

## ORIGINAL REPORT

# INTRA-RATER RELIABILITY AND AGREEMENT OF MUSCLE STRENGTH, POWER AND FUNCTIONAL PERFORMANCE MEASURES IN PATIENTS WITH HIP OSTEOARTHRITIS

Theresa Bieler, MSc<sup>1,2</sup>, S. Peter Magnusson, MDSc<sup>1,2</sup>, Michael Kjaer, MDSc<sup>2</sup> and Nina Beyer, PhD<sup>1</sup>

From the <sup>1</sup>Musculoskeletal Rehabilitation Research Unit, Department of Physical & Occupational Therapy, and <sup>2</sup>Institute of Sports Medicine Copenhagen, Copenhagen University Hospital, Bispebjerg, and Center for Healthy Aging, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

**Objective:** To investigate the reliability and agreement of measures of lower extremity muscle strength, power and functional performance in patients with hip osteoarthritis at different time intervals, and to compare these with the same measures in healthy peers.

**Design:** Intra-rater test-retest separated by 1, 2, or 2.5 weeks in patients, and 1 week in healthy peers.

**Subjects:** Patients with hip osteoarthritis (age range 61–83 years) with 1 ( $n=37$ ), 2 ( $n=35$ ), or 2.5 weeks ( $n=15$ ) between tests, and 35 healthy peers (age range 63–82 years).

**Methods:** Maximal isometric hip and thigh strength, leg extensor power, and functional performance (8-foot Up & Go, stair climbing, chair stand and 6-min walk) were measured in patients, and quadriceps strength, leg extensor power and functional performance were measured in healthy peers. Systematic error, reliability and agreement were calculated.

**Results:** Most hip strength measurements for the most symptomatic extremity, and nearly all strength measurements for the least symptomatic lower extremity, declined after 1 week ( $p<0.05$ ), but not after a 2.5-week interval. In healthy peers, quadriceps strength was unchanged. Regardless of the time interval, leg extensor power was unchanged, while functional performances improved at retest for all participants.

**Conclusion:** In patients with hip osteoarthritis leg extensor power is unaffected by the time interval between tests, in contrast to muscle strength and functional performance.

**Key words:** reproducibility of results; osteoarthritis; hip.

*Correspondence address:* Theresa Bieler, Musculoskeletal Rehabilitation Research Unit, Department of Physical & Occupational Therapy, Building 10, Copenhagen University Hospital, Bispebjerg, Bispebjerg Bakke 23, DK-2400 Copenhagen NV, Denmark. E-mail: Theresa.Bieler@regionh.dk

J Rehabil Med 2014; 46: 997–1005

Accepted Apr 25, 2014; Epub ahead of print Aug 6, 2014

## INTRODUCTION

Osteoarthritis (OA) is characterized by pain during physical activity, which may lead to reduced physical activity, resulting in muscle weakness and functional limitations (1). In comparison

with the general population, patients with hip OA experience more limitations in mobility and activities of daily living (2) and a higher level of all-cause mortality, which increases with the severity of walking disability (3). While patients with knee OA primarily display quadriceps weakness (4), data indicate that all muscles in the lower limb may be affected in hip OA (5).

The recommended management for patients with hip OA includes exercise, education, analgesic medication and, if necessary, weight reduction (6). However, there has been little research into the effect of exercise in patients with hip OA (7). In addition, there is limited documentation on the reproducibility of outcome measures to assess muscle function and functional performance in patients with symptomatic hip OA only.

Reproducibility can be distinguished in population-specific reliability and agreement. Reliability assesses whether subjects can be distinguished from each other, despite measurement errors. Agreement assesses the measurement error for repeated measurements (8). An important component of measurement error is systematic bias (e.g. learning effect) (9). Results for repeated performance tests depend on effort and motivation, which can be influenced by prior test experience, desire to improve, decline due to fatigue, and loss of motivation with repeated trials (9, 10). Consequently, the time interval between tests may affect reproducibility (11). In athletes, a time interval longer than 1 day is recommended between repeated measures of performance tests to allow for adequate recovery (9); however, the time interval in reproducibility studies of functional performance in patients with hip or knee OA varies greatly (minutes to weeks) (12). A very short time between test and retest of maximal muscle strength in patients with hip OA may increase the risk of delayed onset muscle soreness, which typically subsides 3–5 days after unaccustomed exercise, and muscle weakness, which may persist for up to 21 days after a single bout of eccentric exercise (13).

Pua et al. (14) documented good to excellent intra-rater reproducibility (intraclass correlation coefficient (ICC) 0.84–0.97, coefficient of variation (CV) 16–8%) of maximal isometric hip muscle strength in persons with hip OA, when tested with at least 1 week (median 19 days) between tests.

However, hip abductor strength improved, which indicates a bias due to a learning effect. Compared with muscle strength, muscle power appears to be more closely related to impaired functional performance in older individuals (15). In a mixed population of patients awaiting hip or knee replacement, a learning effect has been demonstrated for tests of muscle power (16) and functional performance (16, 17), but, to our knowledge, it is not known whether this is the case in a population that only includes patients with hip OA. Finally, some studies have indicated that the reproducibility of muscle strength and functional performance measures may be poorer in patients compared with healthy individuals, but comparing patients with hip OA with healthy controls has yielded conflicting results (18, 19).

The purpose of our study was to investigate the test-retest intra-rater reliability and agreement of maximal isometric hip and thigh muscle strength, leg extensor power (LEP), and 6 functional performance measures in patients 60 years of age and older with hip OA, with time intervals between test and retest of 1, 2 and 2.5 weeks, respectively. A further aim was to compare intra-rater reliability and agreement of maximal isometric knee extensor strength, LEP, and functional performance measures in patients with hip OA and healthy peers.

## MATERIAL AND METHODS

### Participants

This study is part of a randomized trial investigating the effects of exercise in patients with hip OA. Inclusion criteria were: home-dwelling individuals 60 years of age and older with symptomatic hip OA who met the clinical criteria of hip OA according to the American College of Rheumatology (20), but not awaiting hip joint replacement. Exclusion criteria were: (i) symptomatic OA of the knee or the big toe; (ii) knee or hip joint replacement; (iii) other types of arthritis; (iv) previous hip fracture; (v) co-morbidity that prevented exercising; (vi) treatment related to hip problems within the last 3 months; (vii) inability to use public transportation; and (viii) exercising regularly twice or more weekly. All participants provided informed consent. The study was approved by the Danish Ethics Committee of the Capital Region (H-C2009-042).

Patients were recruited through general practitioners and specialists in Greater Copenhagen and advertisements in local newspapers. To reduce the bias due to potential learning effect on the interpretation of the effect of exercise all patients in the randomized trial performed the baseline tests twice (14, 16, 17). The initial data indicated that retest results appeared to be lower in some patients when isometric hip and thigh muscle strength were measured with short time between the 2 baseline tests. Accordingly, to explore the importance of the time interval between tests, we selected all patients with 1 week (6–8 days), 2 weeks (11–15 days) and 2.5 weeks (16–20 days), respectively, between baseline tests for this study. Furthermore, to avoid inter-tester variation we included only those patients who were examined by the same tester.

