Osteoarthritis and Cartilage 25 (2017) 1792–1796

Osteoarthritis and Cartilage

Brief Report

Reliability and measurement error of the Osteoarthritis Research Society International (OARSI) recommended performance-based tests of physical function in people with hip and knee osteoarthritis

F. Dobson † *, R.S. Hinman ‡, M. Hall ‡, C.J. Marshall ‡, T. Sayer ‡, C. Anderson †, N. Newcomb §, P.W. Stratford ||, K.L. Bennell ‡

† Department of Physiotherapy, School of Health Sciences, University of Melbourne, Victoria, Australia
‡ Centre for Health, Exercise and Sports Medicine, Department of Physiotherapy, School of Health Sciences, University of Melbourne, Victoria, Australia
§ Institute of Evolutionary Medicine, University of Zurich, Zurich, Switzerland
|| School of Rehabilitation Science, Faculty of Health Sciences, McMaster University, Hamilton, Ontario, Canada

ARTICLE INFO

Article history:
Received 7 March 2017
Accepted 14 June 2017

Keywords:
Osteoarthritis
Performance-based tests
Physical function
Outcomes
Reliability
Measurement error

SUMMARY

Objective: To estimate the reliability and measurement error of performance-based tests of physical function recommended by the Osteoarthritis Research Society International (OARSI) in people with hip and/or knee osteoarthritis (OA).

Design: Prospective repeated measures between independent raters within a session and within-rater over a week interval. Relative reliability was estimated for 51 people with hip and/or knee OA (mean age 64.5 years, standard deviation (SD) 6.21 years; 47% females; 36 (70%) primary knee OA) on the 30s Chair Stand Test (30sCST), 40m Fast-Paced Walk Test (40mFPWT), 11-Stair Climb Test (11-step SCT), Timed Up and Go (TUG), Six-Minute Walk Test (6MWT), 10m Fast-Paced Walk Test (10mFPWT) and 20s Stair Climb Test (20sSCT) using intra-class correlation coefficients (ICC). Absolute reliability was calculated using standard error of measurement (SEM) and minimal detectable change (MDC).

Results: Measurement error was acceptable (SEM < 10%) for all tests. Between-rater reliability was: optimal (ICC > 0.9, lower 1-sided 95% CI > 0.7) for the 40mFPWT, 6MWT and 10mFPWT; sufficient (ICC >0.8, lower 1-sided 95% CI > 0.7) for 30sCST, 20sSCT; unacceptable (lower 1-side 95% CI < 0.7) for 11-step SCT and TUG. Within-rater reliability was optimal for 40mFPWT, and 6MWT; sufficient for 30sCST and 10mFPWT and unacceptable for 11-step SCT, TUG and 20sSCT.

Conclusions: The 30sCST, 40mFPWT, 6MWT and 10mFPWT, demonstrated, at minimum, acceptable levels of both between and within-rater reliability and measurement error. All tests demonstrated sufficiently small measurement error indicating they are adequate for measuring change over time in individuals with knee/hip OA.

Introduction

Feasible, reliable and valid measurement of treatment outcomes is critical to both research and clinical practice[1]. The Osteoarthritis Research Society International (OARSI) recommends a set of performance-based measures of physical function representing typical activities relevant to individuals diagnosed with hip or knee OA[2,3]. These tests are the 30-s Chair Stand Test (30sCST), 40m Fast-paced Walk Test (40mFPWT), a Stair Climb Test (SCT), Timed up and Go Test (TUG) and six-minute Walk Test (6MWT). In consideration of some limitations related to the complexity of the 40mFPWT and scoring of the 11-step SCT, a modified short distance walking test, 10m Fast-paced Walk Test (10mFPWT) and a time-based SCT, 20-second Stair Climb Test (20sSCT) were suggested as alternate tests to overcome limitations associated with existing tests[2].

