Measurement properties of the Disability Rating Index in patients undergoing hip replacement

Authors

Helen Parsons[1§], Julie Bruce[2], Juul Achten[2], Matthew L Costa[2] and Nicholas R Parsons[1]

[1] Statistics and Epidemiology, Division of Health Science, Warwick Medical School; University of Warwick; Coventry; CV4 7AL, UK.

[2]Warwick Clinical Trials Unit, Division of Health Sciences, University of Warwick, Coventry CV4 7AL, UK

[§]Corresponding author. Email address: h.parsons@warwick.ac.uk

Short title

Measurement properties of the DRI

Abstract

Objectives: To establish and validate the measurement properties of the Disability Rating Index (DRI) in a population of adults undergoing hip replacement.

Methods: 126 adults participating in a randomized controlled trial completed the Oxford Hip Score, Harris Hip Score, DRI and EQ-5D questionnaires at four time points. The structural validity of the DRI was assessed using principal component analysis. Cronbach's α was used to determine the internal consistency, and scale reliability was also assessed. Correlation between the DRI and the other functional and health-related quality of life scales was used to check criterion validity. The DRI responsiveness was estimated and finally, the interpretability of the scale was also assessed by checking for edge effects.

Results: Results of analyses showed that the DRI was internally consistent (Cronbach's α = 0.92), had good association with both function specific and general health-related quality of life scores and was sensitive to change (smallest detectable change = 2.7). No evidence of edge effects was found. Furthermore, structural assessment of the DRI revealed two novel subscales representing "simple tasks" and "difficult tasks".

Conclusions: The DRI is structurally valid, responsive and concurs with functional assessment in adults undergoing hip replacement.

Keywords

Disability evaluation, patient reported outcomes, validity, DRI

Key messages

The DRI is a valid, reliable patient reported outcome for patients undergoing hip replacement.

The DRI may have wider applications for assessing lower limb function in research studies.

Introduction

Accurate assessment of health-related quality of life and function from the patient perspective is vital when determining the effectiveness of health interventions. Therefore, it is important to use measurement scales which have been correctly tested and validated to avoid the introduction of bias [1, 2]. Whilst many guidelines for validation are available [3-6], it is often difficult to establish whether a particular scale has been rigorously validated and if so, for which patient populations [7].

This paper focuses on the Disability Rating Index (DRI), a patient reported outcome measure developed by Salén *et al* in 1994 [8], with the aim to measure physical disability of respondents within a clinical setting. The DRI evaluates a respondent's disability by assessing activity and participation limitations. The DRI has been used in patients undergoing lumbar disc surgery [9], rehabilitation [10]; women with pelvic girdle pain [11]; upper extremity disorders [12], and patients with chronic pain [13]. The DRI is also a core measure within the Swedish Quality Registry for Pain Rehabilitation [14]. The Salén study validated the DRI on healthy persons (n = 1092) and a heterogeneous clinical population (n=366) with varying levels of disability, including elderly arthritis patients waiting for hip (n=23) or knee (n=24) replacement surgery and wheelchair-bound patients with multiple sclerosis (n=16). The validation consisted of assessing test-retest, inter- and intra- rater reproducibility, internal consistency, construct validity (including face validity) and responsiveness. The scale was found to be highly reliable (Cronbach's α of 0.95 in a one day interval test), acceptable (high compliance in all populations) and correlated well (ρ=0.69) with the patient's rating of their disability. However, the authors did not investigate the structure of the instrument using factor analysis, clinically relevant changes, or consider whether any subscales may exist. In their systematic review, Grotle *et al* [15] recommended that the DRI could be used "without further validation" for patients with lower back pain, but noted that the assessment of the factor structure was "inadequate".

Here, we aim to validate the DRI in a population of younger adults undergoing hip replacement surgery. After a hip replacement, patients value improvements in their ability to participate in activities of daily life, alongside a reduction in pain and stiffness from the hip [16-18]. Therefore the DRI may be a useful measurement tool in this group of patients, and indeed for many other patient groups in whom participation in daily activities is a key outcome. The World Health Organization (WHO) defines disability as "an umbrella term for impairments, activity limitations and participation restrictions" [19], also implying that disability rating could be useful in a wide range of different patient groups. We examined the DRI for internal consistency and structure using principal component analysis and investigated the minimum clinically important change in the DRI occurring after surgical intervention for hip arthritis. The international initiative OMERACT (Outcome Measures in Rheumatology) [2] aims to improve outcome measurement in rheumatology, and scrutinises measures on the three criteria: "truth, discrimination, and feasibility" [20]. Our analyses will help establish both the truth and discrimination criteria.

