bmj.com/site/about/resources/checklist.pdf](http://bmjopen.bmj.com/site/about/resources/checklist.pdf) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

**ARTICLE DETAILS**

<table>
<thead>
<tr>
<th>TITLE (PROVISIONAL)</th>
<th>The ICF Generic Set as new standard for the system wide assessment of functioning in China: A multi-center prospective study on metric properties and responsiveness applying item response theory</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUTHORS</td>
<td>Ehrmann, Cristina; Proding, Birgit; Stucki, Gerold; Cai, Wenzhi; Zhang, Xia; Liu, Shan; Liu, Shouguo; Li, Jianan; Reinhardt, Jan</td>
</tr>
</tbody>
</table>

**VERSION 1 – REVIEW**

| REVIEWER          | Martin Weigl  
|                  | Department of Orthopaedics, Physical Medicine and Rehabilitation, University Hospital, LMU, Munich |
| REVIEW RETURNED   | 20-Feb-2018 |

**GENERAL COMMENTS**

It is an important topic, because the application of the ICF generic core set in acute hospitals would allow a reliable, valid, standardized and practicable assessment of functioning and disability. The statistics are described precisely. However, the statistics are rather complicated and go beyond my statistical knowledge. Accordingly, I suggest a review by a statistician.

Minor comments:

Page 5: Was the study protocol including statistical analysis defined a priori and available for the public? I would suggest commenting on this issue in the methods.

Page 6: The ICF categories were accompanied by "clinically meaningful descriptions". I suggest providing an example for these meaningful descriptions, and not only providing the reference.

Page 6: "... functioning ... was assessed by the same trained Nurse". Accordingly, interrater reliability was not an issue in this study, but it could be a problem in clinical practice. I suggest commenting on this issue in the discussion, probably it is a limitation of the generalizability of this study.

Patient Characteristics: I recommend reporting if patients with surgery were included. This is important, because for example a cancer patients may have no pain at the entry, but may have pain at discharge due to surgery.

Table 3. Reliability - PSI was below 0.70. This appears to be only a moderate reliability. I suggest commenting on this fact in the discussion.

Conclusion: I suggest providing more details for the need of future studies. For example, it would be very interesting if the results could be confirmed in other health care systems and other regions of the world.

| REVIEWER | Mari Klokkerud |
The introduction is well written and provides a good basis for the rationale for the aims of the study. The method section needs further improvement; Design and setting is well described, but all numbers regarding the population (like number excluded) should be removed from the method section and presented in the result section. The study population is also described in a sufficient way. The “Measures and procedure section” needs to be improved. It is difficult to understand on what basis the assessors (nurses) have scored the ICF generic set? My understanding of the description is that they have used both observation and “all relevant available information for each patient”, the latter may include a diversity of anamnesis, tests, PROMS, examinations etc, which include a mixture of different assessors/raters view/scorings and may provide a threat to the reliability of the ICF generic set scoring. This needs to be clarified as it is important both considering reliability and feasibility of the set. This also needs to be addressed in the discussion section, as reliability of a measurement is an important prerequisite for responsiveness. Previous studies, mentioned in the introduction section, addressed the problem of using the qualifier scale of the ICF Generic set. It would be interesting to know whether this study has added more knowledge or solution to this, as this could be both an important barrier to implementation of the set in routine clinical practice and a threat to the reliability of the results. The description of how the Rash analysis is conducted is comprehensive and needs to be reviewed by a statistician or researcher that is expert in Rach methodology.

Responsiveness is a debated term and the methods of how to evaluate this are many. It would have strengthened the article if this was discussed and if the authors have reflected upon the strengths and limitations of the method that you have chosen. COMIN recommendations should be mentioned and referred. This is an important study because it examines psychometric properties of a potentially useful instrument. The large study population strengthens the possibility of reliable conclusions from the statistical analysis, but potential methodological weaknesses should be more thoroughly described, and the conclusion section should also reflect this. Tables and figures are relevant and described in a sufficient way.
are less familiar with the ICF. In my opinion the following issues needs to be addressed:
- The components “body functions and structures”, “activities and participation”, and “environmental factors”, supplemented by “personal factors”.
- Each component, except personal factors, is composed of different domains and each domain is differentiated into subcategories
- The ICF uses an alphanumeric system
- Each letter is followed by a numeric code starting with the chapter number. This is tailed by the second level (three-digit) and the third and fourth level
- An individual person’s health can be documented by selecting the appropriate codes or categories. These can be complemented by assessment criteria that include a numerical code to specify the extent of functioning or disability in that category.

