Clin Orthop Relat Res (2012) 470:3499–3505
DOI 10.1007/s11999-012-2534-x

Clinical Orthopaedics and Related Research®
A Publication of The Association of Bone and Joint Surgeons®

CLINICAL RESEARCH

Reliability of Bucholz and Ogden Classification for Osteonecrosis Secondary to Developmental Dysplasia of the Hip

Andreas Roposch MD, MSc, FRCS, John H. Wedge OC, MD, FRCS(C), Georg Riedl MD

Received: 2 February 2012 / Accepted: 1 August 2012 / Published online: 18 August 2012

Abstract

Background Osteonecrosis is perhaps the most important serious complication after treatment of developmental dysplasia of the hip (DDH). The classification by Bucholz and Ogden has been used most frequently for grading osteonecrosis in this context, but its reliability is not established and unreliability could affect the validity of studies reporting the outcome of treatment.

Questions/Purpose We established the interrater and intrarater reliabilities of this classification and analyzed the frequency and nature of disagreements.

Methods Three pediatric hip surgeons, a musculoskeletal pediatric radiologist, and three orthopaedic trainees graded 39 radiographs (hips) according to the Bucholz and Ogden classification, blinded to any clinical data. Ratings were repeated after 2 weeks. Interrater reliability and intrarater reliability were determined using the simple kappa statistic. Grading was compared among raters, the nature and frequency of disagreements established, and subgroup analyses performed.

Results Interrater reliability was 0.34 (95% CI = 0.28, 0.40) for all raters, and 0.31 (0.20 to 0.43) for the three surgeons. The best interrater reliability was observed between the radiologist and a surgeon with a kappa of 0.51 (0.30, 0.72). Intrarater reliability estimates ranged from 0.44 to 0.69. Raters disagreed regarding the grade of osteonecrosis in 26 of 39 hips (67%), with seven of 26 disagreements (27%) involving confusion between Grades I and II.

Conclusions The interrater reliability was lower than expected, considering the raters' experience. Distinguishing between Grades I and II was the most frequently observed problem. We believe that the low reliability was a result of an ambiguous classification scheme rather than the variability among the raters. Outcome studies of DDH based on this classification should be interpreted with caution. We recommend the development of a new classification with better prognostic ability.

Level of evidence Level III, diagnostic study. See the Guidelines for Authors for a complete description of levels of evidence.

The institution of one or more of the authors (AR) has received, in any one year, funding from Arthritis Research UK.
All ICMJE Conflict of Interest Forms for authors and *Clinical Orthopaedics and Related Research* editors and board members are on file with the publication and can be viewed on request.
This work was performed at the Institute of Child Health, University College London, London, England.

A. Roposch (✉), G. Riedl
Department of Orthopaedic Surgery, Great Ormond Street Hospital for Children, Institute of Child Health, University College London, London WC1N 3JH, England, UK
e-mail: a.roposch@ich.ucl.ac.uk

J. H. Wedge
Division of Orthopaedic Surgery, Hospital for Sick Children, University of Toronto, Toronto, Canada

Introduction

Osteonecrosis of the capital femoral epiphysis is a major complication of treatment for developmental dysplasia of the hip (DDH) with reported incidences ranging from 6% to 48% [14]. This irreversible condition is associated with subsequent hip pain and declining hip function in childhood [12]. Premature arthritis requiring hip arthroplasty as early as during the third decade is common with severe forms [12]. For these reasons, osteonecrosis is considered one of the most important quality indicators of DDH treatment.

The diagnosis of osteonecrosis secondary to DDH is established using radiography. Among several classifications systems [7, 9, 15], the one by Bucholz and Ogden [1] has been used most widely [14]. It is a four-grade system that is based on the morphologic changes of the proximal femur that occur as a result of ischemic necrosis [1].

Although widely used, the reliability of this classification has never been established. We realized in previous studies [12–14, 16] that classifying hips as described by Bucholz and Ogden [1] can be challenging and confusing with questionable reliability. Reliability, or reproducibility, is a prerequisite of any scientific measure. A measure is regarded as reliable if the same result is obtained when a test is repeated by the same (intrarater reliability) or different (interrater reliability) investigators [18]. Potential sources of measurement inconsistency relate to the operator performing the measurement (ie, the observers' skills in evaluating a radiograph for the presence of osteonecrosis) and the measure itself (ie, the classification system) [18]. For example, ambiguously defined conversion criteria (ie, the criteria by which to convert morphologic features of the hip into one of four grades) or the lack of explicit criteria to distinguish a mild Grade II from a severe Grade I are potential sources to affect reliability.