Healthy participants were recruited through advertisements in local newspapers. Inclusion criteria were: home-dwelling, 60 years of age and older, no self-reported mobility problems, joint pain, and morbidities.

### Methods

Test conditions were standardized with regard to location, test protocol and test order. In patients with hip OA all tests were conducted by the same experienced physiotherapist, who was blinded to the results of the previous measurements. Both test and retest were conducted over 2 days with the following test order: day A – maximal isometric

strength of the knee extensors, knee flexors, hip external rotators, hip internal rotators, and hip flexors, and 15 s marching on the spot test (MOS) (total time 2–2.5 h); day B – maximal isometric strength of the hip abductors and adductors, 8-foot Up & Go test (UG), leg extensor power, timed stair climb test (TSC), timed 5 chair stands (5CT), 30 s chair stand test (30sCS) and 6-min walk test (6MW) (total time 1.5 h). The time between test day A and B was 1–3 days for the first test and 1–2 days for the retest.

In the healthy participants, tests were conducted in 1 day and repeated exactly 7 days later. These tests comprised maximal isometric knee extensor strength, leg extensor power and the 6 functional performance measures (total time 2 h). Four testers, who were trained to complete the test battery in the same standardized way, performed the tests. To avoid inter-tester variation, the same tester carried out all tests in the same individual.

**Muscle function assessments.** The least symptomatic lower extremity (LE) was tested first and, to ensure the correct direction of force production, a sub-maximal practice trial was performed. Participants crossed their hands against their chest and were instructed to perform at their maximum ability. Standardized verbal encouragement and visual feedback were given. Measurements were repeated until a decrease in output occurred. The highest value was used for data analysis. A minimum of 3 repetitions for the strength measurement and 5 repetitions for power measurement was required. Self-reported hip pain prior to and during each measurement was assessed using a numeric pain rating scale from 0 to 10 (21).

**Maximal isometric hip muscle strength.** Maximal isometric hip muscle strength measurements were conducted with a handheld dynamometer (HHD) (JTech Power Track II commander) (14). All tests were made tests (22), and contractions lasted 5 s separated by a rest period of 60 s. The HHD force pad was placed 5 cm above the medial and the lateral malleolus, respectively, except for hip flexor measurements, in which the pad was placed 5 cm above the patella. External and internal rotators and flexors were measured with the patient seated in a straight-back chair with hips and knees flexed at 90°. Stabilization belts were placed across the waist and the ipsilateral thigh distally, but when measuring flexors strength no stabilization belts were used. Abductors and adductors were measured with the patients in the supine position and the hips in the neutral position. Stabilization belts were applied across the pelvis and distally across the contra-lateral thigh.

**Maximal isometric thigh muscle strength.** Maximal isometric thigh muscle strength measurements were conducted with the Good Strength device (Version 3.14 Bluetooth; Metitur Ltd, Finland) (23). The participants were seated with hips flexed at 90° and knees flexed at 60°. Stabilization belts were placed across the waist and distally across the ipsilateral thigh, and the transducer was placed 5 cm above the malleoli. The contraction lasted 5 s and each trial was separated by a rest period of 60 s.

**Leg extensor power.** LEP (force × velocity) measurements were conducted with the Leg Extensor Power Rig (Queen's Medical Centre, Nottingham University, UK) (24) and measured during unilateral lower limb extension. The participants were in a seated position and a single explosive lower limb extension accelerated a flywheel from rest. The participants were instructed to kick the pedal as hard and fast as possible. The extension movement took 0.25–0.4 s, and the final angular velocity of the flywheel was used to calculate the mean LEP during the push (24). Each trial was separated by a rest period of 20 s.

**Functional performance assessments.** A stopwatch, a timer, a tally counter, a straight-back chair without armrest (seat height 44.5 cm), a long measuring tape and 2 cones were used for the 6 functional performance tests. After verbal and visual instructions of the procedures, the participants conducted a sub-maximal practice trial to ensure the correct technique, except for the 6MW test. In all tests, participants were instructed to perform at their maximum.

Table I. Participant characteristics

Characteristics	Older adults with hip OA			Healthy peers
	1	2	2.5	1
Time between tests, weeks, <i>n</i>	1	2	2.5	1
Participants, <i>n</i>	37	35	15	35
Sex, male/female, <i>n</i>	11/26	8/27	8/7	17/18
Age, years, mean (SD)	68 (4)	68 (6)	71 (5)	71 (5) <sup>a</sup>
Height, cm, mean (SD)	168 (8)	169 (7)	170 (11)	171 (9)
Weight, kg, mean (SD)	77 (16)	79 (14)	78 (17)	77 (14)
BMI, kg/m <sup>2</sup> , mean (SD)	28 (6)	27 (5)	27 (4)	26 (3)

<sup>a</sup>The healthy older adults were significantly (*p*=0.038) older than the patients.

OA: osteoarthritis; SD: standard deviation; BMI: body mass index.

Fifteen seconds marching on the spot test (MOS): the number of knee raises completed in 15 s (19). Marking tape on the wall midway between the participant's patella and iliac crest was used to monitor correct knee height (25).

Eight-foot Up & Go test (UG): the time to rise from a chair, walk as quickly as possible around a cone placed 8 feet in front of the chair and return to sit on the chair (25). The best result of 2 timed trials was used for data analysis.

Timed stair climbing test (TSC): the time to ascend and descend a flight of 10 steps (step height 16.3 cm, step depth 35.8 cm) without using the handrail (26). The best result of 2 timed trials was used for data analysis.

Thirty seconds chair stand test (30sCS) and timed 5 chair stands test (5CT): the time to complete the first 5 stands (26) and the number of stands completed in 30 s were measured (27); the participants crossed their hands against their chest.

Six-minute walk test (6MW): the walking (not running) distance completed in 6 min was measured on a 30-m flat course marked with 2

cones (26, 28). For safety reasons the tester walked a couple of metres behind the participant during the test. The participants were told when half of the time had passed and when there was 2 min remaining.

Statistical analysis

Data are presented as means and standard deviations (SD), and the difference in values between test 1 and test 2 (retest) are presented as mean (SD) and 95% confidence interval. Self-reported pain prior to and during the muscle function measurements, and the number of repetitions are presented as median and interquartile range (IQR). A number of statistical methods were employed to assess reproducibility (9). Initially a paired *t*-test was used to explore systematic bias relative to random error (9). Intra-class correlation coefficient (ICC, 2-way random model, consistency definition) was calculated to describe the reliability. Agreement or within-subject variation meaning the random variation in a measure when the individual is tested many times (10) was described with 3 parameters: (i) the standard error of the measurement (SEM) (8), which describes the variation in the same units as the original measurement, (ii) the coefficients of variation (CV) (29), which describes the variation in percent, and (iii) the minimal detectable change at the 90% confidence level (MDC<sub>90</sub>) (30). The data was analysed using SPSS, version 20.