<table>
<thead>
<tr>
<th>REVIEWER</th>
<th>John Starr</th>
</tr>
</thead>
<tbody>
<tr>
<td>University of Edinburgh, UK</td>
<td></td>
</tr>
</tbody>
</table>

| REVIEW RETURNED   | 27-Jun-2018                    |

<table>
<thead>
<tr>
<th>GENERAL COMMENTS</th>
<th>Item response models are well suited to testing the validity of questionnaire scales. Ranch models are well established. I have a few comments:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) The authors do not justify a parametric compared to a non-parametric polytomous analysis. A non-parametric analysis may well provide better data fit and make fewer assumptions though would not allow the construct of a transformed interval score from the raw scores. However, such a transformed score is not much of an advantage.</td>
<td></td>
</tr>
<tr>
<td>2) The scoring moved from a 5-point scale in the pilot to an 11-point scale. This raises major questions about how raters scored each item. With so many points, the differentiation between a score of 6 and a score of 7 might be very limited. Some reporting of an evaluation of this for each item is important therefore, such as item threshold estimates allowing the reader to evaluate discrimination of scores for each item. If there was little discrimination in the mid range, say, one might suggest that the range of score (from 0-10) might sensibly be reduced. That is, increasing the potential range of item scores from the pilot may actually have led to a less valid scale.</td>
<td></td>
</tr>
<tr>
<td>3) There is heterogeneity in this sample which results in differential item functioning between groups. This is a well recognised problem with Rasch models, but it does raise questions about validity of the scale as it is scored here should it be applied to populations with a different case-mix. This is acknowledged by the authors as a limitation.</td>
<td></td>
</tr>
<tr>
<td>4) Some items failed the local independence assumption for Rasch models. This relates to the unidimensionality of the latent trait(s) and the authors ended up splitting item groups into testiest for different conditions. This suggests that an overarching models is not the optimal solution for data fit. Again, Rasch modelling does not set such a high priority on data fit.</td>
<td></td>
</tr>
</tbody>
</table>
| 5) Floor effects. These are clear in the raw data and also are stated to be present in change scores. Mixed regression models
are a reasonable approach, but I think that some model which can
deal with a mixture of present/absent and a scale of how much
present would be more useful here. Perhaps a Tobit regression
analysis might be considered.

6) There are many scales that claim timesaver generic function.
The authors cite a single reference to support the choice of the
ICF generic set with this regard. As someone who looks after
patients with functional deficits in hospital, the items chosen in this
analysis appear to miss the criterion of face validity. For example,
they removed the remunerative employment item as inappropriate
to a hospital setting, but retain the carrying out daily routine item.
However, what happens to a person in hospital can hardly be
termed 'routine'. For example, daily routine might include
shopping, gardening etc. Similarly, sensation of pain is hardly a
function. It may well relate to function (and it does given the local
dependence with other items) but it is not about being able to do
things (function) itself. The choice of items to be included is an
area that the authors must justify a priori, not just statistically,
especially given that they have discarded one item themselves
indicating that they are bound by the ICF Generic Set per se.

7) The ethics sta-
tement is confusing. First it states that the study is
approved by the Chinese Association of Rehabilitation, which I
would be surprised if this were an institution with the appropriate
responsibility to decide rather than provide an opinion. Then it
states that, consequently, no hospital ethics approval was
required. And finally we note that all study participants consented.
I would find it helpful to have some clarification of which authority
gave ethical approval, perhaps with some details of this.

