Variations in the Use of Diagnostic Criteria for Developmental Dysplasia of the Hip

Andreas Roposch MD, MSc, FRCS,
Liang Q. Liu MBBS, PhD, Evangelia Protopapa MSc

Received: 15 September 2012 / Accepted: 31 January 2013 / Published online: 21 March 2013
© The Association of Bone and Joint Surgeons® 2013

Abstract

Background Wide variation exists in reported prevalence estimates and management standards of developmental dysplasia of the hip (DDH). Discrepancies in diagnosticians' opinions may explain some of this variation.

Questions/purposes We sought to determine (1) the consistency with which pediatric orthopaedic surgeons rate the importance of diagnostic criteria for DDH, and (2) whether there were geographic differences in how the diagnostic criteria were rated by surgeons.

Methods One hundred ninety-seven of 220 members of the European Paediatric Orthopaedic Society and 100 of 148 members of the British Society of Children's Orthopaedic Surgery treating children with DDH participated in this cross-sectional study across 35 countries (15 regions). Each rated 37 items in four domains that specialists previously had identified as the most important features associated with DDH in early infancy. We determined

consistency using the intraclass correlation coefficient (ICC; two-way random-effects model) interpreted as poor (0–0.40), acceptable (0.41–0.74), or good (≥ 0.75).

Results Poor consistency among surgeons was found in rating the 37 diagnostic criteria (ICC, 0.33; 95% CI, 0.24–0.45). Consistency was poor for three domains (patient characteristics/history: ICC, 0.29; 95% CI, 0.16–0.58; ultrasound: ICC, 0.26; 95% CI, 0.14–0.52; radiography: ICC, 0.34; 95% CI, 0.12–0.95) and acceptable for one (clinical examination: ICC, 0.50; 95% CI, 0.33–0.73). Surgeons in particular regions appeared to have a concept of DDH diagnosis that distinguished them from specialists of other regions; consistency in eight regions was greater (ICC ≥ 0.40) than consistency among all 15 regions.

Conclusions The consistency of specialists in rating diagnostic criteria for DDH was lower than expected, and there was considerable geographic variation in terms of how specialists assigned importance ratings of the diagnostic criteria; these findings are somewhat counter-intuitive, given the frequency with which this condition is diagnosed. These inconsistencies could explain, partly, the widely differing prevalence estimates and management standards of DDH.

The institution of one or more of the authors (AR) has received, during the study period, funding from Bupa Foundation and Great Ormond Street Hospital Children's Charity.

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.

Each author certifies that his or her institution approved the human protocol for this investigation, that all investigations were conducted in conformity with ethical principles of research, and that informed consent for participation in the study was obtained.

A. Roposch (✉), L. Q. Liu, E. Protopapa
Institute of Child Health, University College London,
30 Guildford Street, London WC1N 3EH, UK
e-mail: a.roposch@ich.ucl.ac.uk

A. Roposch, L. Q. Liu, E. Protopapa
Department of Orthopaedic Surgery, Great Ormond
Street Hospital for Children, London, UK

Introduction

The diagnosis of developmental dysplasia of the hip (DDH) in early infancy remains controversial [25, 32], especially because abnormal clinical and ultrasonic findings can resolve spontaneously in this age group [12, 28]. Clinicians disagree about the importance of diagnostic findings obtained from physical examination and diagnostic imaging, and about the importance of potential risk factors [21].

Prevalence estimates for DDH in this age group range from two to 28 per 1000 when applying criteria obtained from a physical examination and 40 to 60 per 1000 [15, 18, 26, 29] when using ultrasonographic criteria. The wide range in prevalence estimates could be attributed to intrinsic differences in the populations studied or, perhaps more likely, to the way the diagnosis is made.