We therefore (1) determined the interrater reliability and intrarater reliability of the Bucholz and Ogden classification for osteonecrosis secondary to DDH and the frequency and nature of disagreements; (2) determined whether combining Bucholz and Ogden Grades I and II osteonecrosis would modify the reliability; and (3) compared the interrater reliability of experienced clinicians and orthopaedic trainees.

Materials and Methods

Three pediatric hip surgeons (AR, JHW, NC), a musculoskeletal pediatric radiologist (AO), and three orthopaedic trainees (GR, OO, SM) graded the radiographs. Of the three pediatric hip surgeons, two were at the end-stage of their career and one was 8 years postfellowship training. The radiologist had 7 years of experience as an attending musculoskeletal radiologist. To contrast the reliability of experienced professionals, we included an orthopaedic senior resident, an orthopaedic fellow, and a medical student. The radiographs chosen were those of children who had osteonecrosis develop secondary to a closed or open reduction for DDH. There were 31 girls (79.5%) and eight boys (20.5%). There were 13 (33%) right and 26 (67%) left hips. The age of the patients at reduction was 1.9 ± 1.3 years (range, 0.5–5.5 years), with a mean followup of 9.4 ± 2.7 years (range, 2.6–14.7 years). The children were 11.5 ± 2.6 years old (range, 3.3–16.8 years) when the radiographs were taken.

All raters familiarized themselves with the classification by studying the monograph by Bucholz and Ogden [1] individually and again in a group meeting to agree on the definitions of the classification. Several radiographs (not included in this study) were interpreted in the group before each of the raters graded 39 radiographs individually. The three orthopaedic surgeons had used the classification for several years before this study. The medical student wrote a thesis on the subject, therefore had researched the topic.

Two of us (AO, AR) screened the followup radiographs of 222 children who had undergone surgical treatment for DDH in infancy, for the presence of osteonecrosis using the criteria by Bucholz and Ogden. Grades of severity were not assigned at this stage; instead radiographs were graded with a binary response, ie, "osteonecrosis present/absent". A third expert (JHW) was consulted to resolve disagreements. This process resulted in 89 radiographs depicting features of osteonecrosis and 39 were chosen randomly for the reliability study. This step ensured that all the included hips showed osteonecrosis as per Bucholz and Ogden's definition and none of the hips was normal. We elected not to include normal hips because the Bucholz and Ogden classification focuses on four grades of osteonecrosis. This consensus grading only established a yes or no answer and no attempt was made to grade osteonecrosis.

A standard radiographic protocol was used in all patients. These were radiographs of the pelvis taken with the patient in the supine position, centered on the hips, and with both feet in 15° internal rotation, made depending on the age of the patient at 60 to 80 kV, 4 to 40 mA, and a focus-to-film distance of 150 cm on a digital imaging system (5000R CR, Fuji, Bedford, UK). Images were analyzed electronically (Sienet Sky, Siemens AG Medical Solutions, Erlangen, Germany), blinded to patient history and current symptoms. To determine the intrarater reliability, two ratings were obtained at 2-week intervals with the radiographs in changed order. Interrater reliability was based on the first rating.

In the Bucholz and Ogden classification, Grade I changes refer to irregular ossification or hypoplasia of the femoral head but normal ossification of the metaphysis (Fig. 1). In Grade II, the femoral head will grow into valgus (Fig. 2). For Grade III, the entire metaphysis is involved resulting in shortening of the femoral neck with trochanteric overgrowth (Fig. 3). Grade IV is characterized by varus of the proximal femur (Fig. 4).

Sample size calculation was based on the primary outcome with the aim of showing a reliability that was at least 0.60; the power was set to 80%, $\alpha = 0.05$, and $\beta = 0.20$. Using this approach sample sizes of $n = 4$ raters and $k = 39$ radiographs per rater were calculated for each group [4]. Data were categorical with mutually exclusive categories, and raters were independent; therefore Cohen's simple

Fig. 1 Grade 1 osteonecrosis is seen on this radiograph. The distinguishing features include a hypoplastic epiphysis of the proximal femur.

