Journal of Experimental and Clinical Medicine
https://dergipark.org.tr/omujecm

Research Article

J Exp Clin Med
2022; 39(2): 463-466
doi: 10.52142/omujecm.39.2.30

Comparison of the interobserver reliability of ultrasonography and radiography in diagnosis of developmental dysplasia of the hip

Fatih İlker CAN[1,*], Sacit TURANLI[2], Hakan ATALAR[2]

[1]Orthopedics and Traumatology Clinic, Muğla Training and Research Hospital, Muğla, Turkey
[2]Department of Orthopedics and Traumatology, Faculty of Medicine, Gazi University, Ankara, Turkey

Received: 18.11.2021 • **Accepted/Published Online:** 17.01.2022 • **Final Version:** 18.03.2022

Abstract

We aimed to compare the interobserver reliability of ultrasonography (USG) and x-ray (XR) and to calculate the sensitivity and specificity of these tests in the diagnosis of developmental dysplasia of the hip (DDH). This retrospective study was conducted among 150 USG and 300 XR images of infants examined for DDH between January 2013 and June 2015. The sonographic angle measurements and hip classifications of each USG and XR were carried out by five orthopedic surgeons specialized in pediatric orthopedics. Both USG and XR showed almost perfect agreement between five observers (κUSG=0.936, κXR=0.927, respectively, p<.0001). In patients under the age of 6 months, the interobserver reliability was almost perfect for USG and substantial for XR (κUSG=0.957, κXR=0.809, respectively, p<.0001). In patients older than 6 months, although interobserver agreement of both tests were almost perfect (κ>0.81), XR showed slightly higher agreement than USG (κXR=0.912, κUSG=0.885, respectively, p<.0001). In the diagnosis of DDH, both USG and plain x-rays are effective radiological tests because they offer high interobserver reliability. However, the surgeon's practical training and experience significantly affect the reliability in the evaluation of pediatric hip USG and XR.

Keywords: Developmental dysplasia of the hip, ultrasonography, plain x-ray, interobserver reliability

1. Introduction

Developmental dysplasia of the hip (DDH) contains a broad spectrum of aberrant development of the acetabulum and proximal femur (1, 2). The incidence of DDH in routine screening has been reported as 5-30/1000 (3). Treatment in the first months of the infant is simpler and has a better prognosis, hence early diagnosis and treatment is critical (4). As the infant gets older, the potential for harmonious development of the hip decreases and reduction becomes more challenging. Therefore, the assessment of plain X-ray (XR) and ultrasonography (USG) and the staging discrepancies depending on these assessments may cause critical varieties in diagnosis and treatment. High consistency is crucial among the surgeons amid the diagnosis.

In the sonographic examination defined by Graf, the relationship between the femoral head and the acetabulum is evaluated using angular measurements of pediatric hip (5). XR is also frequently used in diagnosis, however less preferred in patients younger than 6 months due to ionizing radiation and the difficulty of imaging the non-ossified cartilage structures of the hip (6). Since XR findings may be affected by the position of the pelvis, different measurements may occur among surgeons (7). Due to the varieties in evaluation encountered in both USG and XR, inconsistencies may occur between measurements, and standardization in staging and treatment may deteriorate. While many orthopedic disorders can be diagnosed using physical examinations rather than precise radiological measurements, the accuracy of the radiological assessment is much more decisive than the physical examination findings in DDH (8, 9).

Although there are studies evaluating the efficacy of USG in the diagnosis and treatment of DDH in the literature (10, 11, 12, 13), no study has been observed comparing the interobserver reliability between USG and XR in terms of measurement and DDH staging criteria. Therefore, we aimed to compare the interobserver reliability of USG and XR imaging techniques and to calculate sensitivity and specificity of these tests in the diagnosis of DDH.