RESULTS

A total of 87 patients with hip OA and 35 healthy peers were included. Participant characteristics are shown in Table I.

Patients with hip osteoarthritis

Muscle strength measurements. In the 1-week group significantly lower retest values were detected for hip external rotators, flexors and abductors in the most symptomatic LE; and for all the hip

Table II. Reliability and agreement of maximal isometric hip muscle strength measurements in the most symptomatic limb and the least symptomatic lower limb in patients with hip osteoarthritis

	The most symptomatic limb			The least symptomatic limb		
	Weeks between test and retest			Weeks between test and retest		
Hip muscle strength (N)	1 ( <i>n</i> =37)	2 ( <i>n</i> =35)	2.5 ( <i>n</i> =15)	1 ( <i>n</i> =37)	2 ( <i>n</i> =35)	2.5 ( <i>n</i> =15)
<i>External rotators</i>						
Test/retest, mean (SD)	87.9 (26.8)/ 80.5 (29.8)	76.9 (24.6)/ 72.9 (27.1)	95.0 (33.7)/ 90.8 (30.9)	95.0 (30.3)/ 87.5 (28.9)	79.2 (21.7)/ 75.0 (26.3)	101.6 (30.4)/ 97.8 (26.0)
Difference, mean (SD)	-7.4 (18.7)*	-4.0 (15.3)	-4.2 (11.7)	-7.4 (13.6)** [-11.9- to -2.9]	-4.1 (15.9)	-3.7 (12.6)
[95% CI]	[-13.7 to -1.2]	[-9.3 to 1.3]	[-10.7 to 2.2]		[-9.6 to 1.3]	[-10.8 to 3.2]
Pain <sup>a</sup>						
Test/retest, mean (SD)	0 (0-0), 0 (0-1)/ 0 (0-0), 0 (0-2)	0 (0-0), 0 (0-2)/ 0 (0-0), 0 (0-0)	0 (0-0), 0 (0-0)/ 0 (0-0), 0 (0-0)	0 (0-0), 0 (0-0)/ 0 (0-0), 0 (0-0)	0 (0-0), 0 (0-0)/ 0 (0-0), 0 (0-0)	0 (0-0), 0 (0-0)/ 0 (0-0), 0 (0-0)
ICC	0.782	0.824	0.935	0.894	0.782	0.900
SEM (N)	13.2	10.8	8.2	9.6	11.2	8.9
CV (%)	16.7	14.8	9.2	11.9	14.9	9.1
MDC <sub>90</sub> (N)	30.9	25.2	19.3	22.4	26.2	20.8
<i>Internal rotators</i>						
Test/retest, mean (SD)	84.0 (24.4)/ 81.2 (29.3)	72.3 (32.4)/ 70.4 (33.0)	92.3 (35.9)/ 89.8 (36.9)	94.4 (25.9)/ 90.8 (27.4)	84.5 (24.6)/ 83.0 (29.4)	103.0 (37.2)/ 92.2 (33.9)
Difference, mean (SD)	-2.8 (15.7)	-1.9 (16.0)	-2.4 (13.4)	-3.6 (13.9)	-1.5 (22.0)	-10.8 (16.2)*
[95% CI]	[-8.0 to 2.5]	[-7.4 to 3.4]	[-9.9 to 5.0]	[-8.2 to 1.0]	[-9.1 to 6.1]	[-19.8 to -1.8]
Pain <sup>a</sup>						
Test/retest, mean (SD)	0 (0-0), 0 (0-3)/ 0 (0-0), 0 (0-2)	0 (0-0), 0 (0-3)/ 0 (0-0), 0 (0-2)	0 (0-0), 0 (0-0)/ 0 (0-1), 0 (0-2)	0 (0-0), 0 (0-0)/ 0 (0-0), 0 (0-0)	0 (0-0), 0 (0-0)/ 0 (0-0), 0 (0-0)	0 (0-0), 0 (0-0)/ 0 (0-0), 0 (0-0)
ICC	0.830	0.880	0.932	0.864	0.670	0.896
SEM (N)	11.1	11.3	9.5	9.8	15.6	11.4
CV (%)	13.5	15.8	10.2	10.8	18.4	13.8
MDC <sub>90</sub> (N)	25.9	26.4	22.1	22.9	36.3	26.7

Table II. *Contd.*

	The most symptomatic limb Weeks between test and retest			The least symptomatic limb Weeks between test and retest		
	1 (n=37)	2 (n=35)	2.5 (n=15)	1 (n=37)	2 (n=35)	2.5 (n=15)
<b>Hip muscle strength (N)</b>						
<i>Flexors</i>						
Test/retest, mean (SD)	117.0 (33.4)/ 109.2 (36.7)	106.6 (50.6)/ 92.3 (35.2)	128.4 (63.2)/ 119.3 (50.8)	126.8 (41.7)/ 119.1 (40.1)	117.9 (48.9)/ 102.5 (43.6)	128.2 (45.5)/ 120.6 (44.8)
Difference, mean (SD) [95% CI]	-7.8 (19.5)* [-14.3 to -1.3]	-14.3 (26.8)** [-23.5 to -5.0]	-9.1 (32.7) [-27.2 to 9.0]	-7.7 (19.7)* [-14.3 to -1.1]	-15.5 (23.1)** [-23.4 to -7.5]	-7.6 (24.2) [-21.0 to 5.7]
<b>Pain<sup>a</sup></b>						
Test/retest, mean (SD)	0 (0-0), 0 (0-2)/ 0 (0-0), 0 (0-2)	0 (0-0), 0 (0-2)/ 0 (0-1), 0 (0-2)	0 (0-0), 0 (0-0)/ 0 (0-0), 0 (0-1)	0 (0-0), 0 (0-0)/ 0 (0-0), 0 (0-0)	0 (0-0), 0 (0-0)/ 0 (0-0), 0 (0-0)	0 (0-0), 0 (0-0)/ 0 (0-0), 0 (0-0)
ICC	0.845	0.811	0.837	0.884	0.875	0.857
SEM (N)	13.8	19.0	23.1	13.9	16.3	17.1
CV (%)	13.0	21.6	18.8	12.1	17.7	14.0
MDC <sub>90</sub> (N)	32.2	44.2	54.0	32.5	38.1	39.9
<i>Abductors</i>						
Test/retest, mean (SD)	64.7 (23.8)/ 60.0 (24.2)	54.4 (24.5)/ 50.3 (24.6)	71.2 (24.9)/ 73.8 (29.0)	74.3 (25.3)/ 69.2 (25.1)	60.1 (23.8)/ 56.3 (24.2)	74.7 (26.5)/ 76.9 (30.3)
Difference, mean (SD) [95% CI]	-4.7 (13.0)* [-9.1 to -0.38]	-4.1 (9.1)* [-7.2 to -0.9]	2.6 (14.0) [-5.1 to 10.3]	-5.0 (11.1)** [-8.7 to -1.3]	-3.7 (11.1) [-7.3 to 0.1]	2.2 (15.6) [-6.4 to 10.9]
<b>Pain<sup>a</sup></b>						
Test/retest	0 (0-1), 2 (0-3)/ 0 (0-2), 1 (0-3)	0 (0-1), 0 (0-2)/ 0 (0-1), 0 (0-3)	0 (0-0), 0 (0-3)/ 0 (0-1), 0 (0-2)	0 (0-0), 0 (0-0)/ 0 (0-0), 0 (0-0)	0 (0-0), 0 (0-0)/ 0 (0-0), 0 (0-0)	0 (0-0), 0 (0-0)/ 0 (0-0), 0 (0-0)
ICC	0.853	0.931	0.867	0.902	0.894	0.849
SEM (N)	9.2	6.4	9.9	7.8	7.8	11.0
CV (%)	15.5	13.3	13.4	12.0	14.0	14.2
MDC <sub>90</sub> (N)	21.5	15.0	23.1	18.3	18.3	25.7
<i>Adductors</i>						
Test/retest, mean (SD)	72.7 (27.8)/ 72.0 (27.8)	59.5 (30.0)/ 58.5 (30.9)	82.4 (30.6)/ 78.4 (29.5)	78.9 (22.3)/ 74.1 (23.7)	63.5 (29.8)/ 62.6 (28.8)	84.3 (25.0)/ 80.6 (24.6)
Difference, mean (SD) [95% CI]	-0.7 (14.6) [-5.6 to 4.1]	-1.0 (11.6) [-6.0 to 3.0]	-3.9 (13.1) [-11.1 to 3.3]	-4.8 (12.4)* [-9.0 to -0.7]	-0.9 (9.9) [-4.3 to 2.5]	-3.7 (8.8) [-8.6 to 1.1]
<b>Pain<sup>a</sup></b>						
Test/retest, mean (SD)	0 (0-2), 1 (1-3)/ 0 (0-2), 0 (0-3)	0 (0-1), 0 (0-2)/ 0 (0-1), 0 (0-1)	0 (0-1), 0 (0-2)/ 0 (0-2), 0 (0-2)	0 (0-0), 0 (0-0)/ 0 (0-0), 0 (0-0)	0 (0-0), 0 (0-0)/ 0 (0-0), 0 (0-0)	0 (0-0), 0 (0-0)/ 0 (0-0), 0 (0-0)
ICC	0.863	0.927	0.905	0.855	0.943	0.937
SEM (N)	10.3	8.2	9.2	8.8	7.0	6.2
CV (%)	12.0	13.8	11.7	12.1	11.0	7.9
MDC <sub>90</sub> (N)	24.1	19.1	21.6	20.5	16.3	14.5