VERSION 1 – AUTHOR RESPONSE

RESPONSES TO REVIEWER#1’s COMMENTS:

Martin Weigl, Institution and Country: Department of Orthopaedics, Physical Medicine and
Rehabilitation, University Hospital, LMU, Munich

General comments
It is an important topic, because the application of the ICF generic core set in acute hospitals would
allow are reliable, valid, standardized and practicable assessment of functioning and disability. The
statistics are described precisely. However, the statistics are rather complicated and go beyond my
statistical knowledge. Accordingly, I suggest a review by a statistician.

Minor comments:
Comment 1: Page 5: Was the study protocol including statistical analysis defined a priori and
available for the public? I would suggest commenting on this issue in the methods.

Thank you for this comment. The study protocol including statistical analysis has been defined a priori
and was available to the participating hospitals. We have made the following changes in the Method
section:

The study protocol was available to the participating hospitals.

Comment 2: Page 6: The ICF categories were accompanied by "clinically meaningful descriptions". I
suggest providing an example for these meaningful descriptions, and not only providing the reference.
We have included an example of these meaningful descriptions in addition to the reference in the Measures and procedures section:

Six out of seven categories of the ICF Generic Set (excluding d850-remunerative employment) accompanied by clinically meaningful descriptions (14) 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”.

Comment 3: Page 6: “... functioning .. was assessed by the same trained Nurse”. Accordingly, interrater reliability was not an issue in this study, but it could be a problem in clinical practice. I suggest commenting on this issue in the discussion, probably it is a limitation of the generalizability of this study.

We note several Two major limitations of to our study need to be considered. First, most of the patients had neurological and musculoskeletal conditions, which may limit the generalizability to other 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 generalizability 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.

Comment 4: Patient Characteristics: I recommend reporting if patients with surgery were included. This is important, because for example a cancer patients may have no pain at the entry, but may have pain at discharge due to surgery.

A total of 915 patients with surgery were included in this study. We have introduced the following text in the Results section:

A total of 915 patients underwent a 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).

Comment 5: Table 3. Reliability - PSI was below 0.70. This appears to be only a moderate reliability. I suggest commenting on this fact in the discussion.

Thank you for your suggestion. We have included a comment referring to the reliability of the ICF Generic 6 Set in the Discussion section:

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, 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.

Comment 6: Conclusion: I suggest providing more details for the need of future studies. For example, it would be very interesting if the results could be confirmed in other health care systems and other regions of the world.

We have included more text in the Conclusion section to emphasize the need of confirming our results in other health care systems and other regions of the world:

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. Future studies are needed to evaluate the validity of ICF Generic 6 Set. 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, e.g. 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 health care systems and countries.

RESPONSES TO REVIEWER#2’s COMMENTS:

Reviewer Name: Mari Klokkerud
Institution and Country: Diakonhjemmet Hospital, Dep. Rheumatology National Advisory Unit on Rehabilitation in Rheumatology, Norway

Comment 1: The introduction is well written and provides a good basis for the rationale for the aims of the study.

Thank you for your appreciation.

Comment 2: The method section needs further improvement: Design and setting is well described, but all numbers regarding the population (like number excluded) should be removed from the method section and presented in the result section.

Thank you for your suggestion. We now present all numbers regarding the population in the Result section.

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/or study endpoint. Descriptive information for the sample is provided in Table 1. From the 4510 adult patients included in this study, more than half were male and the mean age was about 58 years.

and delete them from the Study Design and Setting section:

This was a prospective multi-center study conducted from November 5 2014 to February 28 2015. Included were 4510 patients admitted to the departments of Pulmonology, Cardiology, Neurology, Orthopedics, 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). Children, 308 adults with no defined medical diagnosis and 54 adults with no complete data at admission and study endpoint were excluded from this study.

Comment 3: The study population is also described in a sufficient way.

Thank you for your appreciation.