Fig. 3 Grade III osteonecrosis is seen in this radiograph. The changes include trochanteric overgrowth, shortening of the femoral neck, and asphericity of the femoral head.

Fig. 2 The pelvic radiograph shows Grade II osteonecrosis. Damage to the lateral aspect of the growth plate results in valgus alignment of the proximal femur. Typically, a horizontally oriented growth plate is seen as a result of this irreversible growth disturbance.

Fig. 4 Grade IV osteonecrosis is shown. Damage to the medial aspect of the growth plate leads to a growth disturbance resulting in varus alignment of the proximal femur. As a result of the varus alignment, relative trochanteric overgrowth is seen. The femoral head appears hypoplastic.

kappa statistic [3] was used as a measure of agreement for comparisons between two raters. An overall kappa was calculated as described by Fleiss [6] to determine the reliability among three or more raters. Interrater reliability was established from the first reading process. To contrast the interrater statistics of the four clinicians to a group of less experienced professionals, we established interrater coefficients of a senior orthopaedic resident, orthopaedic

research fellow, and graduate medical student. They prepared for this task in the same way as the clinicians did. Our a priori assumption was that interrater coefficients of this group of trainees should not exceed those of the four experienced clinicians. To test the effect of a three-grade scheme on reliability we derived interrater coefficients for a modified classification scheme that combined Grades I and II into one category. All analyses were performed using SAS 9.2 statistical package (SAS Institute Inc, Cary, NC, USA).

Results

Among the four experienced raters, the interrater reliability was low with a kappa statistic of 0.34 (95% CI = 0.28–0.40). Subgroup analysis showed that the interrater reliability was highest between the radiologist and one of the surgeons with a kappa of 0.51 (0.30–0.72) (Table 1). The best intrarater reliability was the result of a surgeon's rating with a kappa of 0.69 (0.49–0.89). Grade II osteonecrosis was diagnosed most often by each of the raters but the remainder of the distributions of the grades varied among raters (Fig. 5). Of the 39 hips rated, all raters agreed in 13 (33%) and disagreed in 26 (67%) instances (Table 2). On 16 (62%) occasions, the disagreements were on two levels (ie, for the same hip the raters assigned two different grades), and on 10 occasions (38%) the disagreements were on three levels (ie, for the same hip the raters assigned three different grades). There was no four-level disagreement. The most common two-level disagreement included confusion between Grades I and II with 27% (seven of 26) of all disagreements. The most common three-level disagreement involved Grades I, II, and IV with 23% (six of 26) of all disagreements (Fig. 6).

Collapsing the classification from four to three categories by combining Grades I and II improved the interrater reliability to a kappa of 0.42 (0.35, 0.50) among the four experienced raters.

In contrast to the kappa value of 0.34 for the experienced observers, the trainees' interrater reliability ranged from 0.12 to 0.30 (Table 3).

Table 1. Results of clinician's reliability coefficients based on interrater and intrarater studies of 39 hips.

Rater	Kappa	95% confidence interval
Interrater		
All four raters	0.34	0.28–0.40
All three surgeons	0.32	0.20–0.43
Surgeon 1 – Surgeon 2	0.46	0.25–0.67
Surgeon 1 – Surgeon 3	0.22	0.02–0.43
Surgeon 2 – Surgeon 3	0.22	0.03–0.40
Radiologist – Surgeon 1	0.34	0.13–0.55
Radiologist – Surgeon 2	0.31	0.09–0.53
Radiologist – Surgeon 3	0.51	0.30–0.72
Intrarater		
Surgeon 1	0.69	0.49–0.89
Radiologist	0.61	0.39–0.83
Surgeon 2	0.52	0.30–0.73
Surgeon 3	0.44	0.21–0.67

Discussion

Osteonecrosis is one of the most frequent and perhaps the most serious complication of treatment of DDH. It is