Methods

Patient Sample

All outcome data were collected within the Warwick Arthroplasty Trial (WAT) (trial registration ISRCTN33354155, UKCRN 4093), described in full elsewhere [21, 22]. Briefly, 126 patients, aged over 18 (ranging from 40 to 69 years old; median age = 57.5 years), 52 (41%) of whom were female, were recruited to compare the effectiveness of total hip replacement with resurfacing hip replacement in patients with severe hip arthritis. Patients were recruited from a large University teaching hospital in the UK. The WAT study was approved by the Coventry research ethics committee (no 07/Q2802/26) on 9 May 2007.

The primary outcome measurements were the Oxford Hip Score [16] and the Harris Hip score [23]; secondary outcome measures included the DRI and EuroQol EQ-5D [24]. Data were collected at baseline (preoperatively), 3 months, 6 months and 12 months after surgery.

The DRI

The DRI questionnaire is self-administered and quick to complete (<5 minutes); it is a single scale comprising of 12 items measured using a visual analogue scale (VAS 0-100). Low scores denote little or no disability, which then increases with the total score. The items are broadly grouped into three distinct sections as follows: (i) Q1-4, Basic activities of daily life: Dressing, Outdoor walks, Climbing stairs, Sitting longer time, (ii) Q5-8, Daily physical activities: Standing bent over a sink, Carrying a bag, Making a bed, Running and (iii) Q9-12, Work related/more vigorous activities: Light work, Heavy work, Lifting heavy objects, and Participating in exercise/sports.

Comparison Instruments

1. Harris Hip Score (HHS)

The Harris Hip Score [23] (HHS) is a clinician based outcome measure, developed in the 1960s to assess function before and after hip surgery. The maximum score (best function) is 100.025, whilst the minimum score (worst function) is 0. A total score of less than 70 is considered indicative of poor function, 70–79 is considered fair, 80–89 is good and 90–100 is excellent function. The tool is known to suffer from ceiling effects [25]. For the WAT study, a change of 7 points was considered to be a minimal clinically important difference (MCID) [21].

2. Oxford Hip Score (OHS)

The Oxford Hip Score [16] is a 12 item patient reported outcome measurement designed to measure hip function. At the point of administration, patients are asked to recall hip function (6 items) and hip pain (6 items, including overall pain level) over the last four weeks. Each item is scored from 0 (worst function) to 4 (best function) with the overall score ranging from 0–48 where 48 represents the best function. A total score of greater than 41 represents excellent function, 34–41 is good, 27–

33 is fair and less than 27 is poor function [26]. A change of 5 points was defined to be a MCID in this population [21].

3. EuroQol (EQ-5D)

The EQ-5D(3L) was developed by the EuroQol Group [24] to give a standard measure of health related quality of life. It is a patient reported outcome measurement and is split into two components: a health status classification (5 item score) and self-rated health (scored by a VAS). The five domains surveyed for the health status are (i) mobility, (ii) self-care, (iii) usual activities, (iv) pain, and (v) anxiety. Responses to the health status classification system are converted into an overall score using a utility algorithm for the UK population [27]. Health status is anchored at 1 (perfect health) to 0 (death); for the UK tariff, health states range from 1.0 (no health problems) to -0.594 (severe health problems). A change of approximately 0.07 points was defined to be a MCID for EQ-5D [28].

Statistical Methods

All analyses were conducted in the statistical software R [29]. Data summaries and descriptive statistics were estimated at all data collection points. Due to the small amount of missing trial data (<4% of all possible participant responses were missing, with no incomplete responses), all analyses were conducted using complete cases only.

Structural Validity

Structural validity is concerned with the construction of a measurement instrument, identifying underlying elements such as the existence of subscales and examining the behaviour of individual items which comprise the instrument. Conceptually, it is a subtype of construct validity, in that it is a check that all aspects of the latent traits are measured. Internal consistency assesses the similarity of items within a test and the relation with the overall score. Terwee et al [6] suggest that analyses should be performed on a sample size of 7 times number of items (this equates to 84 for the DRI) or 100 participants, whichever is greater.

As no information was available on the structure of the DRI, principal component analysis was utilised to investigate the factor composition of the DRI. To ensure that the instrument structure was not biased by repeated responses, it was decided that each participant should be entered into the structural analysis once. However, limiting the analysis to one particular time point was not deemed to be an appropriate sub-set, as not all levels of hip function were represented at a single time point (i.e. respondents entered the study with low levels of function and improved post-operatively). Hence, for each participant in the WAT study (n=126), an instance of the DRI was randomly sampled from the four time-points (baseline, 3 months, 6 months and 12 months). This then yielded a subsample consisting of data from 34 study participants at baseline (27%), 38 at 3 months (30%), 29 at 6 months (23%) and 25 at 12 months (20%). The mean DRI score in the subsample was 37.6 (range; 0 to 91.6). The subsample was then entered into a model with varmax rotation. A scree plot

was used to determine the appropriate number of principal components (PCs). Absolute item loadings greater than 0.45 were considered to represent sufficient correlation with a PC to warrant inclusion within a subscale.