Comment 4: The “Measures and procedure section” needs to be improved. It is difficult to understand on what basis the assessors (nurses) have scored the ICF generic set? My understanding of the description is that they have used both observation and “all relevant available information for each patient”, the latter may include a diversity of anamnesis, tests, PROMS, examinations etc, which include a mixture of different assessors/raters view scorings and may provide a threat to the reliability of the ICF generic set scoring. This needs to be clarified as it is important both considering reliability and feasibility of the set. This also needs to be addressed in the discussion section, as reliability of a measurement is an important prerequisite for responsiveness.

We have changed the Measures and procedure section where we have described on what basis the assessors scored the ICF Generic Set:

In rating each category, the assessors considered previous data routinely collected in the hospital department in question: Instead the assessors were asked to consider all relevant available information for each patient in order to rate each category, these could include information from anamnesis, clinical examinations, single item scales like visual analogue scale for pain, or standardized assessment tools such as the Barthel Index that were routinely used in the hospital department in question.

The interrater reliability of the ICF Generic 6 Set raw score and the interval ICF Generic 6 Set has been evaluated in a follow up study as described in the Discussion section. See our answer to Reviewer 1, comment 3.

Comment 5: Previous studies, mentioned in the introduction section, addressed the problem of using the qualifier scale of the ICF Generic set. It would be interesting to know whether this study has added more knowledge or solution to this, as this could be both an important barrier to implementation of the set in routine clinical practice and a threat to the reliability of the results.

We find this comment very important. We believe that the 11 point numeric rating scale has contributed to the improvement in the PSI reliability of the ICF Generic 6 Set, in contrast to the pilot study. However, future research is needed to confirm this statement. See our answer to Reviewer 1, comment 5.

Comment 6: Responsiveness is a debated term and the methods of how to evaluate this are many. It would have strengthened the article if this was discussed and if the authors have reflected upon the strengths and limitations of the method that you have chosen. COMIN recommendations should be mentioned and referred.

We have introduced COSMIN recommendations for testing the Responsiveness of the ICF Generic 6 Set in the Discussion section:

With regards 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 (41) since such data was not
available. In line with the results of the pilot study, we would expect moderate to large treatment
effects.

Reference section:

41. COSMIN-manual. Available from:

To emphasize that we tested the resposiveness, we have changed the Title and the Abstract of the
study.

Title:
The ICF Generic Set as new standard for the system wide assessment of functioning in China: A
multi-center prospective study on metric properties and responsiveness applying item response
theory
The ICF Generic Set as a new standard for the system wide assessment of functioning: A multi-center
prospective study from Mainland China

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

Comment 7: This is an important study because it examines psychometric properties of a potentially
useful instrument. The large study population strengthens the possibility of reliable conclusions from
the statistical analysis, but potential methodological weaknesses should be more thoroughly
described, and the conclusion section should also reflect this.

We have considered your comment and we have added more text to the Conclusion section as
described in our response to Reviewer 1, comment 6:

Comment 8: Tables and figures are relevant and described in a sufficient way

Thank you for your appreciation.

RESPONSES TO REVIEWER#3’s COMMENTS:
Reviewer Name: Rachel Sommer
Institution and Country: University Medical Center Hamburg-Eppendorf

General comment: The study "The ICF Generic Set as a new standard for the system wide
assessment of functioning: A multi-center prospective study from Mainland China" aimed to examine
the psychometric properties of the ICF Generic Set when used in routine clinical practice to assess
functioning. The manuscript is well written and the study clearly described. The results are of interests
and important.

Thank you for your appreciation.

Comment 1: I only would recommend to describe some more conceptual issues of the ICF and its use
in the introduction section for readers who are less familiar with the ICF. In my opinion the following
issues needs to be addressed:
-the components “body functions and structures”, “activities and participation”, and “environmental factors”, supplemented by “personal factors”.
- Each component, except personal factors, is composed of different domains and each domain is differentiated into subcategories.
- The ICF uses an alphanumeric system.
- Each letter is followed by a numeric code starting with the chapter number. This is tailed by the second level (three-digit) and the third and fourth level.
- An individual person’s health can be documented by selecting the appropriate codes or categories. These can be complemented by assessment criteria that include a numerical code to specify the extent of functioning or disability in that category.