Significant difference between test and retest \* $p < 0.05$ ; \*\* $p < 0.01$ .

<sup>a</sup>Median (interquartile range) pain rating before, and during the measurement.

N: newton; 95% CI: 95% confidence interval; ICC: intraclass correlation coefficient; SEM: standard error of the measurement; CV: coefficients of variation; MDC<sub>90</sub>: minimal detectable change at the 90% confidence level; SD: standard deviation.

muscles in the least symptomatic LE except the internal rotators (Table II). In the 2-week group, significantly lower retest values were detected for hip flexors and abductors in the most symptomatic LE, and hip flexors in the least symptomatic LE. In the 2.5-week group, significantly lower retest value was only detected for the internal rotators in the least symptomatic LE.

There was no significant difference in the most symptomatic LE between test and retest values for the thigh muscle strength irrespective of the time interval between the tests (Table III). In the least symptomatic LE, the retest values for both knee extensors and knee flexors were significantly lower in the 1-week group, while only the retest value for the knee extensors was significantly lower in the 2-week group (Table III). For hip and thigh muscle strength measurements the reliability ranged from ICC=0.67 to 0.94 and the agreement from CV=22% to 6% (Ta-

bles II and III). Overall, the number of repetitions for hip and thigh muscle strength measurements was 4 (3-7) (median (IQR))

In general, self-reported pain was 0 (0-0) prior to and during all muscle strength measurements of the least symptomatic LE and thigh muscle strength measurements of the most symptomatic LE (Tables II-III). In contrast, some patients reported pain during hip muscle strength measurements of the most symptomatic LE (Table II).

*Leg extensor power measurements.* There were no significant differences between test and retest for LEP values (Table III). The reliability parameter ranged from ICC=0.93 to 0.96 and the agreement parameter from CV=11% to 7% (Table III). The number of repetitions for the power measurements ranged from 6-8 (5-10). Overall, self-reported pain was 0 (0-0) prior

Table III. Reliability and agreement of maximal isometric thigh strength and leg extensor power measurements in the most symptomatic limb and the least symptomatic lower limb in patients with hip osteoarthritis

