



Impact of scan quality on AI assessment of hip dysplasia ultrasound

Abhilash Rakkundeth Hareendranathan¹ · Baljot Chahal¹ · Siyavash Ghasseminia^{1,2} · Dornoosh Zonoobi² · Jacob L. Jaremko^{1,2}

Received: 30 October 2020 / Accepted: 15 January 2021 / Published online: 5 March 2021
© Società Italiana di Ultrasonologia in Medicina e Biologia (SIUMB) 2021

Abstract

Aims Early diagnosis of developmental dysplasia of the hip (DDH) using ultrasound (US) is safe, effective and inexpensive, but requires high-quality scans. The effect of scan quality on diagnostic accuracy is not well understood, especially as artificial intelligence (AI) begins to automate such diagnosis. In this paper, we developed a 10-point scoring system for reporting DDH US scan quality, evaluated its inter-rater agreement and examined its effect on automated assessment by an AI system—MEDO-Hip.

Methods Scoring was based on iliac wing straightness and angulation; visibility of labrum, os ischium and femoral head; motion; and other artifacts. Four readers from novice to expert separately scored the quality of 107 scans with this 10-point scale and with holistic grading on a scale of 1–5. MEDO-Hip interpreted the same scans, providing a diagnostic category or identifying the scan as uninterpretable.

Results Inter-rater agreement for the 10-point scale was significantly higher than holistic scoring ICC 0.68 vs 0.93, $p < 0.05$. Inter-rater agreement on the categorisation of individual features, by Cohen's kappa, was highest for os ischium (0.67 ± 0.06), femoral head (0.65 ± 0.07) and iliac wing (0.49 ± 0.12) indices, and lower for the presence of labrum (0.21 ± 0.19). MEDO-Hip interpreted all images of a quality > 7 and flagged 13/107 as uninterpretable. These were low-quality images (3 ± 1.2 vs. 7 ± 1.8 in others, $p < 0.05$), with poor visualization of the os ischium and noticeable motion. AI accuracy in cases with quality scores ≤ 7 was 57% vs. 89% on other cases, $p < 0.01$.

Conclusion This study validates that our scoring system reliably characterises scan quality, and identifies cases likely to be misinterpreted by AI. This could lead to more accurate use of AI in DDH diagnosis by flagging low-quality scans likely to provide poor diagnosis up front.

Keywords Hip ultrasound · Scoring system · Interobserver study · 3D ultrasound · Artificial intelligence

Introduction

Developmental dysplasia of the hip (DDH) affects 1–3% of infants [1]. Undiagnosed DDH is thought to account for more than one-third of hip replacement surgeries in the < 60 age group [2]. The annual economic burden of the surgical treatment of DDH in the United States alone is estimated to be more than US\$1 billion². If diagnosed during infancy, DDH can be treated with simple soft-bracing (using a Pavlik Harness). The success rate of Pavlik Harness treatment

is very high ($> 90\%$) in infants under 7 weeks of age but decreases thereafter [3]. Despite the obvious advantages of early diagnosis [4, 5], universal screening for DDH at birth is not recommended or performed in most countries, due in part to the high variability in assessment [6].

Physical examinations for DDH via Barlow and Ortolani maneuvers have poor sensitivity beyond the neonatal period and miss cases of mild DDH [7–9]. Ultrasound is more sensitive to mild DDH, harmless and portable, thus being ideally suited to hip examination in infants for DDH diagnosis and treatment [10]. Currently, two-dimensional ultrasound (2DUS) is used for DDH diagnosis [11]. Scans are assessed based on Graf criteria, which measure the angle between the ilium and the acetabular roof (called the Alpha angle), as shown in Fig. 1. Alpha angles $> 60^\circ$ are considered normal and $< 43^\circ$ are considered severely

✉ Abhilash Rakkundeth Hareendranathan
hareendr@ualberta.ca

¹ Department of Radiology and Diagnostic Imaging,
University of Alberta, Alberta, Canada