* Address correspondence and reprint requests to: F. Dobson, Department of Physiotherapy, The University of Melbourne, Victoria, 3010, Australia.
E-mail addresses: fdobson@unimelb.edu.au (F. Dobson), ranash@unimelb.edu.au (R.S. Hinman), halm@unimelb.edu.au (M. Hall), c.marshall@unimelb.edu.au (C.J. Marshall), tsayer@student.unimelb.edu.au (T. Sayer), mranderson808@gmail.com (C. Anderson), nranewcomb@gmail.com (N. Newcomb), stratfor@mcmaster.ca (P.W. Stratford), k.bennell@unimelb.edu.au (K.L. Bennell).

http://dx.doi.org/10.1016/j.joca.2017.06.006
1063-4584/

For a measure to be useful it must demonstrate sufficient relative and absolute reliability[4]. Relative reliability refers to the degree to which a measurement is free from error and remains consistent when measured repeatedly[5]. Absolute reliability refers to the systematic and random error of a measurement that is not attributable to true changes expressed as measurement error[5]. Although reliability estimates for different populations are available for some OARSI recommended tests[6,7], comprehensive OA-specific reliability estimates for all five tests is limited. This study estimates the relative reliability and absolute measurement error of the OARSI recommended tests together with two other clinically-relevant complementary performance-based tests in people with hip and/or knee OA.

Methods

This prospective reliability study received University human research ethics approval (HECS 123908). All participants provided written informed consent. In this study, "between-rater reliability" refers to repeated measures between independent raters within one session, whilst "within-rater reliability" refers to repeated measures by one rater over a week interval. Both designs incorporate test-retest reliability.

Participants

Participants were sourced from a research database of community volunteers. To be eligible, participants were required to: (1) be age > 40 years; (2) have hip or knee pain on most of days of the past month; (3) be able to ambulate independently in the community; and (4) be able to read and follow instructions in English. Participants were also required to fulfil clinical diagnostic criteria for knee or hip OA, established by the American College of Rheumatology[8,9]. Participants were not eligible if they had: (1) previous hip or knee joint replacement; (2) hip or knee surgery in the past 6 months; and/or (3) neurological disorders which may interfere with walking and balance.

Procedures

Participants were tested at two separate testing sessions approximately 1 week apart. In session 1, participants performed seven performance-based tests with two independent raters to examine the reliability between raters within a test session. One rater (Rater A) was a research physiotherapist with 12-years clinical experience who tested all participants. To increase the generalisability of our findings, Rater B was formed from a pool of three raters, two final-year graduate physiotherapy students and one physician with 1-year clinical experience, each who tested a sub-set of >15 participants each. All raters received a half-day training session on the study protocol and how to administer the tests. The initial testing order of the tests and rater were both randomised, and then repeated in the same order with the alternate rater with 5 min rest between each rater's assessments. Baseline demographics, an 11-point Pain Numeric Rating Scale (NRS) to assess the average level of knee/hip pain experienced over the past week and, as appropriate, the Knee injury and Osteoarthritis Outcome Score[10], or the Hip disability and Osteoarthritis Outcome Score to assess patient-reported symptoms and disability related to knee or hip OA[11] were collected.

In session 2, participants repeated the performance tests in the same order as the first test occasion with Rater A (blinded to previous results). Participants completed a self-reported global change scale (GCS) and only those reporting no or slight change (from a five-point scale: much worse, slightly worse, no change, slightly better and much better) were retested and analysed to ensure their condition was stable over the week. A 1-week interval between test sessions limited rater recall of test scores and the potential for real change in clinical status.

A description of each of the OARSI recommended performance-based test (30sCST, 40mFPWT, SCT, TUG and 6MWT) including set up, equipment, preparation (environment, participant, and tester), procedures, verbal instructions and scoring are available on the OARSI website: http://oarsi.org/research/physical-performance-measures. This resource was used to familiarise Raters on the test administration and the lead author (FD) instructed Raters on the administration of the 10mFPWT and 20sSCT. An 11-step SCT was conducted for this study. The 10mFPWT was performed using a straight 10 m pathway where participants were timed walking as quickly as possible, without running. The 20sSCT, was performed by ascending and descending an 11-step flight of stairs as quickly and safely as possible in 20 s. A supervised practice trial was conducted to check safety and understanding for both tests.