Internal consistency

Using the subsample, the internal consistency of the whole DRI, as well as any identified subscales (extracted PCs), were calculated using Cronbach's α. Values between 0.70 and 0.95 were considered acceptable, indicating items have good to excellent agreement with the score total [6].

Criterion validity

Criterion validity assesses if the instrument has good correlation with other externally validated and well accepted measures [3] (where possible, the "gold standard" measurement), with defined hypotheses stated *a priori* [6]. To investigate external validity of the DRI, Pearson's correlation coefficient was used to assess the association between the DRI and the three other instruments, with an expectation of a moderate to high positive correlation ($\rho \geq 0.7$); as poor hip function was likely to be the main cause of disability for this patient group and poor health-related quality of life is associated with disability.

Reliability, responsiveness and interpretability

The intra-class correlation coefficient (ICC) was used to measure reliability through agreement for continuous measures [3, 4, 6], where reliability is defined as the amount that a scale is influenced by random error [3]. Reliability is ideally calculated using a test-retest format in a short time frame [6], but here we compare "stable" respondents between 6 and 12 months; a respondent was considered stable if the change in clinician-reported HHS between the two time points was less than the MCID (7 points). This subsample of 66 patients (56% of the valid total cohort), was used to assess reliability (ICC) and the absolute measurement error or agreement (standard error of the model; SEM) using a two-way random effects model [6, 30]. An ICC larger than 0.7 is required for good reliability [30].

The interpretability of an instrument relates to how easily qualitative interpretations can be assigned to an instrument's results, whilst its responsiveness is the ability to detect changes in the latent traits under study, usually over time. The COSMIN checklist [4] recommends that any edge effects (ceiling or floor effects) and a MCID are reported as part of interpretability.

Here, we report the responsiveness as the smallest detectable change (SDC), which is the smallest change in scores which can be shown to be greater than measurement error. The SDC can be calculated for both individuals and for groups as $SDC_{ind} = SEM \times 1.96 \times \sqrt{2}$ or $SDC_{group} = 1.96 \times \sqrt{2} \times SEM/\sqrt{n}$, where *n* is the group size [6]. A MCID was calculated by comparing changes in DRI between 6 and 12 months for three groups of patients defined by the HHS: "stable" (change in

HHS <7 points), "improved" (HHS increase by ≥7 points) and "worsened" (HHS decrease ≥7 points). To be considered appropriate, the MCID must be greater than the SDC.

Results

Descriptive statistics of each of the four instruments completed at each time point in the WAT study are shown in Table 1. Item level responses of the DRI can be found in Supplementary Table S1, available at *Rheumatology* online.

Structural Validity and internal consistency

DRI Structure

After examination of the scree plot (Figure 1), it was clear that two factors were sufficient to describe the data as these explained 66% of the total variation within the data. The first component (PC1) accounted for the majority (41%) of the total variation, with the second component (PC2) explaining the remainder. The DRI items were well explained by the model as all item communalities (item correlation with other items of that PC) were greater than 0.5 (Table 2). Hence the two PC model was a good approximation to the DRI data.

Although the DRI was broadly split into three sections by the original authors, the two principal components imply the possible existence of two distinct subscales. Items and their corresponding PCs are shown in Table 2, where it can be seen that sections 1 and 2 correspond to PC1, with section 3 corresponding to PC2. Item 8 (*Running*) and item 9 (*Light work*) are, however, found in a different PC to the rest of their section. This grouping is also reflected in the overall item means (Table 2), with the items in PC1 scoring around 30 points and items in PC2 scoring much higher.

Cronbach's alpha for the full scale was 0.92 (95% CI; 0.89, 0.94), indicating high reliability. The reliability remained high if any item was removed (see Table 2 for details), indicating high inter-item correlation. This is also reflected in the high correlations of individual items and the overall DRI score, the lowest of which, Item 8, was 0.59 (95% CI; 0.46, 0.69). Item standard deviations were all similar, ranging from 23 to 38 points (Table 2).

Subscales

As each item loaded onto a single PC, PC1 and PC2 were both considered subscales of the DRI. As shown in Table 2, PC1 (subscale S1) consisted of all items grouped as "*basic activities of daily life*" and three out of the four items grouped as "*daily physical activities*". These items are all relatively simple tasks. PC2 (subscale S2), however, comprises more complex and difficult tasks: all of the "*work related/vigorous activities"* and *running* from the "*daily physical activities*" group. This lends itself to headings of the subscales of "*Simple tasks"* for S1 and *"Difficult tasks"* for S2.