Diagnostic practices for DDH in early infancy differ among countries [5, 32]. For example, the UK screening guidelines for DDH include clinical criteria and recommend use of ultrasonography only for a selected group, but in other countries the diagnosis is based solely on ultrasonographic criteria [30]. To the best of our knowledge, no studies regarding variability in diagnostic practices for DDH have been published. We theorized that inconsistencies in diagnostic practices among pediatric orthopaedic surgeons exist for DDH and that these inconsistencies could largely explain the wide variation in prevalence estimates and could be linked to widely differing standards in the management of DDH. Variability in physician opinion has been identified as a major factor affecting quality of care [8, 11, 19], usually for the worse [8]. The aims of this study were to (1) quantify the consistency with which pediatric orthopaedic surgeons of different geographic regions rate the importance of diagnostic criteria of DDH in early infancy, and (2) describe variations in consistency among surgeons based on geographic regions.

Materials and Methods

The study was approved by the institutional review board, the European Paediatric Orthopaedic Society (EPOS), and the British Society of Children's Orthopaedic Surgery (BSCOS). It was part of a larger-scale research project with the aim to establish a novel diagnostic index for DDH. In previous research [23], we compiled 37 criteria that were considered the most important features associated with DDH in infants not older than 8 weeks by conducting a survey of all members of EPOS and BSCOS, key informant interviews, and a literature review. EPOS is one of the two largest learned societies for pediatric orthopaedics, with members from more than 30 countries. The 37 criteria were grouped into four domains: patient characteristics/history, clinical examination, ultrasonography, and radiography (Table 1).

We surveyed all members of EPOS and BSCOS who treat patients with DDH using the tailored design method [7]. Of 261 members of EPOS who were contacted, 41 surgeons did not treat DDH, had retired, or were deceased, leaving 220 eligible participants. Of these, 197 (90%) responded and 156 (71%) returned fully completed questionnaires. The response rate for members of BSCOS was 68% (100 of 148).

We asked participants to rate each of the 37 criteria for their relative importance in the diagnosis of DDH in infants not older than 8 weeks. For the purpose of this study, we defined DDH as a condition that warrants either treatment or followup with an orthopaedic surgeon. Respondents rated the 37 criteria on a 10-cm VAS with the anchors "completely unimportant" and "extremely important." They were asked to consider all the listed criteria regardless of other abnormalities and to rate each criterion as if it was the only finding occurring in one patient. We recognized this approach does not reflect clinical practice and could affect the ratings in that some criteria could be rated too high and others too low. However, we assumed, because each criterion was rated in isolation, the relative importance rating would be stable [14]. We chose this approach because our study was not designed to determine how surgeons make clinical decisions regarding the diagnosis of DDH but to quantify the degree of consistency in assigning importance to a set of 37 predefined items. By contrast, in a holistic approach two or more factors are assessed simultaneously and in all possible combinations. While this approach is closer to the clinical setting, for the purpose of this study, an explicated approach was more appropriate because we were solely interested in the consistency of surgeon-based VAS ratings. A sample size of 37 items with 156 raters for each item achieves 81% power to detect an intraclass correlation coefficient (ICC) of 0.80 under the alternative hypothesis when the ICC under the null hypothesis is 0.69 using an F-test with a significance level of 0.05 [31].

We used three methods to analyze the data [13]: (1) a ratio scale measurement of the perceived value on the VAS provided a continuous outcome; (2) the continuous outcome was converted to an ordinal outcome (by rounding the 10-cm VAS scale to round numbers) with the aim to place less weight on the precise position on the VAS; and (3) a binary outcome was created by determining whether each criterion was in a respondent's top 10 most important criteria. This third measure was calculated against the possibility that low interrater agreement was more a function of disagreement on the least important items than on disagreement regarding which items were more important [13]. For each scoring method, we calculated the ICC for all 37 criteria as a whole and for each domain. We defined 15 regions based on geography, common heritage, and feasibility and calculated an ICC for each region to study variations by area. ICC is used to assess agreement of quantitative measurements in the sense of consistency. The concept of consistency is defined as the agreement of two quantitative measurements in settings where neither one is assumed correct [10]. The participating surgeons were multiple raters of the criteria, and because all items were evaluated by all raters, we used the Case 2 model according

Table 1. The 37 criteria in four domains rated by study participants for their importance in the diagnosis of DDH