Data analysis

SPSS (version 22, SPSS Chicago, IL USA) was used to perform statistical analyses. Data were checked for normality. Descriptive analyses were performed and the frequencies of minimum and maximum scores were reviewed to investigate ceiling or floor effects.

Relative within-rater reliability was calculated using intra-class correlation coefficients ($ICC_{2,1}$) with 95% confidence intervals (CI) for a two-way random effects model and absolute agreement. Relative between-rater reliability was calculated using $ICC_{1,1}$ with 95% CI for a one-way random effects model as only one rater (Rater A) assessed all participants in the first session, whilst Rater B was from a pool of three raters, who each assessed a sub-set of participants in session 1. Interpretation of ICC values was based on single measures and inspection of the lower one-sided 95% CI set at a minimum acceptable level of 0.70. Pre-set point estimates ICC values of 0.8 or more indicated sufficient reliability and values 0.90 or more indicated optimal reliability[1,4]. Absolute reliability (measurement error) was expressed as the standard error of measurement (SEM), calculated as the square root of the mean square error term from the ANOVA. To quantify the inherent variability in 90% of unchanged participants, the MDC_{90} was estimated, calculated as SEM × 1.65 (z score of 90 % interval) × $\sqrt{2}$. As such, changes greater than the minimal detectable change (MDC) were interpreted as true change. A 95% CI for the SEM was calculated according to recommended methods[12]. As the units of measurement varied, SEM percentage (SEM%) and MDC_{90} percentage (MDC_{90}%) were calculated as to SEM% = (SEM/mean) × 100 and MDC% = (MDC_{90}/mean) × 100. A SEM% ≤10% was accepted as an acceptably small measurement error[13].

Sample size

Sample size calculations were based on *a priori* set levels of sufficient and minimal acceptable limits of reliability for clinical measurement. A minimum of 50 participants were required to estimate a sufficient ICC of 0.80 and a minimal acceptable lower one-sided 95% confidence limit of 0.70[5].

Results

Participant characteristics are presented in Table I. Interpretation was based on normally distributed data. No ceiling or floor effects were detected for any physical function test at any test occasion. Eight participants were excluded (n = 3 unwell and

Table I
Participant characteristics presented as mean (SD), min–max, unless otherwise stated

Characteristic	$N = 51$
Age, years	64.5 (6.2), 51–81
Female gender, *n* (%)	24 (47)
Height, m	1.72 (0.09), 1.55–1.89
Body mass, kg	84.47 (17.74), 48.0–119.0
BMI, kg/m^2	28.54 (4.84), 19.83–38.51
Duration of symptoms, years	6.9 (5.8), 1–36
*Primary hip OA involvement, *n* (%)	15 (30)
*Primary knee OA involvement, *n* (%)	36 (70)
†Unilateral OA, (%)	19 (37)
†Bilateral OA, *n* (%)	32 (63)
‡NRS pain, (0–10)	5.3 (1.8), 1-9
§KOOS score, (0–100) [$n = 36$]	
Pain	58.6 (11.6), 36.1–86.1
Symptoms	55.0 (15.4), 14.3–92.9
ADL	67.6 (13.9), 23.5–92.6
Sport	33.3 (18.2), 0.0–80.0
QOL	42.7 (12.2), 18.8–68.8
§HOOS score, (0–100) [$n = 15$]	
Pain	57.2 (11.7), 35.0–80.0
Symptoms	59.3 (14.1), 35.0–80.0
ADL	61.1 (11.7), 39.7–77.9
Sport	47.9 (17.3), 6.3–75.0
QOL	49.2 (13.3), 25.0–68.8

BMI = body mass index, HOOS = hip dysfunction and osteoarthritis outcome score, KOOS = knee dysfunction and osteoarthritis outcome score, QOL = Quality of Life, ADL = Activities of daily living.
* Primary refers to most symptomatic joint in case of OA in more than one joint type.
† Refers to hip and knee OA.
‡ 0 = no pain, 10 = worst possible pain.
§ 0 = extreme problems, 100 = no problems.

unable for session 2; $n = 2$ moderately better; $n = 2$ moderately worse; $n = 1$ much worse). Data for between-rater reliability within a single session and within-rater reliability of repeated measures over a 1-week interval are presented in Table II. Results remained largely unchanged when including all 59 participants in the between-rater analyses, suggesting that the excluded participants did not influence between-rater reliability appreciably.