Thigh muscle strength and muscle power	The most symptomatic limb			The least symptomatic limb		
	Weeks between test and retest			Weeks between test and retest		
	1 (n=37)	2 (n=35)	2.5 (n=15)	1 (n=37)	2 (n=35)	2.5 (n=15)
<i>Extensors</i>						
Test/retest, mean (SD)	327.1 (113.0)/ 314.3 (110.4)	297.61 (89.6)/ 281.91 (107.7)	374.91 (140.2)/ 366.61 (113.7)	364.91 (114.2)/ 336.81 (110.5)	331.91 (89.9)/ 315.31 (102.6)	384.21 (137.0)/ 390.51 (97.6)
Difference, mean (SD)	-12.8 (45.9)	-15.81 (56.4)	-8.71 (72.9)	-28.11 (38.6)**	-16.61 (47.2)*	6.31 (64.0)
[95% CI]	[-28.1 to 2.5]	[-35.1 to 3.6]	[-49.1 to 31.9]	[-41.0 to -15.3]	[-32.8 to -0.37]	[-29.1 to 41.8]
Pain <sup>a</sup>						
Test/retest, mean (SD)	0 (0-0), 0 (0-0)/ 0 (0-0), 0 (0-0)	0 (0-0), 0 (0-0)/ 0 (0-0), 0 (0-0)	0 (0-0), 0 (0-0)/ 0 (0-0), 0 (0-0)	0 (0-0), 0 (0-0)/ 0 (0-0), 0 (0-0)	0 (0-0), 0 (0-0)/ 0 (0-0), 0 (0-0)	0 (0-0), 0 (0-0)/ 0 (0-0), 0 (0-0)
ICC	0.916	0.838	0.837	0.941	0.881	0.855
SEM (N)	32.5	39.9	51.5	27.3	33.4	45.3
CV (%)	10.4	14.1	13.5	9.5	10.8	11.3
MDC <sub>90</sub> (N)	75.7	93.1	120.3	63.7	77.9	105.6
<i>Flexors</i>						
Test/retest, mean (SD)	147.91 (56.4)/ 143.91 (54.9)	119.81 (43.8)/ 116.91 (51.6)	180.31 (59.3)/ 176.31 (63.7)	160.91 (57.4)/ 148.51 (59.9)	121.61 (40.4)/ 121.71 (55.6)	166.81 (53.8)/ 175.71 (60.1)
Difference, mean (SD)	-4.01 (38.1)	-2.91 (33.1)	-4.01 (21.9)	-12.41 (28.9)*	0.1 (35.5)	8.81 (29.8)
[95% CI]	[-16.7 to 8.7]	[-14.3 to 8.5]	[-16.1 to 8.1]	[-22.0 to -2.8]	[-12.1 to 12.3]	[-7.7 to 25.3]
Pain <sup>a</sup>						
Test/retest, mean (SD)	0 (0-0), 0 (0-0)/ 0 (0-0), 0 (0-0)	0 (0-0), 0 (0-0)/ 0 (0-0), 0 (0-0)	0 (0-0), 0 (0-0)/ 0 (0-0), 0 (0-0)	0 (0-0), 0 (0-0)/ 0 (0-0), 0 (0-0)	0 (0-0), 0 (0-0)/ 0 (0-0), 0 (0-1)	0 (0-0), 0 (0-0)/ 0 (0-0), 0 (0-0)
ICC	0.765	0.761	0.937	0.879	0.733	0.864
SEM (N)	26.9	23.4	15.5	20.4	25.1	21.1
CV (%)	18.3	19.6	8.5	14.2	20.3	12.4
MDC <sub>90</sub> (N)	62.9	54.6	36.1	47.7	58.6	49.2
<i>Muscle power</i>						
Test/retest, mean (SD)	114.21 (56.2)/ 117.01 (54.5)	112.71 (45.6)/ 110.31 (42.4)	131.21 (53.9)/ 130.41 (55.3)	136.41 (56.0)/ 140.51 (61.8)	131.31 (48.9)/ 132.41 (52.6)	142.91 (41.9)/ 144.51 (51.9)
Difference, mean (SD)	2.81 (17.3)	-2.41 (15.8)	-0.81 (15.1)	4.11 (18.9)	1.11 (18.1)	1.71 (18.1)
[95% CI]	[-3.0 to 8.5]	[-7.8 to 3.0]	[-9.2 to 7.6]	[-2.2 to 10.4]	[-5.2 to 7.3]	[-8.4 to 11.7]
Pain <sup>a</sup>						
Test/retest, mean (SD)	0 (0-0), 0 (0-2)/ 0 (0-0), 0 (0-1)	0 (0-1), 0 (0-1)/ 0 (0-0), 0 (0-1)	0 (0-1), 0 (0-1)/ 0 (0-0), 0 (0-0)	0 (0-0), 0 (0-0)/ 0 (0-0), 0 (0-0)	0 (0-0), 0 (0-0)/ 0 (0-0), 0 (0-0)	0 (0-0), 0 (0-0)/ 0 (0-0), 0 (0-0)
ICC	0.951	0.936	0.962	0.949	0.937	0.926
SEM (W)	12.2	11.2	10.7	13.4	12.8	12.8
CV (%)	10.6	10.0	7.9	9.8	9.6	8.7
MDC <sub>90</sub> (W)	28.5	26.1	24.9	31.2	29.9	29.9

Significant difference between test and retest; \* $p < 0.05$ ; \*\* $p < 0.01$ .

<sup>a</sup>Median (interquartile range) pain rating before, and during the measurement.

N: newton; W: watt; 95% CI: 95% confidence interval; ICC: intraclass correlation coefficient; SEM: standard error of the measurement; CV: coefficients of variation; MDC<sub>90</sub>: minimal detectable change at the 90% confidence level.

to the measurements, while some patients reported pain during the measurements of the most symptomatic LE (Table III).

*Functional performance measures.* Significantly, higher retest values were detected for all the functional performance measures except the UG and the 6MW test in the 2-week group and the TSC in the 2.5-week group (Table IV). The reliability parameter ranged from ICC = 0.82 to 0.96 and the agreement parameter from CV = 12% to 4% (Table IV).

*Healthy peers*

No significant differences were detected for maximal isometric knee extensor strength and LEP (Table V), but significantly higher retest values were detected for all functional performance measures except the timed stair climbing test (Table VI).

DISCUSSION

This intra-rater test-retest study in patients with hip OA in general showed good to excellent reliability for both hip muscle strength measured with HHD (ICC = 0.67–0.94) and thigh muscle strength measured with the Good Strength device (ICC = 0.73–0.94), while agreement ranged from poor to good (CV = 22–8%). However, the reproducibility of the muscle strength measurements, in general, was affected by the time interval between tests; from a general systematic decline for the 1-week interval to no change for the 2.5-week interval. In contrast, we showed excellent reliability (ICC = 0.93–0.96) and good agreement (CV = 11–8%) of LEP, and irrespective of the time interval between tests the results were unchanged. Finally, although both reliability (ICC = 0.82–0.96) and agree-

Table IV. Reliability and agreement of functional performance measures in patients with hip osteoarthritis

Weeks between test and retest	Performance measures					
	MOS (n)	UG (s)	TSC (s)	5CT (s)	30sCS (n)	6MW (m)
<b>1 (n = 37)</b>						
Test/retest, mean (SD)	21.0 (5.4)/ 22.7 (6.2)	5.74 (1.47)/ 5.52 (1.23)	10.04 (2.07)/ 9.73 (2.17)	10.14 (2.63)/ 9.34 (2.61)	13.7 (3.6)/ 14.9 (4.7)	511.81 (102.62)/ 529.03 (107.87)
Difference, mean (SD)	1.6 (2.8)**	-0.22 (0.61)*	-0.32 (0.76)*	-0.78 (1.28)**	1.1 (2.1)**	17.22 (29.27)**
[95% CI]	[0.7 to 2.6]	[-0.42 to -0.02]	[-0.57 to -0.07]	[-1.22 to -0.34]	[0.4 to 1.8]	[7.46 to 26.98]
ICC	0.888	0.900	0.936	0.880	0.877	0.961
SEM	2.0	0.43	0.54	0.91	1.5	20.69
CV (%)	10.3	7.3	5.6	10.7	11.6	4.6
MDC <sub>90</sub>	4.6	1.00	1.25	2.11	3.5	48.30
<b>2 (n = 35)</b>						
Test/retest, mean (SD)	20.0 (4.8)/ 21.8 (5.1)	5.98 (1.31)/ 5.95 (1.32)	10.56 (2.61)/ 10.11 (2.22)	11.37 (3.02)/ 10.49 (2.70)	12.6 (3.4)/ 13.5 (3.5)	520.86 (97.57)/ 526.86 (99.69) <sup>a</sup>
Difference, mean (SD)	1.8 (2.8)**	-0.04 (0.54)	-0.45 (0.86)**	-0.89 (1.40)**	0.9 (1.3)**	6.01 (27.85)
[95% CI]	[0.9 to 2.7]	[-0.22 to 0.15]	[-0.75 to -0.16]	[-1.39 to -0.38]	[0.5 to 1.4]	[-3.71 to 15.72]
ICC	0.846	0.917	0.937	0.867	0.925	0.960
SEM	2.0	0.38	0.61	0.99	0.9	19.7
CV (%)	11.0	6.3	6.6	11.0	8.7	3.9
MDC <sub>90</sub>	4.6	0.89	1.42	2.31	2.1	45.95
<b>2.5 (n = 15)</b>						
Test/retest, mean (SD)	20.9 (5.7)/ 22.9 (4.1)	5.32 (0.84)/ 5.10 (0.88)	8.92 (1.30)/ 8.70 (1.35)	9.99 (2.12)/ 9.38 (1.94)	14.2 (3.4)/ 15.1 (3.6)	537.15 (71.67)/ 555.34 (74.09)
Difference, mean (SD)	2.1 (3.0)*	-0.22 (0.32)*	-0.23 (0.62)	-0.61 (0.89)*	0.9 (1.5)*	18.2 (30.3)*
[95% CI]	[0.4 to 3.7]	[-0.40 to -0.04]	[-0.58 to 0.11]	[-1.10 to -0.12]	[0.1 to 1.8]	[1.41 to 34.98]
ICC	0.821	0.931	0.890	0.905	0.911	0.914
SEM	2.1	0.23	0.44	0.63	1.1	21.43
CV (%)	11.5	5.1	5.2	7.7	8.3	4.5
MDC <sub>90</sub>	5.0	0.53	1.02	1.47	2.5	50.00