We have introduced the following description of the ICF in the Introduction section:

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 using more than 1400 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-, third- or fourth-level categories.

RESPONSES TO REVIEWER#4’s COMMENTS:
Reviewer Name: John Starr
Institution and Country: University of Edinburgh, UK
General comment: Item response models are well suited to testing the validity of questionnaire scales. Ranch models are well established.

Comment 1: The authors do not justify a parametric compared to a non-parametric polytomous analysis. A non-parametric analysis may well provide better data fit and make fewer assumptions though would not allow the construct of a transformed interval score from the raw scores. However, such a transformed score is not much of an advantage.

In the Statistical analysis section we have argued for the choice of a parametric polytomous analysis. The 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, i.e. 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).

Reference section:

Comment 2: The scoring moved from a 5-point scale in the pilot to an 11-point scale. This raises major questions about how raters scored each item. With so many points, the differentiation between a score of 6 and a score of 7 might be very limited. Some reporting of an evaluation of this for each item is important therefore, such as item threshold estimates allowing the reader to evaluate discrimination of scores for each item. If there was little discrimination in the mid range, say, one might suggest that the range of score (from 0-10) might sensibly be reduced. That is, increasing the potential range of item scores from the pilot may actually have led to a less valid scale.

Thank you for this comment. The movement from a 5 point scale to an 11 point scale showed a slight improvement in the reliability of the ICF Generic 6 Set in terms of PSI. Moreover, we collected additional data on 703 patients with functioning rated by two independent nurses at each time point. A follow up study will report the results of the interrater reliability analysis. See our answer to Reviewer 1, comment 5. We have added the following text in the Discussion section:

Further research studying the interrater reliability, convergent validity, known group validity, and predictive validity of the ICF Generic 6 Set is underway.

We have also added text in the Introduction section to emphasize the difference between the 5-point Likert scale and the 11 point rating scale:

To address the above issues, a large multi-center study was conducted as a follow up. (i) Instead of the generic ICF qualifier scale, where each qualifier is defined, a numeric rating scale from zero (no problem) to ten (complete problem), where only the extremes are defined, was used and (ii) clinically meaningful descriptions of ICF categories developed in a consensus conference were employed (14).

Comment 3: There is heterogeneity in this sample which results in differential item functioning between groups. This is a well recognised problem with Rasch models, but it does raise questions about validity of the scale as it is scored here should it be applied to populations with a different case-mix. This is acknowledged by the authors as a limitation.

We agree with your comment. We have addressed the limitation of having a heterogeneous sample by splitting the Activities and Participation testlet and we have generated different scoring algorithms for those health condition groups.

Comment 4: Some items failed the local independence assumption for Rasch models. This relates to the unidimensionality of the latent trait(s) and the authors ended up splitting item groups into testiest for different conditions. This suggests that an overarching models is not the optimal solution for data fit. Again, Rasch modelling does not set such a high priority on data fit.

Indeed, we have applied a complex analytical approach to fit the data to the Rasch model. We have used the testlet approach for accommodating the presence of the local dependency across the items. This approach is appropriate when the local dependency between items within a testlet is moderated. Methodological speaking, the split of the Activities and Participation testlet did not cause a problem as we could statistically adjust for it. The number of significant t-tests was below 5%, indicating unidimensionality when using the testlet approach. Fitting all data to the Rasch model allowed us to obtain an interval ICF Generic 6 Set score so that it can be used to compare functioning across patients from different health condition groups. We have added the following text in the Discussion section:

The DIF between 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.
Comment 5: Floor effects. These are clear in the raw data and also are stated to be present in change scores. Mixed regression models are a reasonable approach, but I think that some model which can deal with a mixture of present/absent and a scale of how much present would be more useful here. Perhaps a Tobit regression analysis might be considered.

