Clinical Rehabilitation 2005; **19:** 659–661

The reliability of hip scoring systems for total hip arthroplasty candidates: assessment by physical therapists

Linda Kirmit School of Physiotherapy, **Vasfi Karatosun** Department of Orthopaedics and Traumatology, School of Medicine, **Bayram Unver, Serkan Bakirhan, Ayse Sen** and **Zeliha Gocen** School of Physiotherapy, Dokuz Eylül University, Izmir, Turkey

Received 14th September 2003; returned for revisions 25th January 2004; revised manuscript accepted 27th November 2004.

Objective: Hip rating systems have been widely used in assessing severity of hip dysfunction but no uniform method has emerged. The current study was performed to determine the interobserver reliability of five different hip scores on patients with coxarthrosis.
Design: Test reliability among physical therapists for five commonly used hip scores.
Subjects: Thirty-five patients (48 hips) who had coxarthrosis and who were candidates for total hip arthroplasty were included in the study.
Methods: Patients were evaluated preoperatively by three physical therapists using five different hip rating systems; the Harris Hip Score, the Iowa Hip Score, the Charnley Hip Score, the Merle d'Aubigne Hip Score and the American Academy of Orthopaedic Surgeons' Hip Score.
Results: The average age of the patients was 58.8 ± 2.2 years (range 28–76 years). For all scores, an excellent interobserver reliability between the physical therapists were found ($\kappa = 0.77$–0.95). The best correlation between first and second observer was on Harris Hip Score ($\kappa = 0.91$), between second and third was on Merle d'Aubigne Hip Score ($\kappa = 0.95$) and between first and third was on Iowa Hip Score ($\kappa = 0.87$).
Conclusion: There was an excellent interobserver reliability for all hip scores between the physical therapists, suggesting that all these hip scores are suitable for use by physical therapists.

Introduction

The role of physical therapists in the treatment of patients with a diagnosis of osteoarthritis of the hip is to assist in achieving the highest functional outcome possible, irrespective of whether they will eventually have a total hip arthroplasty.[1] Physical therapists, therefore, need measurement tools that accurately assess function and monitor change over time.[2] Many different hip rating scales have been constructed to quantify patients' complaints in evaluating severity of hip dysfunction,[3] but no uniform method has emerged.[4] The outcome results are inconsistent, often giving contrary measures of success in the same patient.[4] Inter-rater reliability is important in clinical settings, because several clinicians may assess the same patient at different times.[5] Although commonly

Address for correspondence: Vasfi Karatosun, Erzene Mah 116/16 Sok No: 8/12 Bornova 35050 Izmir, Turkey.
e-mail: vasfi.karatosun@deu.edu.tr

10.1191/0269215505cr869oa

used by physical therapists, only one report is found in the literature[6] evaluating the interobserver reliability of one hip score recorded by physical therapists. The present study was conducted to evaluate the reliability of five commonly used hip rating scales.

Method

A series of 48 hips of 35 patients (18 women (26 hips) and 17 men (22 hips)), who had coxarthrosis were included in the study. The average age of the patients was 58.8 (range 28–76 years). The indication for surgery was primary coxarthrosis in 42 hips, and secondary in six hips. Patients who had neurological or medical conditions causing locomotor disability and patients who had chronic disease such as obstructive lung disease or coronary artery disease were excluded from the study.

All patients were evaluated preoperatively by three observers (SB, AS, ZG) with a range of 3–5 years' experience, using five commonly used hip rating systems[7,8]: Charnley, Merle d'Aubigne, Harris, Iowa and the American Academy of Orthopaedic Surgeons (AAOS) hip scores. The observers completed all five scores for each patient independently in random order: Observer I was the first observer to examine the first patient, the second to examine the second patient, the third observer to examine the third patient, and the first observer to examine the fourth patient. The observers were given as much time as they needed to complete all five scores. Each observer completed the scores without knowledge of the ratings of the other observers.

University hospital local ethic committee approved the study and informed consent was obtained as required.

To simplify the statistical work only two grades of result were used. Excellent and good results of the rating methods were closed as 'good' and fair, poor and failures were classed as 'poor'.

For the interobserver correlations, kappa values were calculated. Kappa values are generated by setting the observed agreement in relation to the proportion of agreement expected by chance. Values between 0.40 and 0.75 were evaluated as good correlation and values more than 0.75 as excellent correlation.

Results

The correlations of five rating scales between the physical therapists are displayed in Table 1. Although all kappa values were more than 0.75 and rated as excellent, maximum points between the physical therapists were achieved in different rating scales. The best correlation between first and second observer was on Harris Hip Score ($\kappa = 0.91$), between second and third was on Merle d'Aubigne ($\kappa = 0.95$) and between first and third was on Iowa Hip Score ($\kappa = 0.87$). The worst correlation was between first and second observer on Charnley Hip Score ($\kappa = 0.77$), this was also more than 0.75 and rated as excellent (Table 1).