Significant difference between test and retest; \**p*<0.05; \*\**p*<0.01. <sup>a</sup>*n*=34, 1 missing due to Achilles tendon overuse injury.

MOS: 15 s marching on the spot test; UG: 8-foot Up & Go test; TSC: timed stair climb test; 5CT: timed 5 chair stands test; 30sCS: 30 s chair stand test; 6MW: 6-min walk test; 95% CI: 95% confidence interval; ICC: intraclass correlation coefficient; SEM: standard error of the measurement; CV: coefficients of variation; MDC<sub>90</sub>: minimal detectable change at the 90% confidence level.

ment (CV = 12–4%) were moderate to good for the functional performance measures nearly all results improved regardless of the time interval between tests. Significant improvements in the functional performance tests at retest were also detected in the healthy participants, while results on knee extensor strength and LEP were unchanged.

The time-dependent systematic decrease in maximal isometric hip and thigh muscle strength in patients in the 1 and the 2 week groups cannot be explained by self-reported hip

pain. Self-reported hip pain was not an issue regarding assessments of muscle function in the least symptomatic LE. In general, self-reported pain prior to and during the muscle function measurements in the most symptomatic LE was very low, did not change during the test procedures, and were the same at test and retest. We cannot ascertain the mechanism for the decrease in strength, but fear of pain or pain-related fear could potentially play a role. It has been shown that anticipation of pain in clinical pain populations often results in poor

Table V. Reliability and agreement of knee extensor strength and leg extensor power measurements in the dominant and the non-dominant lower limb in healthy older adults with one week between tests

	Knee extensor strength (N)		Muscle power (Watt)	
	Domin	Non-domin	Domin	Non-domin
Test/retest, mean (SD)	392.6 (101.4)/396.7 (105.6)	375.2 (106.4)/371.0 (106.3)	154.0 (56.1)/157.6 (57.0)	152.1 (56.4)/150.5 (53.1)
Difference, mean (SD) [95% CI]	4.1 (47.4) [-12.2 to 20.4]	-4.3 (38.0) [-17.3 to 8.8]	3.6 (22.8) [-4.2 to 11.59]	-1.6 (18.3) [-7.9 to 4.69]
ICC	0.895	0.936	0.919	0.944
SEM	33.5	26.9	16.1	12.9
CV, %	8.4	7.1	10.3	8.4
MDC <sub>90</sub>	78.2	62.7	37.6	30.2

Significant difference between test and retest; \**p*<0.05; \*\**p*<0.01.

N: newton; 95% CI: 95% confidence interval; ICC: intraclass correlation coefficient; SEM: standard error of the measurement; CV: coefficients of variation; MDC<sub>90</sub>: minimal detectable change at the 90% confidence level.

Table VI. Reliability and agreement of functional performance measures in healthy older adults with 1 week between tests

	Performance measures					
	MOS (n)	UG (s)	TSC (s)	5CT (s)	30sCS (n)	6MW (m)
Test/retest, mean (SD)	26.1 (6.4)/ 28.4 (6.8)	5.28 (0.89)/ 5.11 (0.94)	8.99 (1.60)/ 8.91 (1.58)	9.06 (1.60)/ 8.50 (1.91)	15.5 (2.8)/ 16.3 (3.8)	582.72 (83.99)/ 599.03 (102.12)
Difference, mean (SD)	2.4 (3.0)**	-0.17 (0.40)*	-0.08 (0.67)	-0.55 (1.27)*	0.9 (1.9)*	16.30 (30.17)**
[95% CI]	1.3 to 3.4]	[-0.31 to -0.03]	[-0.32 to 0.15]	[-1.00 to -0.11]	[0.19 to 1.5]	[5.94 to 26.67]
ICC	0.896	0.902	0.909	0.737	0.828	0.948
SEM	2.1	0.28	0.49	0.91	1.3	21.33
CV (%)	10.1	5.9	5.3	11.4	9.3	4.1
MDC <sub>90</sub>	4.95	0.66	1.15	2.15	3.1	49.78

Significant difference between test and retest; \* $p < 0.05$ ; \*\* $p < 0.01$ .

MOS: 15 s marching on the spot test; UG: 8-foot Up & Go test; TSC: timed stair climb test; 5CT: timed 5 chair stands test; 30sCS: 30 s chair stand test; 6MW: 6-min walk test; 95% CI: 95% confidence interval; ICC: intraclass correlation coefficient; SEM: standard error of the measurement; CV: coefficients of variation; MDC<sub>90</sub>: minimal detectable change at the 90% confidence level.

behavioural performance, which cannot be accounted for by pain severity (31). However, the decline in strength could also be due to muscle weakness associated with delayed onset muscle soreness, which can persist for up to 21 days after 1 bout of eccentric exercise (13). Recovery of muscle strength after unaccustomed, vigorous, physical activity may be slower in older compared with younger persons (32, 33). Another possible mechanism involves motivational factors (9, 18), since the strength measures accounted for approximately 70% of the total test time.

Our results regarding the isometric hip strength measures in the 2.5-week group are generally consistent with previous results in intra-rater reproducibility of maximal isometric hip muscle strength measures in persons with hip OA (14, 18). However, Pua et al. (14) found a significant improvement in hip abductor strength. In our study, the hip abductor strength was numerically, but not significantly, greater at retest (Table II). Finally, the reproducibility of hip flexor strength was poorer in our study (ICC = 0.84 and CV = 19%) compared with those of Pua et al. (14) (ICC = 0.87 and CV = 11%; median 19 days between tests,  $n = 22$ ) and Arokoski et al. (18) (ICC = 0.98 and CV = 8%; 2–6 weeks between tests,  $n = 9$ ). These differences could be due to differences in testing procedures. We used a HHD-measurement and the seated test position similar to Pua et al., but we did not use a stabilization belt across the waist for the hip flexor strength measurement. In contrast, Arokoski et al. used the supine test position and an Active Isokinetic Rehabilitation System measurement.