Thank you for this comment. Most floor effects disappeared when using the interval ICF Generic 6 Set and the Gaussian mixed effects model did its job in demonstrating sensitivity to change. Indeed, a mixed Tobit regression could be used in cases where floor or ceiling effects play a serious role. We thus added in the limitations section of the Discussion:

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 (42).

and the Reference section:


Comment 6: There are many scales that claim timesaver generic function. The authors cite a single reference to support the choice of the ICF generic set with this regard. As someone who looks after patients with functional deficits in hospital, the items chosen in this analysis appear to miss the criterion of face validity. For example, they removed the remunerative employment item as inappropriate to a hospital setting, but retain the carrying out daily routine item. However, what happens to a person in hospital can hardly be termed ‘routine’. For example, daily routine might include shopping, gardening etc. Similarly, sensation of pain is hardly a function. It may well relate to function (and it does given the local dependence with other items) but it is not about being able to do things (function) itself. The choice of items to be included is an area that the authors must justify a priori, not just statistically, especially given that they have discarded one item themselves indicating that they are bound by the ICF Generic Set per se.

Thank you for your constructive comment. As in the pilot study, we have argued in the Discussion section that Remunerative employment (d850) should be excluded when constructing the interval ICF Generic 6 Set score but nevertheless be assessed. Carrying out daily routine can be assessed in the clinical setting in standardized testing environments. Also, patients carry out activities of daily living such as eating, dressing, and bathing while hospitalized. The relevance of the ICF categories has been confirmed by people with different health conditions and relevant professional bodies.

In practice there is a need of tools to collect functioning information. Although the ICF Generic Set may be criticized for its conciseness, it makes it feasible to collect functioning information routinely and system-wide.

Comment 7: The ethics statement is confusing. First it states that the study is approved by the Chinese Association of Rehabilitation, which I would be surprised if this were an institution with the appropriate responsibility to decide rather than provide an opinion. Then it states that, consequently, no hospital ethics approval was required. And finally we note that all study participants consented. I would find it helpful to have some clarification of which authority gave ethical approval, perhaps with some details of this.
Thanks for this comment. Institutional Review Board approval is only slowly evolving in China. Hospital ethics approval for the data collection was not required at the time of study since data were routinely collected in the hospitals participating in the project of the Chinese Association of Rehabilitation Medicine (the later is actually a government body). Ethical approval is however required for the analysis and publication of the anonymized data for research purposes which was sought and granted. We make the procedure more clear in the respective section now. The project had been approved by the Chinese Association of Rehabilitation Medicine and was exempt from individual hospital ethics approval according to Chinese regulations since it involved non-invasive clinician based assessment of patients based on routinely collected clinical data. The study was nonetheless 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 September 20 2017 (No: NYSZYYEC20170013).

VERSION 2 – REVIEW

<table>
<thead>
<tr>
<th>REVIEWER</th>
<th>Martin Weigl</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>University Hospital Munich, Germany. Department of Orthopedics,</td>
</tr>
<tr>
<td></td>
<td>Physical Medicine and Rehabilitation.</td>
</tr>
<tr>
<td>REVIEW RETURNED</td>
<td>21-Aug-2018</td>
</tr>
</tbody>
</table>

| GENERAL COMMENTS  | The revision has substantially improved the paper.                           |

<table>
<thead>
<tr>
<th>REVIEWER</th>
<th>Mari Klokkerud</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>National Advisory Unit on Rehabilitation in Rheumatology,</td>
</tr>
<tr>
<td></td>
<td>Rheumatology Dep. Diakonhjemmet hospital, Norway</td>
</tr>
<tr>
<td>REVIEW RETURNED</td>
<td>10-Aug-2018</td>
</tr>
</tbody>
</table>

| GENERAL COMMENTS  | The authors of this publication have done a good job improving the          |
|                   | manuscript according to the limitations and need of improvement and        |
|                   | clarifications addressed in the previous review comments. I have no           |
|                   | further comments and think that the paper is now acceptable for publication  |
|                   | in BMJ Open. In the previous review round I asked for statistical review of  |
|                   | the study and that should be done before publication ensuring that the       |
|                   | statistical methods used are correct.                                       |