² MEDO.Ai Inc, Singapore, Singapore

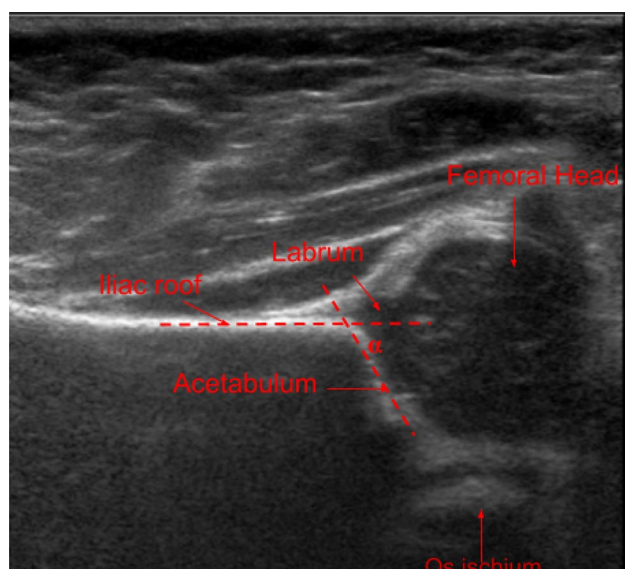


Fig. 1 Coronal Graf plane ultrasound scan of the hip with anatomical landmarks including the iliac roof, labrum, acetabulum, femoral head and os ischium. The alpha angle is measured as the angle between the iliac roof and the acetabulum as shown by the dashed line

dysplastic [12]. A fundamental limitation of this technique is that it requires an ultrasound scan in the Graf Standard Plane with all necessary landmarks, such as the ilium, acetabular roof, labrum and femoral head. Acquiring such high-quality scans requires long hours of expert training and experience. Consequently, scans acquired by novice sonographers are often inadequate, and could result in an incorrect diagnosis. An earlier study on infants and neonates suspected to have DDH showed that slight variations in probe position could result in incorrect diagnosis in ~50% of infants and ~67% of neonates [13].

Compared to 2DUS, 3D ultrasound is more reliable [14], as it visualises a larger anatomical area, which can account for possible variations in probe position. Earlier research has shown that using 3DUS, novice sonographers with minimal training are able to produce images of diagnostic accuracy equivalent to those produced by experts [15]. Semi-automatic interpretation of 3DUS scans reduces the proportion of incomplete scans and resultant need for follow-up by one-third [16]. Various semi-automated [17–19] and fully automated approaches [18, 20–22] have also been proposed for DDH assessment using 3DUS. However, these techniques require 3DUS volumes of adequate quality. Similar to 2DUS, whether scan quality is adequate in 3DUS is also assessed subjectively by the sonographer at the time of scanning. Recently, automatic deep-learning-based techniques for assessment of scan adequacy have also been evaluated on small data sets ($N=25$ patients) [23]. Such end-to-end techniques can be effective, but lack explainability. Regardless of whether

ultrasound scans are assessed by human experts or AI, there is a need for objective measures of scan quality.

In this paper, we devise a 10-point scoring system for assessing scan quality based on US landmarks and characteristics of the US volume. We evaluate the new system in terms of accuracy and inter-observer variability. To assess clinical relevance, we test to what extent scan quality scored by this system predicts the accuracy of an automated AI system (MEDO-Hip).

Methods

Ultrasound scanning

In this scoring exercise, we used a random subset of images previously obtained as part of an institutional health research ethics board approved study. For each subject, we obtained written informed parental consent to perform 3DUS as part of the hip examination. Inclusion criteria were clinical suspicion of DDH (due to risk factors such as hip laxity, asymmetrical skin creases, breech position, female sex, first-born infants and ethnicity). Since DDH can be unilateral or bilateral, we included each hip separately, and in cases of normal hips, we only included one hip per subject in the study.