Figure 2 shows boxplots of the two subscales at each time point, showing that mean scores were higher in S2 than S1. Both subscales reached the lowest estimation of disability (0) at the post-operative time-points, with only S1 reaching 0 preoperatively. S2 (difficult tasks) reached the maximum disability (100) at all data collection points, but the maximum for S1 was 97.7, preoperatively. Subscale S1 had a Cronbach's α of 0.92 (95% CI; 0.89, 0.94), with S2 exhibiting a lower internal consistency of 0.84 (95% CI; 0.77, 0.89). These values both denote good consistency.

Criterion validity

Correlations between DRI and hip function (HHS and OHS), and quality of life (EQ-5D) are shown in

Table 3. The correlations are all negative, as higher values of the HHS, OHS and EQ-5D indicate better function or quality of life, but higher disability for the DRI. The smallest correlation with the HHS was -0.67, however, all confidence intervals include good correlation between the two scales at all times. For the OHS, the smallest correlation was -0.77, indicating good consistency between the two scales. For the DRI and quality of life, the association is fair to good, as the correlation has larger changes between time points and wider confidence intervals. The smallest correlation (-0.59) was at baseline, before hip surgery.

Reliability, responsiveness and interpretability

As shown in Table 1, disability measured using the DRI falls significantly after the cohort undergoes surgery. Compared with baseline preoperative measurements, at the 12 month follow up point, the majority of patients (89.9%) had functional gains as they show a reduction in DRI scores. There were no instances of ceiling effects in the data set as the highest DRI score was 98.5 at baseline (Table 3). There were 9 instances (<2% of all responses) of the minimum DRI recorded, which is far fewer than the 15% of respondents at the lowest score which McHorney and Tarlov [31] suggest as indicative of a floor effect.

The ICC was calculated as 0.86, suggesting good reliability since the majority of the variation in stable respondents occurs between, rather than within subjects. The SEM of the overall DRI score was calculated to be 7.9; SDC_{ind} was 21.9 and SDC_{group} was 2.7. Hence, when using the DRI for clinical assessment of an individual, changes of less than 22 points cannot be distinguished from error, but when using the DRI in groups, much smaller differences are detectable.

The change in DRI between the 6 and 12 month follow-up point by stability group is shown in Figure 3: 66 of the 118 valid respondents (56%) were considered stable, with a mean change in DRI of 0.6 points (sd; 11.2); 26 (22%) were considered to have improved clinically; with a mean DRI decrease of -11.0 (sd; 16.9) and the remaining 26 (22%) had worsened according to their HHS, and had a mean change in DRI of 6.9 (sd; 15.1). This suggests that the smallest difference between clinically stable and unstable patients (i.e. the MCID), in this population is approximately 6.9 points.

Discussion

The DRI is now becoming increasingly important for outcome assessment across a wide spectrum of clinical studies, particularly in musculoskeletal medicine, orthopaedics, surgery and rehabilitation research. The main reason for this is that it provides global, rather than a localized or joint/function specific, outcome assessment. For instance, two large multicentre studies recently funded by the UK National Institute for Health Research (NIHR), WOLLF [32] and FIXDT [33], are both using DRI as the primary outcome in the absence of a more appropriate function specific measure. It is important that we understand the measurements properties of DRI and particularly how it is structured and

relates to other more specific measures. The WAT data provides a particularly rich resource to investigate these issues, and to our knowledge the work presented here is the first structural validation of the DRI. We provide evidence to suggest that the DRI has two separate and distinct subscales, labelled according to the items which they contain as "*Simple tasks"* and *"Difficult tasks".* In the original study by Salén *et al*, the authors sub-divided the instrument into three sections. However, in our population of patients undergoing hip surgery for arthritis, we did not find that these categories aligned to the observed subscales. The original categories of "*daily physical activities"* and "*work related or more vigorous activities*" each had one item split into the other subscale. The problematic items were "*light work*" and "*running*". Further work is required to confirm that this structure is indeed a more general characteristic of the DRI and not an artefact of this particular dataset, or unique to patients suffering from dysfunction of the hip joint.

We have shown that the DRI can detect changes in function postoperatively, with no evidence of edge effects. By comparing the average changes in groups of study participants who experienced clinical changes in function with those who remained clinically stable suggests that the DRI MCID for this population could be as small as 6.9 points. This potential MCID is larger than the SDC_{group}, but not the SDC_{ind}, indicating that in our study population, the DRI was able to assess group changes, but was less sensitive to change on an individual basis. However, this problem of detecting clinically relevant changes for an individual by a PROM is not unique and has been reported elsewhere [34, 35]. Further work is needed to establish a definitive DRI MCID, potentially using multiple methods [7] or anchor points identified as relevant or important to patients [36].