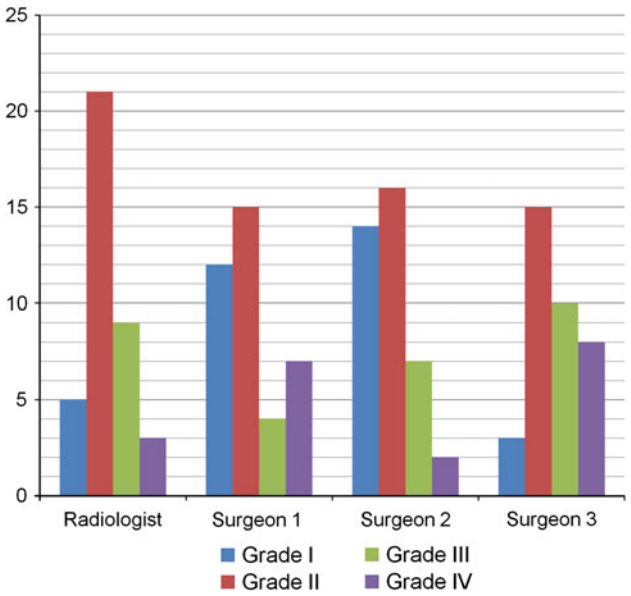

Fig. 5 The graph shows distribution and frequency of grades assigned by each of the raters. Although variability is noted among all the raters, Grade II was assigned most often by all raters. Absolute numbers are shown.

Table 2. Number of disagreements across all radiographs

Disagreement in Bucholz and Ogden grade	Frequency of disagreements	
	Number	Percentage
I and II	16	35%
II and IV	9	20%
I and IV	9	18%
II and III	7	15%
I and III	4	9%
III and IV	1	2%

Five raters evaluated each radiograph, therefore the total number of disagreements exceeds the total number of radiographs.

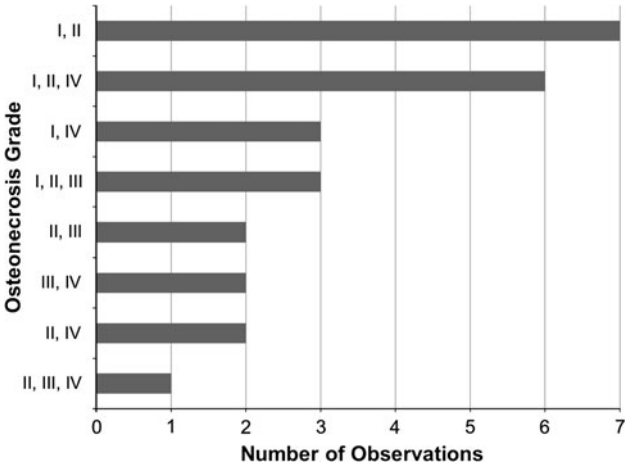

Fig. 6 The graph shows the nature and frequency of the 26 disagreements based on single radiographs. Disagreements occurred on two levels on five occasions and on three levels on three occasions. For example, for the same hip the four raters assigned Grades II, III, and IV on one occasion (bottom bar). The most common confusion was between Grades I and II (top bar) and occurred on seven occasions (27%).

Table 3. Results of trainees' interrater reliability coefficients.

Rater	Kappa	95% confidence interval
Resident – student	0.24	0.08–0.40
Resident – research fellow	0.12	0–0.26
Research fellow – student	0.30	0.12–0.47

essential that grading schemes of osteonecrosis are accurate, especially because interventions are considered based on the type of osteonecrosis [12]. We therefore (1) determined the interrater reliability and intrarater reliability of the Bucholz and Ogden classification for osteonecrosis secondary to DDH and the frequency and nature of disagreements; (2) determined whether combining Bucholz and Ogden Grades I and II osteonecrosis would modify the reliability; and (3) compared the interrater reliability of experienced clinicians and orthopaedic trainees.

There are potential limitations of this study. First, our findings are based primarily on experienced clinicians and are generalizable to this group of professionals only. We acknowledge that radiologists do not use this classification system in general and are not involved with treatment decisions in these cases. However, a musculoskeletal radiologist was chosen to participate in this study to determine the impact of a different training background on reliability. We do not think that much better interrater coefficients would be found in other groups. Second, although it can be difficult to distinguish normal hips from those with Bucholz and Ogden Grade I osteonecrosis, we elected not to include hips without osteonecrosis in our study because the Bucholz and Ogden classification does not include a category for normal hips. Including normal hips would have introduced the potential of falsely underestimating the reliability of the classification. Therefore, we could not quantify whether and to what degree it is problematic to distinguish normal hips from those with Grade I osteonecrosis. Third, we were unable to determine the influence of patient age at the time of examination on reliability. Fourth, the 39 radiographs included were part of a larger pool of radiographs that were graded for the presence or absence of osteonecrosis, partly in consensus. Because some raters were involved in both processes, there is a potential for recall bias. However, we believe the risk is minimal given the large initial number of radiographs.