Domain	Criteria
Physical examination	Leg length discrepancy or positive Galeazzi sign Positive Ortolani test/sign Positive Barlow test/sign Asymmetry of groin or skin crease(s) Any asymmetry of hip abduction (with both hips in flexion) Hip click in a stable (Ortolani- and Barlow-negative) hip Congenital clubfoot or other fixed foot deformities Postural (flexible) foot deformities such as metatarsus adductus/varus Torticollis (any form) Abduction of both (or one) hips limited to 70° (with both hips in flexion) Abduction of both (or one) hips limited to 60° (with both hips in flexion) Abduction of both (or one) hips limited to 45° (with both hips in flexion) Asymmetry in abduction of 20° or more
Ultrasound	On static ultrasound, femoral head coverage $\leq 45\%$ (coronal view) On static ultrasound, femoral head coverage $\leq 50\%$ (coronal view) On static ultrasound, femoral head coverage $\leq 60\%$ (coronal view) On static ultrasound, femoral head coverage $\leq 70\%$ (coronal view) On dynamic ultrasound, displacement of femoral head > 2 mm from the medial aspect of the acetabulum (either coronal or transverse view) Femoral head displaced anatomically with no congruency on stress ultrasound (either coronal or transverse view) Dislocatable hip on dynamic ultrasound (either coronal or axial view) On static ultrasound, α angle $< 60^\circ$ (Graf method) On static ultrasound, α angle $< 55^\circ$ (Graf method) On static ultrasound, α angle $< 50^\circ$ (Graf method) On static ultrasound, α angle $< 45^\circ$ (Graf method)
Patient characteristics/ history	Female sex First-born baby girl First-degree relative was treated in any way for DDH (as per parental report) Any family history (also nonfirst-degree relatives) of DDH in childhood (as per parental report) Birth weight > 4000 g (8.8 pounds) Oligohydramnios (low volume of amniotic fluid on prenatal ultrasound [as per practitioner's definition] in the absence of intrauterine growth restriction, fetal anomaly, or significant maternal comorbidity) during any stage of pregnancy Born by caesarean section Breech presentation born either by vaginal delivery or caesarean section History of breech positioning in utero but born by vertex delivery Multiple births or pregnancies of mother (multiparous mother)
Radiography	Acetabular index $> 30^\circ$ at 8 weeks as a primary imaging modality in this age group Acetabular index $> 25^\circ$ at 8 weeks as a primary imaging modality in this age group Migration of the midpoint of the femoral metaphysis lateral to Perkins' line as a primary modality in this age group (Perkins' line is a vertical line drawn through the most superolateral margin of the ossified acetabulum)

DDH = developmental dysplasia of the hip.

to Shrout and Fleiss [27]. The ICC provides a measure of the extent to which any single rater identified at random would compare with any other randomly selected single rater. The ICC therefore is a measure of the extent to which individual participants would agree on one criterion. The dependence of the ICC on the variance of the population

illustrates that a calculated ICC has no absolute meaning [27]. Naturally, the interpretation of the ICC is somewhat arbitrary and depends on the context. A general and widely accepted guide for its interpretation in health research recommends the following thresholds: 0 to 0.4, poor consistency or large variation in opinion; 0.41 to 0.74,

Table 2. Consistency among members of the EPOS based on three different scoring methods

Scoring method	ICC (95% CI)				
	All 37 criteria	Clinical examination	Ultrasound	Patient characteristics/history	Radiography
Raw data	0.33 (0.24, 0.45)	0.50 (0.33, 0.73)	0.26 (0.14, 0.52)	0.29 (0.16, 0.58)	0.34 (0.12, 0.95)
Ranks	0.32 (0.24, 0.45)	0.49 (0.33, 0.73)	0.26 (0.15, 0.53)	0.28 (0.15, 0.56)	0.31 (0.10, 0.95)
Top 10	0.31 (0.31, 0.31)	0.39 (0.39, 0.40)	0.23 (0.22, 0.23)	0.17 (0.16, 0.18)	0.12 (0.11, 0.13)

EPOS = European Paediatric Orthopaedic Society; ICC = intraclass correlation coefficient.