Between-rater reliability within a single session

In terms of relative between-rater reliability, the 40mFPWT, 6MWT, and 10mFPWT achieved optimal levels (ICC >0.90, lower 1-sided 95% CI > 0.70); the 30sCST and 20sSCT achieved sufficient levels (ICC 0.82–0.86, lower 1-sided 95% CI: 0.73–0.78); whilst the TUG and 11-step SCT did not meet minimal acceptable levels (ICC 0.78, lower 1-sided 95% CI: 0.66–0.67). In terms of absolute between-rater reliability, SEM ranged between 3.6 and 9.3% of the mean test score, demonstrating sufficiently small measurement error in all tests (≤10%).

Within-rater reliability of repeated measures over a 1-week interval

The within-rater reliability test interval was 7 days for 46 (90%) participants, 8 days for 4 (8%) participants and 9 days for one (2%) participant. In terms of relative within-rater reliability, the 40mFPWT and 6MWT achieved optimal levels (ICC >0.90, lower 1-sided 95% CI > 0.70); the 30sCST and 10mFPWT achieved sufficient levels (ICC 0.85–0.88, lower 1-sided 95% CI: 0.70–0.82), whilst the 11-step SCT, TUG, and 20sSCT did not meet minimal acceptable levels (ICC 0.78–0.85, lower 1-sided 95% CI: 0.56–0.68). The SEM of all tests measured over a week interval ranged between 3.3 and 8.1%, demonstrating sufficiently small measurement error. The MDC_{90} estimates ranged between 7.6 and 18.8% of the mean test score.

Discussion

This study estimated the relative reliability and absolute measurement error associated with tests of physical function including those recommended by OARSI for people with knee and/hip OA[2,3]. The sufficiently small measurement error associated with all tests across all occasions, indicate that the OARSI set are appropriate for measuring change in physical function performance in individuals with knee and/or hip OA. In terms of relative reliability, it appeared that the 40mFPWT, 6MWT and 30sCST were more consistent than the 11-step SCT and TUG. The 10mFPWT appears to be an adequate alternate to the OARSI recommended 40mFPWT. Similarly, the 20sSCT appears to be slightly more reliable than the 11-step SCT. Clinicians and researchers can be guided by the MDC estimates, where changes greater than two stands for the 30sCST, 0.19 m/s for the 40mFPWT, 0.2 s for the TUG, 50.23 m for the 6MWT and 0.28 m for the 10mFPWT represent true change 90% beyond measurement error.

Due to limited reliability and measurement error estimates for the recommended performance-based tests of physical function in people with hip and/or knee OA, as well as differences in the terminology of the types of reliability design reported in the literature, direct comparison of our findings is challenging. Nevertheless, a number of consistencies with previously reported estimates of relative reliability in people with end-stage hip and knee OA[14,15] were evident.

Although all tests had sufficiently small measurement error, the 11-step SCT and TUG did not reach our pre-set minimal benchmarks of acceptable reliability. It should be acknowledged that reliability is but one aspect to consider in choosing a suitable test. Tests such as a stair-climb test are relevant to everyday activities that are important to people with OA and hence have face validity. Our findings suggest the 20sSCT may be a better alternate to the 11-step SCT and potentially more feasible as the number of stairs is constrained by the available testing environment.

MDC calculated for the tests can help clinicians and researchers interpret these tests by indicating how much change is real change. However, these values are estimated differently to minimal important change values that instead represent the amount of change that is clinically meaningful to the patient. Further research is required to determine estimates of minimal clinically important change associated with each of the tests.

There are several strengths of the current study. We evaluated OARSI recommended tests that were chosen based on a rigorous consensus process[2]. We provided *a priori* levels of adequate, sufficient and optimal levels of relative reliability upon which to adequately power the study and also to base our interpretations. This allows not only transparency but the ability to compare our data to others. We excluded participants who demonstrated symptomatic change after 1 week from the within-rater analysis. Four raters with varying levels of clinical and research experience were used in the between-rater reliability analyses.