Our analyses showed that there was good association between the DRI and the functional assessment scores, with the OHS correlation greater than the HHS correlation at all times. This may be due to the administration method; the OHS, like the DRI, is self-completed whilst the HHS is administered by a clinician. The correlation between the DRI and the health-related quality of life (EQ-5D) score was not large enough to be considered "good agreement" in the first two data collection points, but increased in later follow up periods. This implies that whilst quality of life and disability may be linked, they are not as closely associated as disability and function in the population under investigation. Furthermore, the EQ5D is a utility based measurement which aims to capture the societal views of health related quality of life.

The WAT study was not designed to validate the DRI, making some aspects of the validation difficult to determine. In particular, MCIDs for each of the subscales were not calculated and the overall MCID given may be an overestimate, as only coarse groupings of "improved" and "worsened" were used, when ideally only patients showing minimal change would be included. Further validation is required using more than 3 anchor points for other patient groups, as well as confirming the MCID. Although the DRI shows potential for use in many different groups of patients, this study was clearly limited to a specific population, and as such caution should be exercised in generalising conclusions and recommendations more widely.

We anticipate that the DRI will correlate well with other functional scores where dysfunction is hypothesised to be a cause of disability e.g. knee osteoarthritis or ankle fracture. Also, the DRI will provide a realistic option for assessing lower limb function for other conditions where no function–specific patient reported outcome measure exists. The DRI may also be of use in studies where the outcome of interest is not restricted to the function of a single joint. For instance, if an intervention aimed to improve mobility in osteoarthritis patients, the DRI could capture change in patients with hip problems, knee problems or a combination of both; an instrument designed to capture single-joint function, although specific, may overlook wider mobility and functional improvement.

In summary, the metrics properties of the DRI explored here lead us to conclude that this outcome measure is robust, reliable and sensitive, and as such we would endorse its use across a wide range of clinical and research settings, including randomized controlled trials, routine patient follow-up, audits and service evaluations.

Competing interests

None declared.

Authors' contributions

MC, JA, JB and NP conceived of the study. HP planned and conducted the analysis. All authors contributed to the interpretation of results, read and approved the final manuscript.

Acknowledgements

This manuscript used data collected during the Warwick Arthroplasty Trial (WAT); which was funded by the Research for Patient Benefit scheme of the National Institute of Health Research (Grant Reference Number PB-PG-0706-10080). The views expressed are those of the authors and not necessarily those of the NHS, the National Institute of Health Research, or the Department of Health.

References

1. Health USDo, Evaluation HSFCfD, Research, Health USDo, Evaluation HSFCfB, Research, et al. Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims: draft guidance. Health and Quality of Life Outcomes. 2006;4(1):79.
2. Tugwell P, Boers M, Brooks P, Simon L, Strand V, Idzerda L. OMERACT: An international initiative to improve outcome measurement in rheumatology. Trials. 2007;8(1):38.
3. Scientific Advisory Committee of the Medical Outcomes Trust. Assessing health status and quality-of-life instruments: attributes and review criteria. Qual Life Res. 2002;11(3):193-205.
4. Mokkink L, Terwee C, Patrick D, Alonso J, Stratford P, Knol D, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Quality of Life Research. 2010;19(4):539-49.
5. Speight J, Barendse SM. FDA guidance on patient reported outcomes. BMJ. 2010;340.
6. Terwee C, Bot S, de Boer M, van der Windt D, Knol D, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. Journal of Clinical Epidemiology. 2007;60:34-42.
7. Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. Journal of Clinical Epidemiology. 2008;61(2):102-9.
8. Salén B, Spangfort E, Nygren A, Nordemar R. The Disability Rating Index: an instrument for the assessment of disability in clinical settings. Journal of Clinical Epidemiology. 1994;47(12)(12):1423-35.
9. Lundin A, Magnuson A, Nilsson O. A stiff and straight back preoperatively is associated with a good outcome 2 years after lumbar disc surgery. Acta orthopaedica. 2009;80(5):573-8.
10. Sjostrom R, Asplund R, Alricsson M. Two-year outcome of a multidisciplinary vocational rehabilitation programme focused on range of motion of the neck and back. Work (Reading, Mass). 2010;37(4):341-8.
11. Grotle M, Garratt AM, Krogstad Jenssen H, Stuge B. Reliability and construct validity of self-report questionnaires for patients with pelvic girdle pain. Physical therapy. 2012;92(1):111-23.
12. Salerno DF, Copley-Merriman C, Taylor TN, Shinogle J, Schulz RM. A review of functional status measures for workers with upper extremity disorders. Occupational and environmental medicine. 2002;59(10):664-70.