We have not been able to identify any reproducibility study of thigh muscle strength in individuals with hip OA. However, a small intra-rater study (34) in 10 patients with hip or knee OA showed reliability parameters of isometric knee extensor strength (ICC = 0.95 and ICC = 0.97; test and retest within 1 week), which is comparable with our results (ICC = 0.84–0.92).

LEP was unchanged at retest regardless of the time interval between tests, and we cannot ascertain the mechanism for this finding. One possible explanation could be that unilateral LE extension involves several muscle groups in the LE, and thus hip extensors and calf muscles may have compensated for the muscle weakness in the other hip and thigh muscles. Furthermore, the measurement of LEP was dynamic in contrast to the

static strength measurement and may resemble a more familiar motor activity compared with the isometric strength measures.

In contrast to our findings a reproducibility study including LEP measurements in patients awaiting hip ( $n = 9$ ) or knee ( $n = 11$ ) replacement (16) showed a systematic improvement at retest 1 week apart. Reliability or agreement parameters were poorer (ICC = 0.72, SEM = 18.3 watt, CV = 21% and MDC<sub>90</sub> = 43 watt) compared with our study (ICC = 0.93–0.96, SEM = 13.4–10.7 watt, CV = 11–8% and MDC<sub>90</sub> = 31.2–24.9 watt). This might be because the patients in that study had end-stage OA and/or because the majority had knee OA, and reliability and agreement parameters are population-specific (11). In our study, reliability and agreement of LEP in patients with hip OA and healthy peers were comparable (Tables III and V). We had expected a greater within-subject variation in the patients because of a greater fluctuation in symptoms and functioning (35).

Several studies have documented learning effects for functional performance measures in healthy older people (36, 37), but although physical function is one of the recommended core outcomes in clinical trials of OA (38), hardly any studies have investigated reproducibility of functional performance measures in patients with symptomatic hip OA only (12). One study on intra-rater reproducibility of functional performance measures in 9 men with hip OA (19) documented a reliability parameter of the marching on the spot test (ICC = 0.85) comparable to our study (ICC = 0.82–0.89), but the agreement parameter was slightly poorer (CV = 15% vs CV = 12–10%), which could be due to the small sample size and the variable time (2–6 weeks) between tests.

The intra-rater reproducibility results of functional performance measures in patients with knee or hip OA have been mixed. Two test-retest studies in patients awaiting total hip or knee replacement demonstrated a significant improvement at retest in the 30-s Chair Stand Test (30–35 min between tests;  $n = 82$ ) (17) and timed 5 chair stand test (1 week between tests;  $n = 20$ ) (16). In contrast, no systematic differences in results were demonstrated in 6-min walk, Timed Up and Go (3 m), timed stair climbing, and a fast self-paced walk when the median interval between test and retest was 178 days ( $n = 150$ ) (30). In this case, a potential learning effect may have disappeared.

In our study, the results for nearly all functional performance measures improved significantly at retest in both patients and healthy and, in contrast to what we expected, the reliability and agreement parameters were comparable (Tables IV and VI). We measured 5CT and 30sCS because both tests have been used as outcome measures in clinical trials of OA (26). The reliability and agreement parameters of the 2 tests are comparable, but a floor effect has been reported for the 5CT (26) and, consequently, the 30sCS appears to be more suitable.

The present study has inherent strengths and limitations. The advantages are inclusion of patients with symptomatic hip OA only, the diversity of outcome measures investigated, and the importance of the timing of tests. The study has documented reliability and agreement for some of the recommended core functional performance measures from clinical practice guidelines to assess activity and participation limitations and effects of treatment programmes in patients with hip OA (12, 26, 39). While intra-rater studies usually limit the generalizability of the results, we believe that it is a strength of the present study because it eliminates inter-rater variation in the results related to the time interval between the tests. The limitations are that it cannot be excluded that the fairly heavy burden of measurements undertaken on both days may have contributed to the systematic decline in muscle strength and the small number of patients in the 2.5-week group may have increased the risk of type 2 errors. A few of the patients were, for practical reasons, tested at different times of the day, which may have added to the variance. Finally, because we only performed the tests twice we are unable to establish whether the test results become more stable at a third trial (9, 36).

In conclusion, to our knowledge, this is the first study that has investigated the intra-rater reliability and agreement of hip and thigh muscle strength, LEP and functional performance measurements in patients with symptomatic hip OA only. Our results indicate that the time interval between test and retest may affect the reproducibility of muscle strength measures and that at least 2 weeks between test and retest are required to avoid a systematic decline. Although the reproducibility of the functional performance measures was good, most test results improved regardless of the time interval between tests, indicating that at least 1 practice trial prior to data collection is needed. The only measurement that seems to be independent of the time interval between testing is leg extensor power, which is also strongly associated with functional performance in older individuals. For that reason, and because the test-retest agreement is within the acceptable range, it may be appropriate and sensitive to measure change over time in patients with hip OA.

#### ACKNOWLEDGEMENTS

This study was supported by independent research grants: Danish Foundation TrygFonden (1190-09), Nordea Foundation (Healthy Ageing grant), Health Foundation (2009B097), Danish Rheumatism Association (R56-Rp2380), Lundbeck Foundation (FP50/2009), School of Physical Therapy in Copenhagen, The Association of Danish Physiotherapists Research Fund.