2. Materials and methods

In this study, 150 USG and 300 XR images of infant patients (mean age; USG=4 months, XR=7 months) were evaluated for DDH screening between January 2013 and June 2015. The study was approved by the local university ethics committee for clinical trials and conducted in accordance with the principles of the Declaration of Helsinki. Printed copies of all 450 images were sent to five orthopedic surgeons who received specific training on pediatric hip evaluation and DDH. The measurements and classifications were carried out by the observers retrospectively. All observers were blinded to the

*Correspondence: dr.fatihcan07@gmail.com

evaluation results of the other.

All USG imaging were performed by the same surgeon who has a pediatric sonography practice certificate. XR imaging consisted of plain radiographs taken in the radiology department. Inclusion criteria were established according to the quality and suitability of the images for evaluation. Suitability for USG was determined according to standard plane criteria, which consisted of a straight iliac wing line, a clear visualization of the acetabular labrum and a complete visualization of the transition point from the ilium to the triradiate cartilage (14). For XR, these criteria were; entirety of the bony pelvis from superior of the iliac crest to the proximal shaft of the femur, symmetrical view of obturator foramen, equal concavity of the iliac wings and greater trochanters of the proximal femur (15).

Observers were asked to make the measurements in each XR and USG and to classify the hip according to the evaluation criteria. In XR, these criteria were; acetabular index measurement, continuity of Shenton-Menard line and the location of the hip in four quadrants formed according to Hilgenreiner's and Perkin's lines. In USG, alpha and beta angle measurements according to Graf method and determination of hip type according to Graf classification were requested from the observers. After the data were compiled, interobserver reliability analysis was performed for both USG and XR, regardless of age and age dependent groups among the patients. In addition, the sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) of both USG and XR were calculated separately for each observer. Any superiority between the imaging tests among the groups in terms of interobserver reliability was addressed.

2.1 Statistical Analysis

Statistical analysis was performed using the SPSS, version 22.0 (SPSS Inc., Chicago, IL, USA) Fleiss kappa test for interobserver reliability. Reliability and consistency evaluation was classified according to the widely used Landis and Koch interpretation of kappa's values ($\kappa \geq 0.81$ equals almost perfect, κ=0.61-0.8 as substantial, κ=0.41-0.6 as moderate, κ=0.21-0.4 as fair, and $\kappa \leq 0.2$ as slight correlation). Assessment of statistical differences between kappa values was calculated with 95% confidence interval (CI) and $p<0.05$ was accepted as statistically significant.

3. Results

Both USG and XR showed almost perfect agreement between five observers when the radiological tests were evaluated regardless of age (κUSG=0.936, κXR=0.927, respectively, $p<.0001$).

When patients were grouped as below and under the age of 6 months, it was observed that in patients under the age of 6 months, the interobserver reliability was almost perfect for USG and substantial for XR (κUSG=0.957, κXR=0.809, respectively, $p<.0001$). In patients older than 6 months, although interobserver agreement of both tests were almost perfect (κ>0.81), XR showed slightly higher agreement than USG (κXR=0.912, κUSG=0.885, respectively, $p<.0001$).

Table 1. Kappa coefficients for interobserver reliability of ultrasonography (USG) and plain X-ray (XR) between five observers

Kappa coefficients for interobserver reliability		
	USG	**XR**
< 6 months of age	0.957	0.809
> 6 months of age	0.885	0.912
All ages	0.936	0.927

Overall, although interobserver reliability of both USG and XR showed statistically significant reliability (Table 1), no statistically significant difference was found between USG and XR in terms of interobserver reliability, based on overlapping 95% confidence intervals.

Table 2. Ultrasonography (USG) consistency calculations between five observers

USG consistency tests for all patients				
	Sensitivity (%)	**Specificity (%)**	**Positive Predictive Value (%)**	**Negative Predictive Value (%)**
1. observer	66.6	53.3	63.1	57.1
2. observer	77.7	46.6	63.6	63.6
3. observer	76	50	82.6	40
4. observer	82.6	60	82.6	60
5. observer	79.1	55.5	82.6	50

According to the calculations made for each observer separately, the sensitivity range of USG between observers was 66.6%-82.6%, the specificity range was between 46.6%-60%, the PPV range was between 63.1%-82.6%, and finally the NPV range was between 40%- 60% (Table 2). On the other hand, the sensitivity of XR among observers was between 71.2%-82.9%, specificity was between 43.7%-66.6%, PPV range was between 78%-88.8%, and NPV range was between 33.3%-70.8% (Table 3).