acceptable consistency; and 0.75 or more, good consistency [22]. Some authors believe lower coefficients are acceptable in the context of research [10, 13], whereas coefficients for judgments on individual patients should reach values of 0.70 to 0.80 [2]. The mean VAS scores of regions were tested for equal variances using the Bartlett test [1]. We compared ICC estimates of regions with use of their 95% CIs, whereas overlapping intervals are considered nonsignificant. To compare the relevance of ultrasonography and clinical examination in the diagnosis of DDH by region, we calculated for each region the difference in ICC estimates of the domains ultrasonography and clinical examination. We calculated Cronbach's α as a measure of stability of the data. A high Cronbach's α in this context suggests another sample of surgeons selected in the same way from the same population would give similar results [6]. The primary analysis was based on the data obtained from members of EPOS, whereas the data

obtained from members of BSCOS were analyzed separately for validation purposes. Our assumption was, if Cronbach's α was high in the EPOS sample, then ICC estimates obtained in the BSCOS sample should be similar.

Results

There was poor consistency among the respondents in rating the 37 items. The ICC for all 37 items was 0.33 (95% CI, 0.24–0.45). Items under the clinical examination domain gave an ICC of 0.50 (0.33–0.73). For the domain patient characteristics/history the ICC was 0.29 (0.16–0.58). For the domain ultrasonography it was 0.26 (0.14–0.52) and it was 0.34 (0.12–0.95) for the domain radiography. The ICC decreased in the binary rating scale with coefficients less than 0.20 (Table 2). The high variability in VAS scores assigned to each item was evident (Fig. 1). In

Fig. 1 A graph shows the variability of the data for each of the 37 items (x-axis) based on the VAS ratings (y-axis) provided by all surgeons. The horizontal line represents the median value. The ends of the box fall at the upper and lower quartiles (the middle 50% of the VAS ratings fall within the range of scores defined by the box). Upper and lower quartiles are indicated by the vertical lines. Outliers are

indicated by a circle, extreme outliers by a star. Each of the 37 items was given the lowest and the highest possible score. Considerable variability is seen for all items. Outliers were observed for items commonly regarded as highly relevant in the diagnosis of DDH (eg, Ortolani and Barlow tests), indicating variability in opinion.

Fig. 2 A graph shows the geographic variation of consistency based on the ICC.

particular, for items 2, 3, 6, 13, and 20 extreme outliers were observed.

A range of ICC values was observed in the geographic regions studied (Fig. 2). The differences in ICC estimates among the 15 regions were as much as 0.30 (Fig. 3); however, the variances in raw VAS scores among the regions did not differ significantly ($p = 0.48$) and the CIs of regional ICCs overlapped with one exception (Fig. 3). In seven regions (46%), consistency in rating the 37 criteria was acceptable with an ICC of 0.40 or greater.

Surgeons of the same region were most consistent regarding clinical examination criteria (Table 3). Differences in ICC estimates greater than 0.20 were found between ratings of clinical examination criteria and ultrasonic criteria in eight regions (53%) (Table 4). In contrast, in five regions (33%), surgeons in each individual region rated clinical criteria and ultrasonic criteria consistently with differences in ICC less than 0.10. The regions with the

least discrepancy in opinion regarding ultrasonography and clinical examination criteria (ICC differences < 0.10) and with an acceptable overall agreement in those domains (ICC > 0.40) were Spain, Portugal, and Turkey. A high variation within regions was found for the domain patient characteristics/history, with 90% of the regions showing an ICC less than 0.40. Similar results were found for the domain radiography, with an ICC less than 0.40 in 12 regions (80%).

The 37 items investigated here showed good internal consistency. Cronbach's α for all 37 criteria was 0.88, indicating another group of surgeons from the same population would report similar results. Cronbach's α was 0.77 for the domain clinical examination, 0.84 for the domain ultrasonography, 0.86 for the domain patient characteristics/history, and 0.75 for the domain radiography.