We performed coronal Graf standard plane 2DUS in both hips using a 12 MHz transducer (L12-5; Philips Healthcare, Andover, MA), as per American College of Radiology recommendations [24]. Along with conventional 2D, we also acquired high-resolution coronal 3DUS using a 13-MHz 3D linear array transducer (13VL5; Philips Healthcare, Andover, MA).

During the scan, the 3DUS probe was positioned with the head resting near the greater trochanter of the infant. The sonographer aligned the probe such that the central slice of the 3D ultrasound volume approximated the Graf standard plane. We performed a 3.2-s automated sweep through $\pm 15^\circ$, generating 256 slices, each of which was 0.13 mm thick and contained 411×192 pixels measuring 0.11×0.20 mm.

The age of infants taking part in this study ranged 26–183 days (68% female). Hips were classified as normal or abnormal based on the orthopedic surgeon's expert opinion. The 'abnormal' category included hips which required treatment for DDH, as well as those which were initially questionably abnormal (e.g., Graf type IIa) but normalised on follow-up without treatment.

Scan quality scoring system

A 10-point scoring was defined based on six imaging features closely linked to Graf analysis: the presence of labrum; os ischium; midportion of the femoral head; straightness of iliac wing, motion artifacts and other artifacts. Scores

and representative examples for each of the features are summarized in Table 1. The alpha angle measurement is intended to be made on slices that contain flat and horizontal ilium, as shown in Table 1 (R1C4). Images captured with an ilium that is not flat and horizontal (Table 1, R1C2, R2C2, R3C2), could result in incorrect measurement. Similarly, the os ischium is clearly visible only in high-quality images (Table 1, R1C4 and R3C4). Movement and imaging artifacts are generally not limited to individual slices, and can be seen in the videos provided as supplementary material.

Hip-scoring exercise

We used 107 images (consisting of 59 normal and 48 abnormal hips) to evaluate the new scoring system. In addition to the 10-point scoring system, we also compared agreement with holistic grading of scan quality on a scale of 1–5. All images were scored by four readers (B.C, a radiology resident with 2 years' experience in hip ultrasound; E.O, a graduate student in radiology with 2 years' experience, and A.H, a research associate with 6 years' experience); and an expert—J.J, lead radiologist, with fellowship training in pediatric and musculoskeletal radiology and 17 years' experience. Readers B.C. and A.H. also scored holistic grading assessments. All images were scored in random order using the Dataturks Labeling Software (www.dataturks.com) hosted on-premise (refer Fig. 2), with readers blinded to each other's assessments.

Assessment of images using MEDO-Hip AI system

Each image was then automatically interpreted using an FDA-approved AI system, MEDO-Hip, which calculated the alpha angle and coverage for all slices. MEDO-Hip uses deep learning to segment the acetabulum, iliac wing and femoral head, and to determine the slice most suitable for diagnosis. An example of measurements overlaid on the most suitable slice is shown in Fig. 3. An expert (JJ) reviewed each MEDO measurement and rated it as acceptable or requiring adjustment (Fig. 3). Clinical diagnosis at the time of scan (0 = normal, 1 = borderline requiring follow-up [e.g. Graf Iia], 2 = dysplastic requiring treatment) was obtained from chart review. We correlated the performance of MEDO-Hip (vs. gold-standard index measurements and clinical diagnosis) against the quality scores obtained from the new scoring system.

Statistics

The descriptive statistics reported are: (1) ICC (3,k) for agreement between users in assessing the overall score, assuming a continuous grading system; and (2) Cohen's kappa, for determining the accuracy of binary

classification as highest quality vs. not highest quality for the categorical scores of each image feature. We tested the significance of differences in ICC or kappa by detecting non-overlapping 95% two-tailed confidence intervals, and the significance of difference in proportions by chi-squared test.

As a potential confounding variable, we separately analysed sub-groups based on clinical diagnosis (normal, borderline or dysplastic). We also used Bland–Altman plots to compare expert assessments to the scores of each of the readers. All calculations were performed using a Python script (Python Version 3.6) developed in-house.