Table 3. Plain X-ray (XR) consistency calculations between five observers

XR consistency tests for all patients				
	Sensitivity (%)	**Specificity (%)**	**Positive Predictive Value (%)**	**Negative Predictive Value (%)**
1. observer	69.8	56.7	73.3	52.7
2. observer	81.3	58.5	70.5	68.5
3. observer	73.6	45.8	81.1	35.4
4. observer	81.1	64.5	83.5	60.6
5. observer	76.6	60.8	86.7	40

4. Discussion

While many orthopedic disorders are commonly diagnosed based on physical examination findings, radiological evaluation and measurements are more decisive in diagnosis of DDH (8, 9). The reduction of the femoral head in acetabulum

is ascertained according to the USG and XR findings and the treatment strategy is determined accordingly. There are a few studies reporting the reliability of USG for the diagnosis of DDH in the literature (5, 16-19) yet sonographic assessments are well-known to be operator-dependent. This means that the quality of the obtained images and their accurate interpretation depend on the experience and knowledge of the sonographer (21). Therefore, the subjective nature of this test may cause different measurements among the surgeons during both sonographic assessment and evaluation (21) and may lead to divergent classifications for diagnosis and treatment of DDH which may raise questions about standardization (10, 22-25). The fact that the sonographies we used in our study were applied by a single person, significantly reduces the possibility of this subjective application. Kolb et al. stated that the reliability of hip sonography with the Graf method is increasing gradually, but different measurements originating from the transducer cannot be avoided, and they reported that transducer inclination creates measurement differences that affect clinical results (28). These studies led to the questioning of the reliability of USG in the diagnosis DDH and the necessity for interobserver reliability studies. Dias et al. investigated the reliability of 62 USG sections among five observers and reported that they observed moderate agreement on alpha angle measurement and poor agreement on beta angle measurement. In their study, the authors generally reported poor interobserver reliability of USG in the diagnosis of DDH (10). When the studies conducted after the 2000s were investigated, it is noticed that the interobserver reliability has begun to increase parallel to the improving experience on USG over the years. Çopuroğlu et al. investigated the reliability of USG among seven observers of 33 pediatric patients and reported high interobserver reliability of alpha angle measurement (22). The authors stated that different classifications can be made from the same USG sections due to different alpha angle measurements, and they also observed many different measurements of the same USG in their very own study. They also claimed that the best results were obtained when USG and plain radiography were evaluated together (22).

In the study of Orak et al. interobserver reliability was investigated among four different observers on 50 infants and meaningful differences were reported between the observers. In this study, the authors reported quite wide range in terms of reliability; 3.6%-44.5% agreement range in alpha angle measurement and 0.9%- 45.3% agreement range in beta angle measurement were reported (23). Likewise, it has also been reported that the variations of positioning the infant and the XR device may cause diversity especially among the distinct experience levels of physicians evaluating the x-rays (26). Ismiarto et al. analyzed the interobserver reliability between junior and senior orthopedic residents for XR using Fleiss kappa test. The kappa value of Tönnis classification among seniors and juniors were 0.715 and 0.577, respectively. The authors claimed that the difference was due to the fact that juniors have less experience than seniors (24). Compared to this study, our study showed higher reliability for XR (κXR=0.927) which was probably due to specific training and over 10 years of experience of our observers in pediatric hip ultrasonography. Singh et al. studied the interobserver reliability of XR and reported a very high reliability with a 0.935 kappa value. In our study, we also obtained high kappa value for XR (κXR=0.927) (27).