<table>
<thead>
<tr>
<th>REVIEWER</th>
<th>Rachel Sommer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>University Medical Center Hamburg</td>
</tr>
<tr>
<td>REVIEW RETURNED</td>
<td>02-Aug-2018</td>
</tr>
</tbody>
</table>

| GENERAL COMMENTS  | The authors adressed all comments made in previous review. The manuscript    |
|                   | is suitbale for publication in its current form.                            |

<table>
<thead>
<tr>
<th>REVIEWER</th>
<th>John Starr</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>University of Edinburgh, Scotland</td>
</tr>
<tr>
<td>REVIEW RETURNED</td>
<td>31-Jul-2018</td>
</tr>
</tbody>
</table>

| GENERAL COMMENTS  | Thank you for the revised version and answers to my previous points. I am     |
|                   | happy with all of these except point 6, the content of the scale. The        |
|                   | authors respond that daily routine can be evaluated in hospital, but I am     |
|                   | not convinced of this. Although it is possible to                            |
assess a wide range of ADLs in hospital environments - and I have worked at hospitals that assessed quite unusual ADLs such as ploughing (both by hand and with horses) and coal mining with specialised environments - these environments are not every day. For example a hospital kitchen will not often be that similar to the one used every day by the person being assessed. Of course, assessment in such environments provides useful information about function but it is not an assessment of daily routine. Even if we accepted that specialised hospital environments allowed assessment of everyday ADLS such as gardening and cooking, are the authors saying that this was done for all the participants and would therefore expected to be done for any future patients to score this scale validly? I think it is essential for readers to understand what is being measured here. Similarly, the authors have not addressed my point about pain. Again, pain is an unpleasant sensation that can impact on function, but it is not a function itself. The paper’s title includes the term “metric properties”: a fundamental property is to know what is being measured. It is certainly not purely functioning and this should be flagged up to avoid any misunderstandings.

VERSION 2 – AUTHOR RESPONSE

RESPONSES TO REVIEWER#4’s COMMENTS:

Reviewer Name: John Starr

Institution and Country: University of Edinburgh, UK

Comment 1: Thank you for the revised version and answers to my previous points. I am happy with all of these except point 6, the content of the scale. The authors respond that daily routine can be evaluated in hospital, but I am not convinced of this. Although it is possible to assess a wide range of ADLs in hospital environments - and I have worked at hospitals that assessed quite unusual ADLs such as ploughing (both by hand and with horses) and coal mining with specialised environments - these environments are not every day. For example a hospital kitchen will not often be that similar to the one used every day by the person being assessed. Of course, assessment in such environments provides useful information about function but it is not an assessment of daily routine. Even if we accepted that specialised hospital environments allowed assessment of everyday ADLS such as gardening and cooking, are the authors saying that this was done for all the participants and would therefore expected to be done for any future patients to score this scale validly? I think it is essential for readers to understand what is being measured here.

Thanks again for this important comment, we acknowledge that from how patients perform in ADL in the hospital, performance in managing daily routine once discharged to their homes and communities can only be inferred since home and community environmental feature differ from those of hospitals. We have introduced this as a fifth limitation of the present study in the respective section of the Discussion.
Fifth, although it is possible to assess how a person manages daily routine along a wide range of 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.

Comment 2: Similarly, the authors have not addressed my point about pain. Again, pain is an unpleasant sensation that can impact on function, but it is not a function itself. The paper’s title includes the term "metric properties": a fundamental property is to know what is being measured. It is certainly not purely functioning and this should be flagged up to avoid any misunderstandings.

We have addressed this now in the Discussion section as follows:

While listed as a body function in WHO’s ICF and ICD-11, it may be debated if 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, i.e. 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.

and the Reference section:


**VERSION 3 – REVIEW**

| REVIEWER       | John Starr  
|----------------|-------------
|                | University of Edinburgh, Scotland |
| REVIEW RETURNED| 21-Oct-2018 |
| GENERAL COMMENTS| Thank you for expanding the Limitations section so that readers will be aware of what exactly the scale does and does not measure. |