Reliability and Validity of Clinical Outcome Measurements of Osteoarthritis of the Hip and Knee - A Review of the Literature

Y. SUN, T. STURMER, K.P. GUNther*, H. BRENNER

Summary
High reliability and validity of clinical rating schemes is crucial for their use as outcome measurements of treatment of hip and knee osteoarthritis. In this paper, we review the empirical evidence on the reliability and validity of commonly used clinical scores. Clinical scores and related reliability and validity studies were identified by systematic literature search. Scores were classified according to the type and joint. Reliability and validity studies were characterized according to design, population, number and qualification of observers, number of measurements, time interval between repeat measurements and results. Reliability and validity studies were reported for only 6 and 15 of the 45 identified clinical scores, respectively. Although comparisons are difficult due to differences in study design, relatively high reliability was reported for most measurements of pain, stiffness, and physical function, while results are less conclusive for clinical signs. Most validity studies focused on the correlation between various scores. Correlation was generally found to be high for overall numerical ratings, but scores often differed with respect to the interpretation of these ratings. Validity has been more comprehensively studied for Lequesne’s scores, WOMAC, and ILAS, and these scores have shown satisfactory responsiveness to different treatment effects. Overall, knowledge on reliability and validity of clinical scores of hip and knee osteoarthritis is limited, underlining the need for further properly designed and conducted studies.

Key words Osteoarthritis, Clinical Assessment, Outcome Measurement, Reliability and Validity

INTRODUCTION
Osteoarthritis (OA) is the most common joint disease and a major public health problem throughout the world (1). OA of the hip and knee joints (cox- and gonarthrosis) is recognized as a major cause of pain, disability, and high social expenditure (1, 2). Treatment is usually aimed at reducing symptoms and preventing impairment and disability. Increasing importance is being placed on the monitoring of outcomes of treatment in clinical studies to investigate the possible therapeutic use of different therapies (such as surgical treatment, physical therapy, or drug therapy). Clinical assessment plays a central role for this purpose.

In the past few decades, a large number of clinical instruments for outcome measurements as well as severity ratings of hip and knee OA have been introduced. Knowledge is limited, however, on the appropriateness of various instruments for clinical and epidemiologic studies, which require a high level of reliability and validity of measurements.

In this paper, which was developed during the preparation of a multi-center study on the epidemiology of hip OA in South Germany, we review the empirical evidence on the reliability and validity of commonly used clinical rating systems of hip and knee OA.

SCOPE OF THIS REVIEW
This review will focus on the following aspects of studies on the inter-rater, intra-rater and test-retest-reliability and of the content and construct validity of commonly used clinical rating systems:
1. General characteristics and the special use of the scores.
2. Clinical items and their weighting included in the different clinical rating systems.
3. Setting and design of reliability studies, such as number and qualification of raters and number and spectrum of patients.
4. Intra-rater, inter-rater and test-retest-reliability for both single clinical items and overall scores.
5. Setting and design of validity studies, such as the qualification of observers and the spectrum of patients.
6. Content and construct validity.

Literature search

In order to comprehensively identify instruments of clinical outcome measurements of osteoarthritis of the hip and knee and studies on their reliability and validity, MEDLINE searches were performed for the time interval from 1984 to 1995, using the following controlled vocabulary: “osteoarthritis”, “index of severity”, “severity”, “clinical rating”, “clinical assessment”, “outcome measurement”, “reliability”, “validity”. Bibliographies and cross-referencing were used for identification of pre-1984 studies and complementation of the literature search.

Conceptual and statistical background for assessing reliability

Three types of reliability are commonly distinguished: 1) inter-rater-reliability indicates to which degree different observers, using a method to assess the same individual, obtain the same result (3). 2) Intra-rater-reliability indicates to which degree the same result is obtained, if the measurement is applied more than once on the same individual, by the same observer (3). 3) Test-retest-reliability indicates to which degree the same result is obtained in repeat applications of self-assessment instruments (3).

Statistical measures of reliability depend on the measurement scale: Pearson's correlation coefficient ($r$) (4) is commonly used to quantify correlation between repeat measurements of continuous variables. These variables should be normally distributed. Spearman's rank correlation coefficient ($r_s$) (4) is often used to assess the reliability of variables that do not follow the normal distribution. An alternative to Spearman's rank correlation coefficient is Kendall's tau ($\tau_p$, $\tau_r$, $\tau_s$), (5) which is somewhat less frequently used. All of these correlation coefficients can take values from $-1$ (maximum possible negative correlation) to $+1$ (perfect positive correlation). A limitation of these correlation coefficients for quantifying reliability is that they do not reflect systematic variation between observers or between measurements. Some authors proposed the use of t-tests for paired comparisons along with the correlation coefficient to reflect systematic variations. It should be noted, however, that the test statistic reflects the size of the sample in addition to the difference between ratings. Furthermore, none of the aforementioned approaches can be used to quantify reliability of more than two measurements per study participant.

An alternative measure of reliability that reflects both systematic and random variation between tests is the intraclass correlation coefficient (ICC) (6). This coefficient is based on the estimation of variance components in analysis of variance. ICC quantifies the proportion of overall variance of ratings that is due to between-subjects variability, and it can therefore take values from 0 (variance entirely due to imperfect reliability) to 1 (variance entirely due to between-subjects variability). ICC can be used for two or more measurements per study participant (6).

Reliability of categorical data is commonly quantified by kappa coefficients ($\kappa$) (7) which quantify the agreement of classification beyond chance agreement. Although primarily developed for dichotomous variables, kappa coefficients can also be applied to variables with more than two categories. For ordinal variables, weighted kappa coefficients are often used in which disagreements are weighted by the magnitude of the discrepancy between ratings. Weighted kappa coefficients are equivalent to Pearson’s correlation coefficient and the intraclass correlation coefficient applied to the categorical data under certain conditions (7). Kappa coefficients have a maximum value of 1 when agreement is perfect. A value of 0 indicates no agreement beyond agreement by chance, and a value below 0 is observed with less than chance agreement. Although the magnitude of kappa depends on a variety of factors other than reliability, such as the marginal distributions of ratings, values greater than 0.75 are generally considered to represent excellent agreement beyond chance, while values below 0.40 are considered to reflect poor agreement (7).

An alternative to the kappa coefficient is the Goodman-Kruskal’s gamma coefficient ($\gamma$) (8), which is less frequently used. The gamma coefficient is defined as the difference between agreement and disagreement of paired observations divided by the sum of agreement and disagreement. It can take values from $-1$ (maximum possible disagreement) to $+1$ (perfect agreement). A limitation of the gamma coefficient for quantifying reliability is that it can only be used for two measurements per study participant. Furthermore, it does not take chance agreement into account.