Discussion

Although hip scoring systems are commonly used by physical therapists, there is only one report in the literature evaluating the reliability of a hip scoring system among physical therapists.[6] However, the study of Söderman and Malchau[6] only evaluated the Harris Hip Score. For the other widely used scoring systems, such as the Merle d'Aubigne Hip Score, our study seems to be the first to investigate their reliability among physical therapists. This point is especially important in

Table 1 Interobserver reliability of five hip scores

	Merle d'Aubigne	Charnley	Harris	Iowa	AAOS
Observer 1 versus observer 2	0.79	0.77	0.91	0.89	0.85
Observer 2 versus observer 3	0.95	0.91	0.87	0.87	0.85
Observer 1 versus observer 3	0.82	0.80	0.82	0.86	0.78

Clinical messages

- Physical therapists can use all five commonly used hip scores with reliability.
- The expertise of the raters can affect the interobserver agreement.

that clinical scoring systems are performed more often by physical therapists.[6,9–11]

In the present study, we found excellent correlation among the physical therapists (Table 1). This is probably due to the experience of the physical therapists (3–5 years) in the present study, for it is well known that the expertise of the raters can affect interobserver agreement.[12,13]

Studies evaluating the reliability of scoring systems should avoid certain points. First, the study should include similar subjects, raters, setting, and the manner in which the evaluations are done.[14] Second, it is not meaningful to compare outcome scores with highly variable postarthroplasty patients, different implants and primary versus revision procedures.[14,15] Third, the so called 'learning effect' should be avoided; for in clinical studies, repeated patient examinations by different observers using multiple evaluations systems may be problematic because of learning effects and decreasing patient co-operation.[13]

In the current study we tried to avoid bias using similar subjects and raters with similar expertise, only including the preoperative cases and evaluating the patients in random order. The weak points of our study may be the relatively low number of subjects and raters. Nevertheless, because this is the first study to evaluate the reliability of five different hip rating scores among physical therapists, we can conclude that physical therapists can reliably use all these commonly used hip scores.

References

1 Lorna K. Case study: physical therapy management of hip osteoarthritis prior total hip arthroplasty (case study). *J Orthop Sports Phys Ther* 1997; **26**: 35–38.

2 Davidson M, Keating JL. A comparison of five low back disability questionnaires: reliability and responsiveness. *Phys Ther* 2002; **82**: 8–24.

3 Harris WH, Sledge CB. Total hip and total knee replacement. *N Engl J Med* 1990; **323**: 801–807.

4 Bach CM, Feizelmeier H, Kaufmann G, Sununu T, Göbel G, Krismer M. Categorization diminishes the reliability of hip scores. *Clin Orthop* 2003; **411**: 166–73.

5 Van Dillen LR, Roach KE. Reliability and validity of the acute care index of function for patients with neurological impairment. *Phys Ther* 1998; **68**: 1098–01.

6 Söderman P, Malchau H. Is the Harris hip score system useful to study the outcome of total hip replacement? *Clin Orthop* 2001; **384**: 189–197.

7 Murray D. The hip. In: Pynset P, Fairbank J, Carr A eds. *Outcome measures in orthopaedics.* Oxford: Butterworth-Heinemann, 1993: 199–215.

8 Goodwin RA. The Austin Moore prosthesis in fresh femoral neck fractures. (A review of 611 postoperative cases). *Am J Orthop Surg* 1968; **10**: 40–43.

9 Hoving JL, Buchbinder R, Green S et al. How reliably do rheumatologists measure shoulder movement? *Ann Rheum Dis* 2002; **61**: 612–15.

10 Flynn JM, Donohoe M, Mackenzie WG. An independent assessment of two clubfoot-classification systems. *J Pediatr Orthop* 1998; **18**: 323–27.

11 Springer BA, Arciero RA, Tenuta JJ, Taylor DC. A prospective study of modified Ottowa ankle rules in a military population. Interobserver agreement between physical therapists and orthopaedic surgeons. *Am J Sports Med* 2000; **28**: 864–68.

12 Liow RYL, Walker K, Wajid MA, Bedi G, Lennox CME. Functional rating for knee arthroplasty: comparison of three scoring systems. *Orthopedics* 2003; **26**: 143–47.

13 Bach CM, Nogler M, Steingruber IE *et al*. Scoring systems in total knee arthroplasty. *Clin Orthop* 2002; **399**: 184–96.

14 Garellick G, Malchau H, Herberts P. Specific or general health outcome measures in the evaluation of total hip replacement: a comparison between the Harris hip score and the Notthingham Health Profile. *J Bone Joint Surg* 1998; **80B**: 600–606.

15 Garellick G, Herberts P, Malchau H. The value of clinical data scoring systems: are traditional hip scoring systems adequate to use in evaluation after total hip surgery? *J Arthroplasty* 1999: **14**: 1024–29.