With a kappa of 0.34 and an upper bound of the 95% confidence interval of 0.40, interrater reliability was lower than expected considering the experience of the raters. Because of the paucity of data provided in previous research (a meta-analysis found that six studies on DDH that reported osteonecrosis as the primary outcome did not produce any reliability estimates [14]), it remains unclear if better reliability has been established elsewhere for this grading scheme. One study [13] reported a kappa of 0.47 between two raters, which is nearly identical to our findings. A recent study [5] reported a weighted kappa statistic of 0.33 for the interrater reliability of the Kalamchi and MacEwan classification [9], which is similar to the Bucholz and Odgen classification. Apart from the clinician, variability in measurement can arise from two other sources: patient and procedure [18]. Variability in the procedure, such as a classification system, may be attributable to the way the procedure has been devised. Variability in this context is caused by a lack of explicit criteria in the classification. Conversion criteria (ie, the criteria by which to convert a radiographic image into one of four grades) are vaguely defined in the classification of Bucholz and Ogden leaving too much room for personal interpretation. Our findings support this observation because although the interrater reliability was low, higher coefficients were found in the intrarater studies. This means that the raters were consistent in their personal interpretations of what constitutes Grades I to IV, but did not agree among themselves. Without explicit and precisely formulated conversion criteria, reliability will suffer [18]. When explicit conversion criteria are available, such as with the Graf classification for hip ultrasound [8], they contribute to improved reliability [11].

Disagreements between the experienced raters occurred in 69% of all hips. Variability also can arise from the lack of discriminative ability of a classification system that is caused by an overlap of the features that define the different categories. Discriminative ability is particularly important when the condition of interest is a continuum

with no clear biological margins. For example, a severe Grade I and a mild Grade II could be confused if the classification does not require explicit criteria that allow one to distinguish and categorize such extremes. Based on our findings we believe the Bucholz and Ogden classification fails to address this problem and its discriminative ability is limited, which, in turn, affects reliability. The most striking example, as seen in our study, was the confusion between Grades II (valgus) and IV (varus). Assuming that the clinicians understood the difference between varus and valgus, this is additional evidence that the scheme has limited discriminative ability and does not fully account for radiographic features that could lead to such confusion.

Raters found it most difficult to discriminate Grade I from Grade II (Fig. 6). This finding is important because some suggested classifying Grade I changes as normal [14] because the radiographic features associated with Grade I are subtle. Following this suggestion, the kappa statistic for interrater reliability improved from 0.34 to 0.42. However, the improvement was marginal, calling into question the overall benefit of a three-grade system regarding reliability; especially as other research suggested that hips with Grade I changes are not the same as normal hips for hip function and such an approach does not seem plausible [12].

Clinicians vary in their observation, extraction, and interpretation of information. To reduce variability among raters and to facilitate producing the best possible reliability, we focused on clinicians with expertise in interpreting radiographic findings of the hip but also contrasted their findings to those of less experienced orthopaedic trainees. Although the four clinicians had prior experience in using this classification [2, 12–14, 16], their intrarater reliability did not reach the recommended 0.75 mark but was still markedly better than the interrater reliability of the trainees. Clinical experience of raters has been associated with improved reliability in previous research [11]. The low interrater reliability among trainees suggests that osteonecrosis is difficult to classify using the criteria by Bucholz and Ogden.

Our study has important clinical implications. First, independent of context, Fleiss viewed reliability coefficients of 0.75 and greater as excellent [6]. Some authors believe lower coefficients are acceptable in the context of research [10, 13], whereas Streiner and Norman suggested reliability coefficients of measures for individual judgments should reach 0.75 [17]. We therefore suggest that the reliability of the Bucholz and Ogden classification is not sufficient enough to be used in day-to-day clinical practice. Second, findings from outcome studies of DDH based on this classification should be interpreted with caution as they might be flawed. Considering that osteonecrosis is one of the main quality indicators in evaluating the care for children with DDH, the reliability of the classification warrants substantial improvement. In subsequent research we will test strategies directed at these two sources for their ability to improve reliability, particularly multirater agreement. In revising the classification, emphasis needs to be placed on prognostic ability. Bucholz and Ogden Grades II, III, and IV, in their severe forms, ultimately lead to a decline in hip function, pain, and degenerative joint disease [5, 7, 12], so perhaps this distinction is all that matters. A classification system that simply grades osteonecrosis as mild and severe might be more practical, discriminative, and predictive of outcome.