Results

Image quality assessment (holistic and 10-point scoring)

Our data consisted of 58 images of normal hips, 38 borderline (e.g. Graf Iia) and 11 dysplastic. All readers were blinded to clinical diagnosis while scoring the images. Holistic scoring of overall image quality from 0 to 5, with a view toward potential diagnostic utility, resulted in an inter-rater ICC of 0.68 (Table 2).

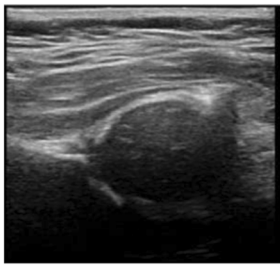
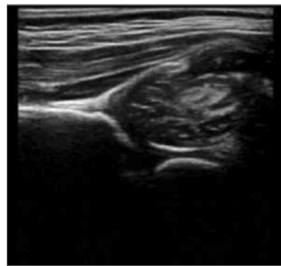
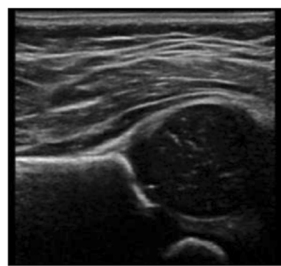
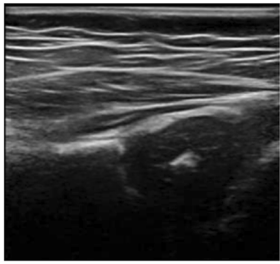
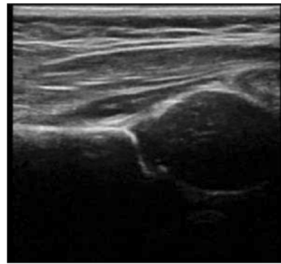
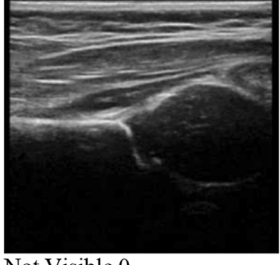

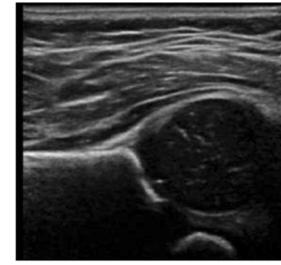

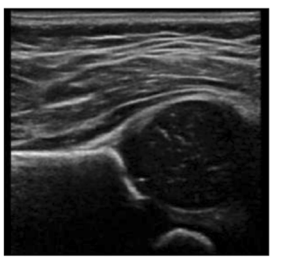
Next, three readers scored the same images using the proposed 10-point scoring system. This gave a higher ICC than holistic scoring in all categories (ICC 0.93 vs 0.68 overall, 0.94 vs 0.72 normal and 0.89 vs 0.65 abnormal; Table 2).

Agreement between readers in assessing each image feature was analysed using Cohen's kappa (Table 3). Kappa values were calculated for each reader vs. expert for each image feature. Agreement between non-expert readers and the expert was high (Kappa > 0.4) in the assessment of the os ischium, the straightness of the iliac wing and the midportion of the femoral head. Agreement was moderate (Kappa > 0.20) for the presence of labrum and other imaging artifacts, features which were rarely absent. Agreement was low for scoring of motion artifacts.

We compared each reader's score against expert scores. Bland–Altman plots comparing the mean score of expert and non-expert readers and the difference in scores are shown in Fig. 4. The scores of readers correlated well with those of the expert.

The new scoring system also clearly differentiated between images of various categories. Examples of images scored as poor (score = 2/10), moderate (score = 6/10) and excellent (score = 10/10) are shown in Fig. 5. The scoring system clearly differentiated Fig. 5a and b, which were both graded at image quality 2 (out of 5) by readers in subjective scoring.