Limitations of the study include that only one rater, albeit an individual with clinical and research experience, was used for our within-rater reliability analysis, limiting the generalisability of these findings. Further, it is possible that reliability could differ between participants with isolated unilateral/bilateral hip or knee OA. However joint specific and laterality sub-group analysis was not conducted due to inadequate statistical power. Lastly, the reliability of Rater B (three raters) is unknown and is a limitation of this study.

In summary, the OARSI recommended set of performance-based tests of physical function demonstrated sufficiently small measurement error indicating they are adequate for measuring change over time in individuals with knee/hip OA. The 10mFPWT appears to be a suitable alternate to the 40mFPWT.

Table II
Between and within-rater reliability and measurement error estimates

Between-rater reliability and measurement error estimates	Functional test	Rater A† Mean (SD)	Rater B Mean (SD)	$ICC_{(1,1)}$ (95% CI)	Lower 1-sided CI	SEM 95% CI	SEM%	
OARSI minimum core set	30sCST (no. of stands)	11.5 (2.7)	11.6 (2.7)	0.86 (0.77–0.92)	0.78	1.0 (0.8–1.3)	8.8	
	40mFPWT (m/s)	1.69 (0.26)	1.69 (0.30)	0.96 (0.93–0.98)	0.93	0.06 (0.05–0.08)	3.6	
	‡11-step SCT (s) [$n = 50$]	12.86 (2.42)	12.81 (2.63)	0.78 (0.65–0.87)	0.67	1.19 (0.99–1.48)	9.3	
	*11-step SCT (s)	13.27 (3.66)	13.18 (3.74)	0.90 (0.83–0.94)	0.85	1.18 (0.99–1.46)	8.9	
OARSI recommended set	§,†TUG (s) [$n = 49$]	8.4 (1.3)	8.3 (1.5)	0.78 (0.63–0.87)	0.66	0.7 (0.6–0.8)	7.9	
	*TUG (s)	8.6 (1.4)	8.3 (1.5)	0.75 (0.60–0.85)	0.63	0.7 (0.6–0.9)	8.5	
	6-min Walk Test (m)	545.1 (83.26)	551.29 (90.01)	0.94 (0.90–0.96)	0.90	21.28 (17.81–26.46)	3.9	
Expert suggestions	10mFPWT (m/s)	1.80 (0.32)	1.74 (0.30)	0.91 (0.83–0.95)	0.85	0.08 (0.07–0.10)	4.5	
	20sSCT (steps)	34.5 (7.4)	35.1 (7.5)	0.82 (0.71–0.89)	0.73	3.2 (2.7–3.9)	9.1	
Within-rater reliability and measurement error estimates	**Functional test**	**Session 1 Mean (SD)**	**Session 2 Mean (SD)**	**$ICC_{(2,1)}$ (95% CI)**	**Lower 1-sided CI**	**SEM 95% CI**	**SEM%**	**MDC_{90} (%)**
OARSI minimum core set	30sCST (no. of stands)	11.5 (2.7)	12.2 (2.4)	0.85 (0.67–0.93)	0.70	0.9 (0.7–1.1)	7.3	2.0 (16.9)
	40mFPWT (m/s)	1.69 (0.26)	1.74 (0.29)	0.92 (0.82–0.96)	0.84	0.07 (0.06–0.09)	4.1	0.19 (9.5)
	‡11-step SCT (s) [$n = 50$]	12.89 (2.42)	11.99 (2.41)	0.78 (0.50–0.89)	0.56	1.00 (0.83–1.24)	8.1	2.33 (18.8)
	*11-step SCT (s)	13.27 (3.66)	12.35 (3.53)	0.90 (0.72–0.95)	0.75	0.99 (0.83–1.23)	7.7	
OARSI recommended set	§,†TUG (s) [$n = 49$]	8.4 (1.3)	8.1 (1.3)	0.81 (0.65–0.89)	0.68	0.5 (0.4–0.7)	6.3	1.21 (14.7)
	*TUG (s)	8.6 (1.4)	8.1 (1.3)	0.75 (0.54–0.86)	0.58	0.6 (0.5–0.8)	7.5	
	6-min Walk Test (m)	545.12 (83.26)	563.85 (84.23)	0.93 (0.77–0.97)	0.79	18.12 (15.16–22.53)	3.3	50.23 (7.6)
Expert suggestions	10mFPWT (m/s)	1.80 (0.32)	1.83 (0.30)	0.88 (0.80–0.93)	0.82	0.10 (0.09–0.13)	5.5	0.28 (12.9)
	20sSCT (steps)	34.5 (7.4)	36.9 (7.6)	0.85 (0.61–0.93)	0.66	2.5 (2.1–3.1)	7.0	7 (16.3)