13. Bromley Milton M, Borsbo B, Rovner G, Lundgren-Nilsson A, Stibrant-Sunnerhagen K, Gerdle B. Is Pain Intensity Really That Important to Assess in Chronic Pain Patients? A Study Based on the Swedish Quality Registry for Pain Rehabilitation (SQRP). PloS one. 2013;8(6):e65483.
14. Nyberg V, Sanne H, Sjolund BH. Swedish quality registry for pain rehabilitation: purpose, design, implementation and characteristics of referred patients. Journal of rehabilitation medicine : official journal of the UEMS European Board of Physical and Rehabilitation Medicine. 2011;43(1):50-7.
15. Grotle M, Brox JI, Vollestad NK. Functional status and disability questionnaires: what do they assess? A systematic review of back-specific outcome questionnaires. Spine. 2005;30(1):130-40.
16. Dawson J, Fitzpatrick R, Carr A, Murray D. Questionnaire on the perceptions of patients about total hip replacement. J Bone Joint Surg Br. 1996;78(2):185-90.
17. Klassbo M, Larsson E, Mannevik E. Hip disability and osteoarthritis outcome score. An extension of the Western Ontario and McMaster Universities Osteoarthritis Index. Scandinavian journal of rheumatology. 2003;32(1):46-51.
18. Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt LW. Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. The Journal of rheumatology. 1988;15(12):1833-40.
19. World Health Organization. The International Classification of Functioning, Disability and Health (ICF). 2001.
20. Boers M, Brooks P, Strand C, Tugwell P. The OMERACT filter for outcome measures in rheumatology. Journal of rheumatology. 1998;25(2):198-9.
21. Costa M, Achten J, Parsons N, Edlin R, Foguet P, Prakas U, et al. Total hip arthroplasty versus resurfacing arthroplasty in the treatment of patients with arthritis of the hip joint: single centre, parallel group, assessor blinded, randomised controlled trial. BMJ. 2012:344.
22. Achten J, Parsons N, Edlin R, Griffin D, Costa M. A randomised controlled trial of total hip arthroplasty versus resurfacing arthroplasty in the treatment of young patients with arthritis of the hip joint. BMC Musculoskeletal Disorders. 2010;11(1):8.
23. Harris W. Traumatic arthritis of the hip after dislocation and acetabular fractures: treatment by mold arthroplasty. An end-result study using a new method of result evaluation. J Bone Joint Surg Am. 1969;54(1):737-55.
24. EuroQoL Group T. EuroQol-a new facility for the measurement of health-related quality of life. Health Policy. 1990;16(3):199-208.
25. Wamper KE, Sierevelt IN, Poolman RW, Bhandari M, Haverkamp D. The Harris hip score: Do ceiling effects limit its usefulness in orthopedics? Acta orthopaedica. 2010;81(6):703-7.
26. Kalairajah Y, Azurza K, Hulme C, Molloy S, Drabu K. Health outcome measures in the evaluation of total hip arthroplasties: a comparison between the Harris Hip Score and the Oxford Hip Score. J Arthroplasty. 2005;20:1037-41.
27. Dolan P. Modeling valuations for EuroQol health states. Medical care. 1997;35(11):1095-108.
28. Walters SJ, Brazier JE. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. Qual Life Res. 2005;14(6):1523-32.
29. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria; 2013.
30. de Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol. 2006;59(10):1033-9.
31. McHorney C, Tarlov A. Individual-patient monitoring in clinical practice: are available health status surveys adequate? Qual Life Res. 1995;4(4):293-307.
32. A randomised controlled trial of standard wound management versus negative pressure wound therapy in the treatment of adult patients with an open fracture of the lower limb [Internet].

Current Controlled Trials. 2012 [cited 20/01/2014]. Available from: http://www.controlled-trials.com/ISRCTN33756652.
33. UK FixDT – A Randomised Controlled Trial for patients with a displaced fracture of the distal tibia, is there a clinical and cost-effectiveness difference between 'locking' plate fixation and intramedullary nail fixation [Internet]. Current Controlled Trials. 2013 [cited 20/01/2014]. Available from: http://www.controlled-trials.com/ISRCTN99771224.
34. Thorborg K, Holmich P, Christensen R, Petersen J, Roos EM. The Copenhagen Hip and Groin Outcome Score (HAGOS): development and validation according to the COSMIN checklist. British journal of sports medicine. 2011;45(6):478-91.
35. Terwee CB, Roorda LD, Knol DL, De Boer MR, De Vet HCW. Linking measurement error to minimal important change of patient-reported outcomes. Journal of Clinical Epidemiology. 2009;62(10):1062-7.
36. Tubach F, Ravaud P, Martin-Mola E, Awada H, Bellamy N, Bombardier C, et al. Minimum clinically important improvement and patient acceptable symptom state in pain and function in rheumatoid arthritis, ankylosing spondylitis, chronic back pain, hand osteoarthritis, and hip and knee osteoarthritis: Results from a prospective multinational study. Arthritis care & research. 2012;64(11):1699-707.