#### REFERENCES

- Holla JF, van der Leeden M, Knol DL, Peter WF, Roorda LD, Lems WF, et al. Avoidance of activities in early symptomatic knee osteoarthritis: results from the CHECK cohort. *Ann Behav Med* 2012; 44: 33–42.
- Fautrel B, Hilliquin P, Rozenberg S, Allaert FA, Coste P, Leclerc A, et al. Impact of osteoarthritis: results of a nationwide survey of 10,000 patients consulting for OA. *Joint Bone Spine* 2005; 72: 235–240.
- Nuesch E, Dieppe P, Reichenbach S, Williams S, Iff S, Juni P. All cause and disease specific mortality in patients with knee or hip osteoarthritis: population based cohort study. *BMJ* 2011; 342: d1165.
- Slemenda C, Brandt KD, Heilman DK, Mazucca S, Braunstein EM, Katz BP, et al. Quadriceps weakness and osteoarthritis of the knee. *Ann Intern Med* 1997; 127: 97–104.
- Loureiro A, Mills PM, Barrett RS. Muscle weakness in hip osteoarthritis: a systematic review. *Arthritis Care Res (Hoboken)* 2013; 65: 340–352.
- Zhang W, Doherty M, Arden N, Bannwarth B, Bijlsma J, Gunther KP, et al. EULAR evidence based recommendations for the management of hip osteoarthritis: report of a task force of the EULAR Standing Committee for International Clinical Studies Including Therapeutics (ESCIIT). *Ann Rheum Dis* 2005; 64: 669–681.
- Fransen M, McConnell S, Hernandez-Molina G, Reichenbach S. Exercise for osteoarthritis of the hip. *Cochrane Database Syst Rev* 2009; (3): CD007912.
- de Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. *J Clin Epidemiol* 2006; 59: 1033–1039.
- Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med* 1998; 26: 217–238.
- Hopkins WG. Measures of reliability in sports medicine and science. *Sports Med* 2000; 30: 1–15.
- Kottner J, Gajewski BJ, Streiner DL. Guidelines for Reporting Reliability and Agreement Studies (GRRAS). *Int J Nurs Stud* 2011; 48: 659–660.
- Dobson F, Hinman RS, Hall M, Terwee CB, Roos EM, Bennell KL. Measurement properties of performance-based measures to assess physical function in hip and knee osteoarthritis: a systematic review. *Osteoarthritis Cartilage* 2012; 20: 1548–1562.
- Pearce AJ, Sacco P, Byrnes ML, Thickbroom GW, Mastaglia FL. The effects of eccentric exercise on neuromuscular function of the biceps brachii. *J Sci Med Sport* 1998; 1: 236–244.
- Pua YH, Wrigley TV, Cowan SM, Bennell KL. Intrarater test-retest reliability of hip range of motion and hip muscle strength measurements in persons with hip osteoarthritis. *Arch Phys Med Rehabil* 2008; 89: 1146–1154.
- Bean JF, Leveille SG, Kiely DK, Bandinelli S, Guralnik JM, Ferrucci L. A comparison of leg power and leg strength within the InCHIANTI study: which influences mobility more? *J Gerontol A Biol Sci Med Sci* 2003; 58: 728–733.
- Villadsen A, Roos EM, Overgaard S, Holsgaard-Larsen A. Agreement and reliability of functional performance and muscle power in patients with advanced osteoarthritis of the hip or knee. *Am J Phys Med Rehabil* 2012; 91: 401–410.
- Gill S, McBurney H. Reliability of performance-based measures in people awaiting joint replacement surgery of the hip or knee. *Physiother Res Int* 2008; 13: 141–152.
- Arokoski MH, Arokoski JP, Haara M, Kankaanpaa M, Vesterinen M, Niemitukia LH et al. Hip muscle strength and muscle cross sectional area in men with and without hip osteoarthritis. *J Rheumatol* 2002; 29: 2185–2195.
- Arokoski MH, Haara M, Helminen HJ, Arokoski JP. Physical function in men with and without hip osteoarthritis. *Arch Phys Med Rehabil* 2004; 85: 574–581.
- Altman R, Alarcon G, Appelrouth D, Bloch D, Borenstein D,



- Brandt K, et al. The American College of Rheumatology criteria for the classification and reporting of osteoarthritis of the hip. *Arthritis Rheum* 1991; 34: 505–514.
21. Downie WW, Leatham PA, Rhind VM, Wright V, Branco JA, Anderson JA. Studies with pain rating scales. *Ann Rheum Dis* 1978; 37: 378–381.
  22. Bohannon RW. Make tests and break tests of elbow flexor muscle strength. *Phys Ther* 1988; 68: 193–194.
  23. Tiainen K, Sipilä S, Alen M, Heikkinen E, Kaprio J, Koskenvuo M, et al. Heritability of maximal isometric muscle strength in older female twins. *J Appl Physiol* 2004; 96: 173–180.
  24. Basseij EJ, Short AH. A new method for measuring power output in a single leg extension: feasibility, reliability and validity. *Eur J Appl Physiol Occup Physiol* 1990; 60: 385–390.
  25. Rikli RE, Jones CJ. Development and validation of a functional fitness test for community-residing older adults. *J Aging Phys Activity* 1999; 7: 129–161.
  26. Bennell K, Dobson F, Hinman R. Measures of physical performance assessments: Self-Paced Walk Test (SPWT), Stair Climb Test (SCT), Six-Minute Walk Test (6MWT), Chair Stand Test (CST), Timed Up & Go (TUG), Sock Test, Lift and Carry Test (LCT), and Car Task. *Arthritis Care Res (Hoboken)* 2011; 63 Suppl 11: S350–S370.
  27. Jones CJ, Rikli RE, Beam WC. A 30-s chair-stand test as a measure of lower body strength in community-residing older adults. *Res Q Exerc Sport* 1999; 70: 113–119.
  28. ATS statement: guidelines for the six-minute walk test. *Am J Respir Crit Care Med* 2002; 166: 111–117.
  29. Vasikaran SD, McCloskey EV, Kahn S, Kanis JA. Intra-individual variation in fasting urinary calcium- and hydroxyproline-creatinine ratios measured in metabolic bone clinic patients as both outpatients and inpatients. *Ann Clin Biochem* 1994; 31: 272–276.
  30. Kennedy DM, Stratford PW, Wessel J, Gollish JD, Penney D. Assessing stability and change of four performance measures: a longitudinal study evaluating outcome following total hip and knee arthroplasty. *BMC Musculoskelet Disord* 2005; 6: 3.
  31. Crombez G, Vlaeyen JW, Heuts PH, Lysens R. Pain-related fear is more disabling than pain itself: evidence on the role of pain-related fear in chronic back pain disability. *Pain* 1999; 80: 329–339.
  32. Chapman DW, Newton M, McGuigan MR, Nosaka K. Comparison between old and young men for responses to fast velocity maximal lengthening contractions of the elbow flexors. *Eur J Appl Physiol* 2008; 104: 531–539.
  33. Dedrick ME, Clarkson PM. The effects of eccentric exercise on motor performance in young and older women. *Eur J Appl Physiol Occup Physiol* 1990; 60: 183–186.
  34. Lin YC, Davey RC, Cochrane T. Tests for physical function of the elderly with knee and hip osteoarthritis. *Scand J Med Sci Sports* 2001; 11: 280–286.
  35. Allen KD, Coffman CJ, Golightly YM, Stechuchak KM, Keefe FJ. Daily pain variations among patients with hand, hip, and knee osteoarthritis. *Osteoarthritis Cartilage* 2009; 17: 1275–1282.
  36. Kervio G, Carre F, Ville NS. Reliability and intensity of the six-minute walk test in healthy elderly subjects. *Med Sci Sports Exerc* 2003; 35: 169–174.
  37. Tager IB, Swanson A, Satariano WA. Reliability of physical performance and self-reported functional measures in an older population. *J Gerontol A Biol Sci Med Sci* 1998; 53: M295–M300.
  38. Pham T, van der Heijde D, Altman RD, Anderson JJ, Bellamy N, Hochberg M, et al. OMERACT-OARSI initiative: Osteoarthritis Research Society International set of responder criteria for osteoarthritis clinical trials revisited. *Osteoarthritis Cartilage* 2004; 12: 389–399.
  39. Dobson F, Hinman RS, Roos EM, Abbott JH, Stratford P, Davis AM, et al. OARSI recommended performance-based tests to assess physical function in people diagnosed with hip or knee osteoarthritis. *Osteoarthritis Cartilage* 2013; 21: 1042–1052.