Although limited number of studies investigating interobserver reliability for both USG and XR are observed in the literature, there is only one study comparing the reliability of these tests in the diagnosis of DDH. This study was conducted in 1990, in which Terjesen et al. examined 312 pediatric hips with USG and XR, it was stated that the authors could make the same diagnosis in 303 of 312 pediatric hips examined. In this study, the authors made a general comparison of the adequacy of diagnosis rather than investigating the reliability of specific measurements (20). In our study, we compared the reliability of USG and XR between five observers using a large series of USG and XR with a total number of 150 USG and 300 XR images, and the observer results were classified as normal, acetabular dysplasia, subluxation and dislocation according to the measurements and the consistency of the diagnosis of DDH was investigated. The interobserver reliability in our study was higher compared to the literature. The possible reasons of this superiority were that all our observers were active performers of pediatric orthopedics and had been performing and evaluating hip sonography according to the Graf method over 10 years. In addition, it is likely that all sonographs were performed by a single surgeon which improves the standardization of the imaging.

In our study, it was observed that the interobserver reliability of USG and XR was high in both infants younger than 6 months and older than 6 months. Although there was no statistically significant difference, we observed that the interobserver reliability of USG was higher in infants younger than 6 months according to Fleiss kappa values, and the interobserver reliability of XR was slightly higher in children older than 6 months. This may be based on that the bony structures become more prominent on direct radiographs around 6 months depending on the ossification mechanism. X-ray assessments may be more consistent after 6 months due to the ossified structures. In our daily routine practices in our clinic, we effectively diagnose and treat developmental hip dysplasia by using both USG and XR examinations together in infants. We prefer USG more frequently, especially in children younger than 6 months, which offers high interobserver reliability.

In the diagnosis of DDH, both USG and plain X-ray imaging techniques are effective methods because they offer high interobserver reliability. However, the surgeon's

experience significantly affects the reliability in the evaluation of pediatric hip USG and XR.

Acknowledgements

None

Declaration of conflicting interests

The author declares that there is no conflict of interest.