MDC_{90} estimates were not calculated between raters as the distribution of the difference scores did not conform to normal distribution.
* Refers to non-normally distributed data.
† Rater A assessed all 51 participants, Rater B is from a pool of three raters assessing >15 participants each; No statistical differences in the mean differences between Rater A and any Rater B were observed for any of the tests evaluated.
‡ One participant removed as considered an outlier.
§ Two participants removed as considered outliers.

Author contributions

FD, RSH and KLB conceived and designed this study. CJM, TS, CA, NN acquired the data. FD, MH and PWS preformed data analyses. FD, RSH, KLB, MH, PWS interpreted the data. FD and MH drafted the manuscript. All authors revised the manuscript for intellectual content and approved the final version.

Competing interests

The authors declare that they have no competing interests.

Role of the funding source

This project was partly funded by OARSI and Australian National Health and Medical Research Council (NHMRC) program grant (#631717) and forms part of an OARSI initiative to develop and recommend a set of physical performance measures for hip and knee OA. KLB is supported by a NHMRC Principal Research Fellowship (APP1058440). RSH is supported by a Future Fellowship from the Australian Research Council (FT130100175). MH is supported by a Sir Randal Heymanson Research Fellowship from The University of Melbourne. TS was supported by an NHMRC Australian Government Research Training Program Scholarship (APP1075881).

References

1. Portney LG, Watkins MP. Foundations of Clinical Research: Applications to Practice. 3rd edn. Upper Saddle River: N.J. Pearson/Prentice Hall; 2009.
2. 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;21:1042–52.
3. McAlindon TE, Driban JB, Henrotin Y, Hunter DJ, Jiang GL, Skou ST, *et al*. OARSI Clinical Trials recommendations: design, conduct, and reporting of clinical trials for knee osteoarthritis. Osteoarthritis Cartilage 2015;23:747–60.
4. de Vet HCW, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol 2006;59:1033–9.
5. De Vet HC, Terwee CB, Mokkink LB, Knol DL. Measurement in Medicine: A Practical Guide. Cambridge University Press; 2011.
6. 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 2011;63(Suppl 11): S350–70.
7. 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;20:1548–62.
8. 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;34:505–14.
9. Altman R, Asch E, Bloch D, Bole G, Borenstein D, Brandt K, *et al*. Development of criteria for the classification and reporting of osteoarthritis. Classification of osteoarthritis of the knee. diagnostic and therapeutic criteria committee of the American Rheumatism Association. Arthritis Rheum 1986;29:1039–49.
10. Roos EM, Roos HP, Ekdahl C, Lohmander LS. Knee injury and Osteoarthritis Outcome Score (KOOS)–validation of a Swedish version. Scand J Med Sci Sports 1998;8:439–48.

11. Thorborg K, Roos EM, Bartels EM, Petersen J, Holmich P. Validity, reliability and responsiveness of patient-reported outcome questionnaires when assessing hip and groin disability: a systematic review. Br J Sports Med 2010;44:1186–96.
12. Stratford PW, Goldsmith CH. Use of the standard error as a reliability index of interest: an applied example using elbow flexor strength data. Phys Ther 1997;77:745–50.
13. Goldberg A, Casby A, Wasielewski M. Minimum detectable change for single-leg-stance-time in older adults. Gait Posture 2011;33:737–9.
14. 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.
15. Gill S, McBurney H. Reliability of performance-based measures in people awaiting joint replacement surgery of the hip or knee. Physiother Res Int 2008;13:141–52.