Figures and Tables

Table 1: Descriptive statistics of Instrument use at each time point.

Instrument	Time point	Values[b]					Change from baseline[a]		
		n	Mean	SD	Min	Max	Mean difference	Significance	Effect size
DRI	Baseline	126	57.5	17.35	11.1	98.5	-	-	-
	3 months	118	34.7	19.31	0.0	89.3	22.6	p<0.01	-1.30
	6 months	120	31.6	22.05	0.0	85.0	26.1	p<0.01	-1.50
	12 months	119	31.4	24.19	0.0	97.1	26.0	p<0.01	-1.50
Harris Hip Score	Baseline	126	49.4	13.80	22.3	78.4	-	-	-
	3 months	119	79.8	17.79	28.0	100.0	30.4	p<0.01	2.20
	6 months	120	85.8	15.09	30.5	100.0	36.3	p<0.01	2.63
	12 months	120	85.2	18.27	27.0	100.0	35.8	p<0.01	2.59
Oxford Hip Score	Baseline	126	19.3	7.82	5	41	-	-	-
	3 months	119	35.9	9.59	7	48	16.6	p<0.01	2.13
	6 months	122	39.4	8.74	12	48	20.2	p<0.01	2.58
	12 months	120	39.2	10.39	6	48	20.0	p<0.01	2.56
EQ5D	Baseline	126	0.3396	0.3357	-0.35	0.76	-	-	-
	3 months	119	0.7092	0.2580	-0.07	1.00	0.368	p<0.01	1.10
	6 months	122	0.7715	0.2668	-0.24	1.00	0.438	p<0.01	1.30
	12 months	120	0.7553	0.3089	-0.18	1.00	0.420	p<0.01	1.25

[a]Significance is of mean change from baseline calculated by a paired t-test. The effect size is calculated as the mean change divided by baseline standard deviation

[b]DRI and HHS values have been rounded to 1dp; no rounding has been applied to EQ5D and OHS values

Table 2: Descriptive statistics for PCA sample (n=126).

DRI Item		PC1 loadings	PC2 loadings	Item Communality	Correlation with total score (95% CI)	Mean (standard deviation)	Cronbach's α if item is removed
Common basic activities of daily life	*Q1: Dressing*	**0.73**	0.24	0.59	0.71 (0.61,0.78)	16 (23)	0.91
	Q2: Outdoor Walks	**0.75**	0.38	0.71	0.82 (0.75,0.87)	30 (31)	0.91
	Q3: Climbing stairs	**0.82**	0.24	0.73	0.78 (0.70,0.84)	26 (31)	0.91
	Q4: Sitting Longer Time	**0.58**	0.43	0.52	0.72 (0.62,0.79)	29 (30)	0.91
More demanding daily physical activities	*Q5: Standing bent over a sink*	**0.77**	0.23	0.65	0.75 (0.66,0.81)	28 (30)	0.91
	Q6: Carrying a bag	**0.80**	0.24	0.69	0.76 (0.68,0.83)	26 (28)	0.91
	Q7: Making a bed	**0.76**	0.19	0.62	0.72 (0.63,0.80)	25 (32)	0.91
	Q8: Running	0.09	**0.82**	0.68	0.59 (0.46,0.69)	76 (33)	0.92
Work-related or more vigorous activities	*Q9: Light work*	**0.79**	0.21	0.67	0.74 (0.65,0.81)	20 (25)	0.91
	Q10: Heavy work	0.43	**0.78**	0.80	0.83 (0.76,0.87)	56 (37)	0.91
	Q11: Lifting heavy objects	0.37	**0.76**	0.71	0.76 (0.68,0.83)	58 (38)	0.91
	Q12: Participating in exercise/sports	0.21	**0.73**	0.57	0.63 (0.51,0.72)	61 (37)	0.92

Note: Loadings with an absolute value above 0.45 are bolded. Cronbach's α for the full scale is 0.92.

Table 3: Edge effects and criterion validity.
Pearson's correlations are shown between quality of life and hip function outcomes and the DRI at each data collection point

Time point	N	HHS Correlation (95% CI)	OHS Correlation (95% CI)	EQ-5D Correlation (95% CI)	DRI Edge effects	
					Range of observed scores	Number of scores at edge (%)
Baseline	126	-0.71, [-0.79, -0.61]	-0.81, [-0.73, -0.86]	-0.64, [-0.73, -0.52]	11.1 - 98.5	0 (0%)
3 months	118	-0.69, [-0.77, -0.57]	-0.79, [-0.72, -0.85]	-0.59, [-0.70, -0.46]	0 - 89.3	3 (2.5%)
6 months	120	-0.67, [-0.76, -0.67]	-0.77, [-0.69, -0.84]	-0.73, [-0.80, -0.63]	0 - 85.0	5 (4.2%)
12 months	119	-0.71, [-0.79, -0.61]	-0.80, [-0.73, -0.86]	-0.76, [-0.83, -0.67]]	0 - 97.1	1 (0.8%)

Figure 1: Scree plot of PCA model of sampled data.