Similar consistency estimates were seen among members of BSCOS. For the BSCOS-wide survey, the ICCs

Fig. 3 A graph shows the ICCs for the 37 criteria for the EPOS-wide survey and the 15 regions. The horizontal lines represent 95% CIs. Italy had significantly better consistency ($p < 0.05$) compared with the Slovak and Czech Republics.

were 0.39 (95% CI, 0.29–0.52) for all 37 criteria, 0.52 (95% CI, 0.35–0.75) for the domain clinical examination, 0.25 (95% CI, 0.14–0.52) for the domain ultrasonography, 0.39 (95% CI, 0.23–0.69) for the domain patient characteristics/history, and 0.37 (95% CI, 0.10–0.95) for the domain radiography.

Discussion

While it is known there is a difference in opinion regarding the management of DDH in early infancy based on geography, to our knowledge surgeons' views on the importance of criteria used for the diagnosis of DDH have not been published before. Consistency in the diagnosis of DDH is important to ensure appropriate treatment and to reduce variations in standards of care. In this study, we determined the consistency with which pediatric orthopaedic surgeons rated the importance of diagnostic criteria for DDH in infants not older than 8 weeks. Our study has important implications for clinicians in that the established consistency estimates enable them to quantify the uncertainty associated with a set of diagnostic criteria for DDH. Consistency was considerably lower than expected, considering the frequency with which DDH is diagnosed by

the participants. We suggest the low consistency estimates mean specialist surgeons do not have the same concept of what constitutes DDH in young infancy [25].

A potential limitation of this study is that it relied on opinions expressed in a survey. However, surveys are an effective means of evaluating physicians' attitudes [20] and physicians do act as they indicate in surveys [9]. Strong concordance has been shown between the ratings of paper patients and actual patients [3]. We thus suggest our results provide an acceptable indicator of actual diagnostic practices of the participants in that the importance, and perhaps the use, of diagnostic criteria for DDH differ substantially. Owing to the limited number of participants from some of the countries studied, we were unable to derive consistency statistics for every single country and defined 15 regions. In rating the 37 criteria, we chose an explicated approach where each criterion was rated in isolation. However, in clinical practice, patients usually present with more than one diagnostic feature and these can occur in several different combinations. Since we did not aim to determine how surgeons make the diagnosis of DDH, an explicated approach seemed more appropriate for this research. In brief, participants of a survey provide their evaluations component by component and the inferences about the whole product (in this study, the consistency of the 37

Table 3. Consistency for members of EPOS based on region and diagnostic domains

Region*	Number of members	ICC					
		All 37 criteria	Clinical examination	Ultrasound	Patient characteristics/history	Radiography	All domains excluding radiography
EPOS members	156	0.33	0.50	0.26	0.29	0.34	0.33
European countries [†]	105	0.35	0.53	0.30	0.28	0.30	0.36
Italy	13	0.55	0.68	0.43	0.30	0.26	0.57
France	15	0.47	0.56	0.45	0.40	0.39	0.47
Norway (1), Denmark (2), Sweden (6), Finland (2)	11	0.46	0.67	0.46	0.23	0.39	0.49
Israel	11	0.45	0.62	0.28	0.36	0.41	0.46
Japan (2), Taiwan (1), South Korea (1)	4	0.43	0.42	0.66	0.41	0.77	0.42
Poland	8	0.40	0.59	0.23	0.25	0.28	0.41
United Kingdom	10	0.40	0.51	0.28	0.32	0.19	0.36
Spain (8), Portugal (1)	9	0.37	0.51	0.44	0.27	0.45	0.39
USA (18), Canada (1), Argentina (1)	20	0.35	0.55	0.24	0.35	0.33	0.36
Turkey	7	0.33	0.47	0.42	0.31	0.33	0.34
Slovenia (2), Croatia (2), Macedonia (1), Serbia (1), Kosovo (1), Armenia (1), Greece (1)	9	0.32	0.47	0.19	0.22	0.34	0.33
Germany (5), Austria (4), Switzerland (8)	17	0.27	0.37	0.29	0.29	0.21	0.26
Egypt (6), United Arab Emirates (1), Palestine (1)	8	0.27	0.31	0.34	0.27	0.39	0.26
Netherlands (5), Belgium (5)	10	0.26	0.42	0.27	0.17	0.04	0.26
Slovak Republic (2), Czech Republic (2)	4	0.18	0.40	0.36	0.06	0.39	0.16