References

1. Dezateux C, Rosendahl K. Developmental dysplasia of the hip. Lancet (London, England). 2007;369(9572):1541-1552.
2. Lee MC, Eberson CP. Growth and development of the child's hip. The Orthopedic clinics of North America. 2006;37(2):119-132, v.
3. Sewell MD, Rosendahl K, Eastwood DM. Developmental dysplasia of the hip. BMJ (Clinical research ed). 2009; 339:b4454.
4. Moraleda L, Albiñana J, Salcedo M, Gonzalez-Moran G. [Dysplasia in the development of the hip]. Revista espanola de cirugia ortopedica y traumatologia. 2013;57(1):67-77.
5. Graf R. The diagnosis of congenital hip-joint dislocation by the ultrasonic Combound treatment. Archives of orthopaedic and trauma surgery. 1980;97(2):117-33.
6. Milligan DJ, Cosgrove AP. Monitoring of a hip surveillance programme protects infants from radiation and surgical intervention. The bone & joint journal. 2020;102-b(4):495-500.
7. Omeroğlu H, Ozçelik A, Inan U, Seber S. Assessment of the correlation between commonly used radiographic parameters in normal, subluxated and dislocated hips. Journal of pediatric orthopedics Part B. 2006;15(3):172-177.
8. Finne PH, Dalen I, Ikonomou N, Ulimoen G, Hansen TW. Diagnosis of congenital hip dysplasia in the newborn. Acta orthopaedica. 2008;79(3):313-320.
9. Sulaiman A, Yusof Z, Munajat I, Lee N, Zaki N. Developmental dysplasia of hip screening using ortolani and barlow testing on breech delivered neonates. Malaysian orthopaedic journal. 2011;5(3):13-16.
10. Dias JJ, Thomas IH, Lamont AC, Mody BS, Thompson JR. The reliability of ultrasonographic assessment of neonatal hips. The Journal of bone and joint surgery British volume. 1993;75(3):479-482.
11. Omeroğlu H, Biçimoğlu A, Koparal S, Seber S. Assessment of variations in the measurement of hip ultrasonography by the Graf method in developmental dysplasia of the hip. Journal of pediatric orthopedics Part B. 2001;10(2):89-95.
12. Roposch A, Graf R, Wright JG. Determining the reliability of the Graf classification for hip dysplasia. Clinical orthopaedics and related research. 2006; 447:119-124.
13. Simon EA, Saur F, Buerge M, Glaab R, Roos M, Kohler G. Inter-observer agreement of ultrasonographic measurement of alpha and beta angles and the final type classification based on the Graf method. Swiss medical weekly. 2004;134(45-46):671-677.
14. AIUM practice guideline for the performance of an ultrasound examination for detection and assessment of developmental dysplasia of the hip. Journal of ultrasound in medicine: official journal of the American Institute of Ultrasound in Medicine. 2009; 28(1):114-119.
15. John Lampignano LEK. Bontrager's Textbook of Radiographic Positioning and Related Anatomy. 2017
16. Berman L, Klenerman L. Ultrasound screening for hip abnormalities: preliminary findings in 1001 neonates. British medical journal (Clinical research ed). 1986;293(6549):719-722.
17. Clarke NM, Clegg J, Al-Chalabi AN. Ultrasound screening of hips at risk for CDH. Failure to reduce the incidence of late cases. The Journal of bone and joint surgery British volume. 1989; 71(1):9-12.
18. Harcke HT, Kumar SJ. The role of ultrasound in the diagnosis and management of congenital dislocation and dysplasia of the hip. The Journal of bone and joint surgery American volume. 1991;73(4):622-628
19. Terjesen T, Bredland T, Berg V. Ultrasound for hip assessment in the newborn. The Journal of bone and joint surgery British volume. 1989;71(5):767-773.
20. Terjesen T, Rundén TO, Tangerud A. Ultrasonography and radiography of the hip in infants. Acta orthopaedica Scandinavica. 1989;60(6):651-660.
21. Sadeghian M, Zarabi V, Noorbakhsh S, Taherinia L, Jafarv M, Ashouri S. Diagnostic value for imaging studies (radiography, ultrasound and CT scan) in pediatric appendicitis in compare with adults. Emerg Med. 2019;2:1-4
22. Copuroglu C, Ozcan M, Aykac B, Tuncer B, Saridogan K. Reliability of ultrasonographic measurements in suspected patients of developmental dysplasia of the hip and correlation with the acetabular index. Indian journal of orthopaedics. 2011;45(6):553-7.
23. Orak MM, Onay T, Çağırmaz T, Elibol C, Elibol FD, Centel T. The reliability of ultrasonography in developmental dysplasia of the hip: How reliable is it in different hands? Indian journal of orthopaedics. 2015;49(6):610-614.
24. Ismiarto YD, Agradi P, Helmi ZN. Comparison of Interobserver Reliability between Junior and Senior Resident in Assessment of Developmental Dysplasia of The Hip Severity using Tonnis and International Hip Dysplasia Institute Radiological Classification. Malaysian orthopaedic journal. 2019;13(3):60-65.
25. Mostofi E, Chahal B, Zonoobi D, Hareendranathan A, Roshandeh KP, Dulai SK, et al. Reliability of 2D and 3D ultrasound for infant hip dysplasia in the hands of novice users. European radiology. 2019;29(3):1489-1495.
26. Bankaoğlu M. Three-dimensional Computerized Tomography and Multiplanar Imaging of Developmental Hip Dysplasia. Sisli Etfal Hastanesi tip bulteni. 2019;53(2):103-109.
27. Singh KA, Ganjwala D, Gupta P, Vazhayil Kottamttavida I, Varma M, Shah H. Reliability of 3 Radiologic Classifications for the Severity of the Developmental Dysplasia of the Hip in Children Older Than 4 Years. Journal of pediatric orthopedics. 2021.
28. Kolb A, Benca E, Willegger M, Puchner SE, Windhager R, Chiari C. Measurement considerations on examiner-dependent factors in the ultrasound assessment of developmental dysplasia of the hip. International orthopaedics. 2017;41(6):1245-1250