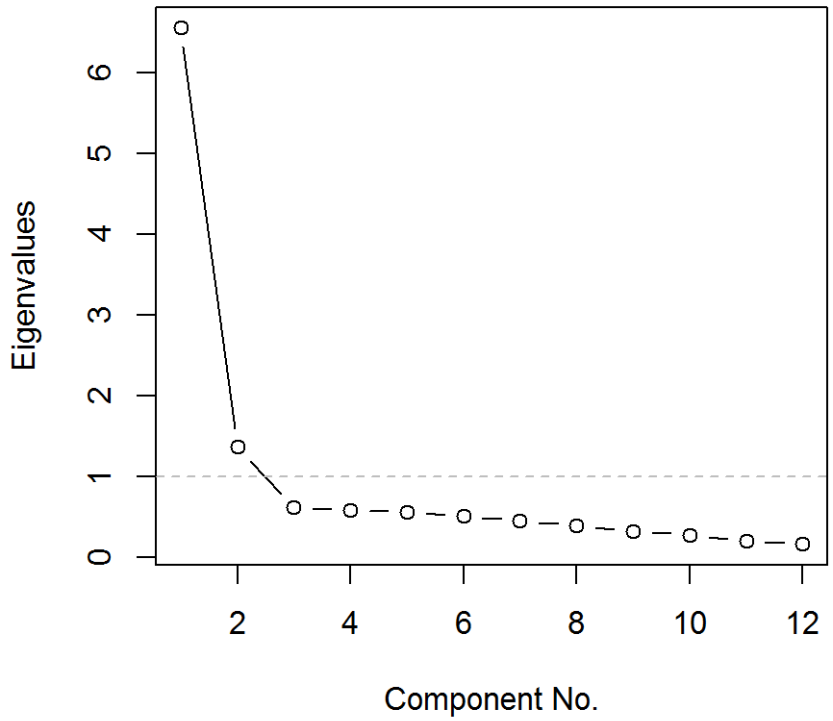

Figure 2: Boxplots of each DRI subscale at each data collection point.
Light grey boxes denote subscale 1, *simple tasks.* Dark grey boxes denote subscale 2, *difficult tasks.*

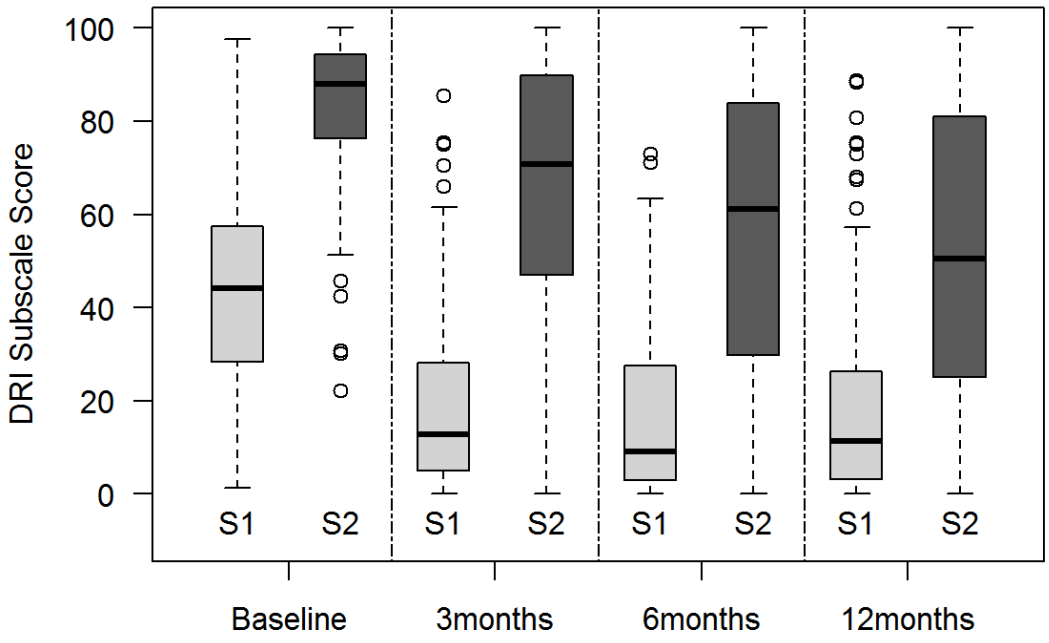

Figure 3: Boxplots of DRI change between the 6 and 12 month follow up point. 66 of the 118 valid respondents (56%) were clinically stable; 26 (22%) had improved; 26 (22%) had worsened.