* Numbers in parentheses denote the number of respondents of respective countries; [†]excluded are Japan, Taiwan, South Korea, USA, Canada, Argentina, Egypt, United Arab Emirates, Palestine, Armenia, Israel, Turkey; EPOS = European Paediatric Orthopaedic Society; ICC = intraclass correlation coefficient.

items) are made from these single evaluations. Explicated methods have been used in health research in the assessment of aspects of respondents' preferences and are particularly appropriate for assessing many attributes, such as 37 different diagnostic criteria [4]. The 37 criteria used in this study were stated the most important in the diagnosis of DDH by an international consensus [23]. Therefore, specialist surgeons should be fairly consistent in rating the importance of each, regardless of context information. In studying reliability, it is perfectly legitimate that researchers determine the impact of different factors in isolation [16]. For such settings, it is essential that all raters independently complete the evaluations in similar test settings [16], which was done in the current study. It is possible that adding context information would have resulted in better consistency estimates but this remains speculative. Determining how pediatric orthopaedic surgeons make the diagnosis of DDH using these 37 criteria warrants a holistic approach, whereby several diagnostic factors are assessed together as they can occur in a patient; we are currently conducting such research.

This study showed generally poor consistency among pediatric orthopaedic specialists in assigning importance to

diagnostic criteria for DDH in young infancy. Consistency was best among criteria related to clinical examination, but none of the consistency estimates reached the recommended standard of 0.70 to 0.80 [2]. The magnitude of regional variation likewise was high, with differences in consistency estimates as large as 0.30 between regions (Fig. 2). Surgeons in each region appeared to have a concept of the diagnosis of DDH that differed from those used by specialists of other regions. Most consistent were surgeons from Italy, France, and Scandinavia, whose consistency estimates exceeded 0.45, suggesting surgeons of these regions agreed on what they consider important diagnostic criteria. Our study failed to show that regional variation in rating the 37 criteria was statistically significant with the exception of Italy, which showed better consistency ($p < 0.05$) than Czech and Slovak Republics (Fig. 2).

The strengths of this study include the involvement of the most appropriate group of diagnosticians, high response rate, and multinational setting. Because our study focused on diagnostic criteria (as opposed to referral criteria), we included only pediatric orthopaedic surgeons who treat DDH. Our results therefore are applicable to such specialists; these might not generalize to nonspecialists, and, because of

Table 4. Differences in consistency between the domains clinical examination and ultrasound for different regions

Region*	Number of members	ICC		
		Clinical examination	Ultrasound	Difference
Poland	8	0.59	0.23	0.36
Israel	11	0.62	0.28	0.34
USA (18), Canada (1), Argentina (1)	20	0.55	0.24	0.31
Slovenia (2), Croatia (2), Macedonia (1), Serbia (1), Kosovo (1), Armenia (1), Greece (1)	9	0.47	0.19	0.28
Italy	13	0.68	0.43	0.25
Japan (2), Taiwan (1), South Korea (1)	4	0.42	0.66	−0.24
UK	10	0.51	0.28	0.23
Norway (1), Denmark (2), Sweden (6), Finland (2)	11	0.67	0.46	0.21
Netherlands (5), Belgium (5)	10	0.42	0.27	0.15
France	15	0.56	0.45	0.11
Germany (5), Austria (4), Switzerland (8)	17	0.37	0.29	0.08
Spain (8), Portugal (1)	9	0.51	0.44	0.07
Turkey	7	0.47	0.42	0.05
Slovak Republic (2), Czech Republic (2)	4	0.40	0.36	0.04
Egypt (6), United Arab Emirates (1), Palestine (1)	8	0.31	0.34	−0.03
EPOS members	156	0.50	0.26	0.24

* Numbers in parentheses denote the number of respondents of respective countries; ICC = intraclass correlation coefficient; EPOS = European Paediatric Orthopaedic Society.

the specialist orientation of the group, also might be considered to represent a best-case scenario. However, to test the effect of sampling on the consistency estimates derived in this study, we conducted a separate survey of all professionals of one region, ie, the UK. Consistency was identical in both surveys. We interpret this as evidence that the pattern of data seen in the EPOS-wide survey is representative of all pediatric orthopaedic surgeons in the respective regions, which is supported by a high Cronbach's α . Another strength of this study is that we investigated 37 criteria that 200 pediatric orthopaedic surgeons from 35 countries rated the most important features indicative of DDH in early infancy [23]. It is therefore unlikely our study omitted any important diagnostic features.