Acknowledgments We thank Amaka Offiah PhD and Nicholas Clarke FRCS for help in grading radiographs, Odeh Odeh MBBS and Sheriffe Montgomery FRCS for participating in the reliability studies of orthopaedic trainees, and Evangelia Protopapa for help with the graphs.

References

1. Bucholz RW, Ogden JA. Patterns of ischemic necrosis of the proximal femur in nonoperatively treated congenital hip disease. *The Hip: Proceedings of the Sixth Open Scientific Meeting of the Hip Society.* St Louis, MO: Mosby; 1978;43–63.
2. Clarke NM, Jowett AJ, Parker L. The surgical treatment of established congenital dislocation of the hip: results of surgery after planned delayed intervention following the appearance of the capital femoral ossific nucleus. *J Pediatr Orthop*. 2005; 25:434–439.
3. Cohen J. Coefficient of agreement for nominal scales. *Educ Pychol Meas.* 1960;20:37–46.
4. Donner A, Eliasziw M. Sample size requirements for reliability studies. *Stat Med.* 1987;6:441–448.
5. Firth GB, Robertson JF, Schepers A, Fatti L. Developmental dysplasia of the hip: open reduction as a risk factor for substantial osteonecrosis. *Clin Orthop Relat Res.* 2010;468:2485–2494.
6. Fleiss JL. The measurement of interrater agreement. *Statistical Methods for Rates and Proportions*. New York, NY: John Wiley & Sons; 1981:212–36.
7. Gage JR, Winter RB. Avascular necrosis of the capital femoral epiphysis as a complication of closed reduction of congenital dislocation of the hip: a critical review of twenty years experience at Gillette Children's Hospital. *J Bone Joint Surg Am.* 1972; 54:373–388.
8. Graf R. Fundamentals of sonographic diagnosis of infant hip dysplasia. *J Pediatr Orthop.* 1984;4:735–740.
9. Kalamchi A, MacEwen GD. Avascular necrosis following treatment of congenital dislocation of the hip. *J Bone Joint Surg Am.* 1980;62:876–888.
10. Marx RG, Connor J, Lyman S, Amendola A, Andrish JT, Kaeding C, McCarty EC, Parker RD, Wright RW, Spindler KP; Multicenter Orthopaedic Outcomes Network. Multirater agreement of arthroscopic grading of knee articular cartilage. *Am J Sports Med.* 2005;33:1654–1657.
11. Roposch A, Graf R, Wright JG. Determining the reliability of the Graf classification for hip dysplasia. *Clin Orthop Relat Res.* 2006;447:119–124.
12. Roposch A, Liu LQ, Offiah A, Wedge JH. Functional outcomes in patients with osteonecrosis secondary to developmental dysplasia of the hip. *J Bone Joint Surg Am.* 2011; 93:e145.

13. Roposch A, Odeh O, Doria AS, Wedge JH. The presence of an ossific nucleus does not protect against osteonecrosis after treatment of developmental dysplasia of the hip. *Clin Orthop Relat Res.* 2011;469:2838–2845.
14. Roposch A, Stohr KK, Dobson M. The effect of the femoral head ossific nucleus in the treatment of developmental dysplasia of the hip: a meta-analysis. *J Bone Joint Surg Am.* 2009; 91:911–918.
15. Salter RB, Kostuik J, Dallas S. Avascular necrosis of the femoral head as a complication of treatment for congenital dislocation of the hip in young children: a clinical and experimental investigation. *Can J Surg.* 1969;12:44–61.
16. Spence G, Hocking R, Wedge JH, Roposch A. Effect of innominate and femoral varus derotation osteotomy on acetabular development in developmental dysplasia of the hip. *J Bone Joint Surg Am.* 2009;91:2622–2636.
17. Streiner DL, Norman GR. *Health Status Measurement Scales*. 3rd ed. New York, NY: Oxford Press; 2003.
18. Wright JG, Feinstein AR. Improving the reliability of orthopaedic measurements. *J Bone Joint Surg Br.* 1992;74:287–291.