To our knowledge, our study is the first to investigate physician attitude, geography, and training backgrounds of diagnosticians as a potential explanation for the wide range of reported prevalence estimates of DDH. The variability identified in our study is in agreement with other studies that showed widely ranging prevalence estimates from 20 to 60 per 1000 [15, 17, 18, 26, 30]. Our study suggests that the differences in prevalence estimates can be attributed, at least in part, to inconsistencies in diagnostic practices.

Of all criteria studied, the participants were most inconsistent about the ultrasonographic criteria; this may be related to the inconsistent evidence in terms of the use

of this diagnostic modality [24, 32]. Surgeons in the same region showed considerably stronger agreement on clinical than on ultrasonographic criteria (Table 4). Mean differences in ICC estimates greater than 0.20 were observed in eight regions (53%), which we interpret as further evidence that the role of ultrasonographic criteria, even within the same region, is controversial. The results of this study emphasize the need to align diagnostic practices by reference to a feasible set of widely accepted diagnostic criteria. Our study was not designed to determine how accurately pediatric orthopaedic surgeons diagnose DDH using the 37 criteria. This question requires a different methodologic approach and is research in progress. Further research to improve the diagnosis of DDH in early infancy is needed and should focus particularly on the role of ultrasonography and risk factors. Reducing variations in diagnostic practices should help improve the widely differing standards in the management of DDH. Standardized, widely accepted diagnostic criteria that are based on the same concept of DDH need to be formulated and followed.

Acknowledgment We thank James G. Wright MD, MPH, FRCS, for valuable comments on a previous version of this manuscript. We also thank all members of EPOS and BSCOS who participated in this study, in particular Fritz Hefti MD, previous Chair of EPOS Scientific Committee, and James Hunter FRCS, former president of BSCOS, who were both essential in promoting this study.

References

1. Bartlett MS. Properties of sufficiency and statistical tests. *Proc R Soc Lond A*. 1937;160:268–282.
2. Bland JM, Altman DG. Cronbach's alpha. *BMJ*. 1997;314:572.
3. Bouma BJ, van der Meulen JH, van den Brink RB, Smidts A, Cheriex EC, Hamer HP, Arnold AE, Zwinderman AH, Lie KI, Tijssen JG. Validity of conjoint analysis to study clinical decision making in elderly patients with aortic stenosis. *J Clin Epidemiol*. 2004;57:815–823.
4. Bridges JF, Joy SM, Gallego G, Kudo M, Ye SL, Han KH, Cheng AL, Blauvelt BM. Needs for hepatocellular carcinoma control policy in the Asia-Pacific region. *Asian Pac J Cancer Prev*. 2011;12:2585–2591.
5. Clinical practice guideline: early detection of developmental dysplasia of the hip. Committee on Quality Improvement, Subcommittee on Developmental Dysplasia of the Hip. American Academy of Pediatrics. *Pediatrics*. 2000;105:896–905.
6. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika*. 1951;16:297–334.
7. Dillman DA. *Mail and Internet Surveys: The Tailored Design Method*. 2nd ed. New York, NY: Wiley; 2000.
8. Eisenberg JM. Physician utilization: the state of research about physicians' practice patterns. *Med Care*. 1985;23:461–483.
9. Epstein AM, McNeil BJ. Relationship beliefs and behavior in test ordering. *Am J Med*. 1986;80:865–870.
10. Feinstein A. *Clinimetrics*. New Haven, CT: Yale University Press; 1987.
11. Franks P, Williams GC, Zwanziger J, Mooney C, Sorbero M. Why do physicians vary so widely in their referral rates? *J Gen Intern Med*. 2000;15:163–168.
12. Gardiner HM, Dunn PM. Controlled trial of immediate splinting versus ultrasonographic surveillance in congenitally dislocatable hips. *Lancet*. 1990;336:1553–1556.
13. Graham B, Dvali L, Regehr G, Wright JG. Variations in diagnostic criteria for carpal tunnel syndrome among Ontario specialists. *Am J Ind Med*. 2006;49:8–13.
14. Graham B, Regehr G, Wright JG. Delphi as a method to establish consensus for diagnostic criteria. *J Clin Epidemiol*. 2003;56:1150–1156.
15. Holen KJ, Tegnander A, Bredland T, Johansen OJ, Saether OD, Eik-Nes SH, Terjesen T. Universal or selective screening of the neonatal hip using ultrasound? A prospective, randomised trial of 15,529 newborn infants. *J Bone Joint Surg Br*. 2002;84:886–890.
16. Karanickolas PJ, Bhandari M, Kreder H, Moroni A, Richardson M, Walter SD, Norman GR, Guyatt GH. Evaluating agreement: conducting a reliability study. *J Bone Joint Surg Am*. 2009;91:99–106.
17. Leck I. Congenital dislocation of the hip. In: Wald N, Leck I, eds. *Antenatal and Neonatal Screening*. 2nd ed. Oxford, UK: Oxford University Press; 2000:398–424.
18. Marks DS, Clegg J, al-Chalabi AN. Routine ultrasound screening for neonatal hip instability: can it abolish late-presenting congenital dislocation of the hip? *J Bone Joint Surg Br*. 1994;76:534–538.
19. McNeil BJ. Shattuck Lecture: hidden barriers to improvement in the quality of care. *N Engl J Med*. 2001;345:1612–1620.
20. Moskowitz AJ, Kuipers BJ, Kassirer JP. Dealing with uncertainty, risks, and tradeoffs in clinical decisions: a cognitive science approach. *Ann Intern Med*. 1988;108:435–449.
21. Patel H. Preventive health care, 2001 update: screening and management of developmental dysplasia of the hip in newborns. *CMAJ*. 2001;164:1669–1677.
22. Portney LG, Watkins MP. Statistical measures of reliability. *Foundations of Clinical Research: Applications in Practice*. 2nd ed. Upper Saddle River, NJ: Princeton-Hall; 2000.
23. Roposch A, Liu LQ, Hefti F, Clarke NM, Wedge JH. Standardized diagnostic criteria for developmental dysplasia of the hip in early infancy. *Clin Orthop Relat Res*. 2011;469:3451–3461.
24. Roposch A, Moreau NM, Uleryk E, Doria AS. Developmental dysplasia of the hip: quality of reporting of diagnostic accuracy for US. *Radiology*. 2006;241:854–860.
25. Roposch A, Wright JG. Increased diagnostic information and understanding disease: uncertainty in the diagnosis of developmental hip dysplasia. *Radiology*. 2007;242:355–359.
26. Rosendahl K, Markestad T, Lie RT. Ultrasound screening for developmental dysplasia of the hip in the neonate: the effect on treatment rate and prevalence of late cases. *Pediatrics*. 1994;94:47–52.
27. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86:420–428.
28. Tegnander A, Holen KJ, Terjesen T. The natural history of hip abnormalities detected by ultrasound in clinically normal newborns: a 6–8 year radiographic follow-up study of 93 children. *Acta Orthop Scand*. 1999;70:335–337.
29. Terjesen T, Bredland T, Berg V. Ultrasound for hip assessment in the newborn. *J Bone Joint Surg Br*. 1989;71:767–773.
30. von Kries R, Ihme N, Oberle D, Lorani A, Stark R, Altenhofen L, Niethard FU. Effect of ultrasound screening on the rate of first operative procedures for developmental hip dysplasia in Germany. *Lancet*. 2003;362:1883–1887.
31. Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. *Stat Med*. 1998;17:101–110.
32. Woolacott NF, Puhan MA, Steurer J, Kleijnen J. Ultrasonography in screening for developmental dysplasia of the hip in newborns: systematic review. *BMJ*. 2005;330:1413.