Table 1 Definitions of Scoring system for hip ultrasound images based on six features—ilium (0–2),labrum(0–1), os ischium(0–2), femoral head(0–1), motion artifact (0–2) and other imaging artifacts (such as limited penetration or excessive image noise) (0–2)

Iliac Wing	 Not straight 0	 Straight (at an angle) 1	 Straight and Horizontal 2
Labrum	 Not Visible 0	 Visible 1	
Os ischium	 Not Visible 0	 Faintly Visible 1	 Clearly Visible 2
Femoral head	 Not Visible 0		

*Movement and imaging artifacts are not limited to individual slices and can be seen in the videos provided as supplementary material

Fig. 2 Graphical user interface (GUI) used for the reading exercise. Using the feature-based scoring system the user answers simpler and more specific questions on each aspect of the image. For example, in this image the user marked a straight ilium, os ischium faintly visible and minor motion artifacts (these can be seen in the video attached as supplementary material)



Fig. 3 MEDO-Hip used Artificial Intelligence to interpret the images by measuring the alpha and coverage. As shown in the image, the AI proposes measurements and the user can modify these by moving the end-points (shown by blue and orange circles) of the line widgets

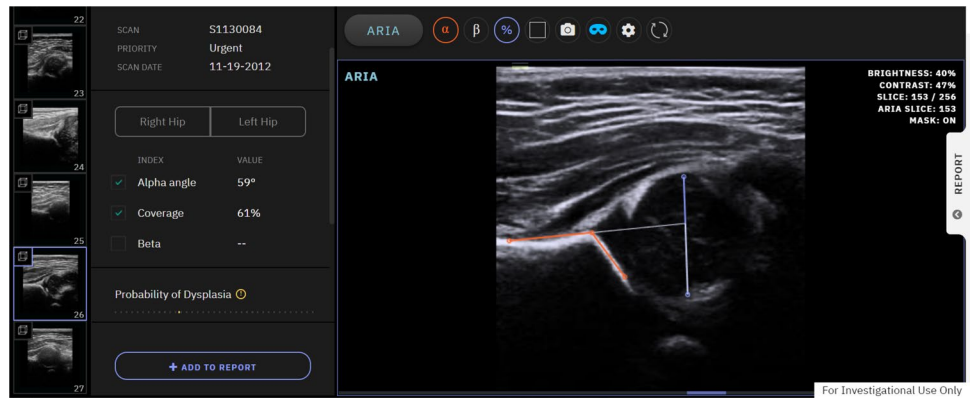


Table 2 Agreement between users in terms of ICC

Scoring technique	ICC [95% CI] (holistic)	ICC [95% CI] (10-point scoring)
All subjects (<i>n</i> = 107)	0.68 [0.64–0.72]	0.93 [0.91–0.96]
Subgroup normal (<i>n</i> = 59)	0.72 [0.68–0.73]	0.94 [0.89–0.96]
Subgroup abnormal (<i>n</i> = 48)	0.65 [0.60–0.68]	0.89 [0.87–0.94]

Note that the agreement was significantly higher when the readers used the scoring system in all categories

Table 3 Agreement between readers of DDH image features quantified using Cohen’s Kappa

Image feature	Kappa (mean ± SD)	Feature prevalence		
		0	1	2
Ilium	0.49 ± 0.12	16% (17/107)	21% (22/107)	63% (68/107)
Labrum	0.21 ± 0.19	6% (6/107)	94% (101/107)	–
Os ischium	0.67 ± 0.06	21% (22/107)	28% (30/107)	51% (55/107)
Femoral head	0.65 ± 0.07	4% (4/107)	96% (103/107)	–
Motion artifacts	0.13 ± 0.14	3% (3/107)	42% (45/107)	55% (59/107)
Other imaging artifacts	0.20 ± 0.06	7% (7/107)	36% (39/107)	57% (61/107)

The corresponding positive prevalence of each feature is also reported

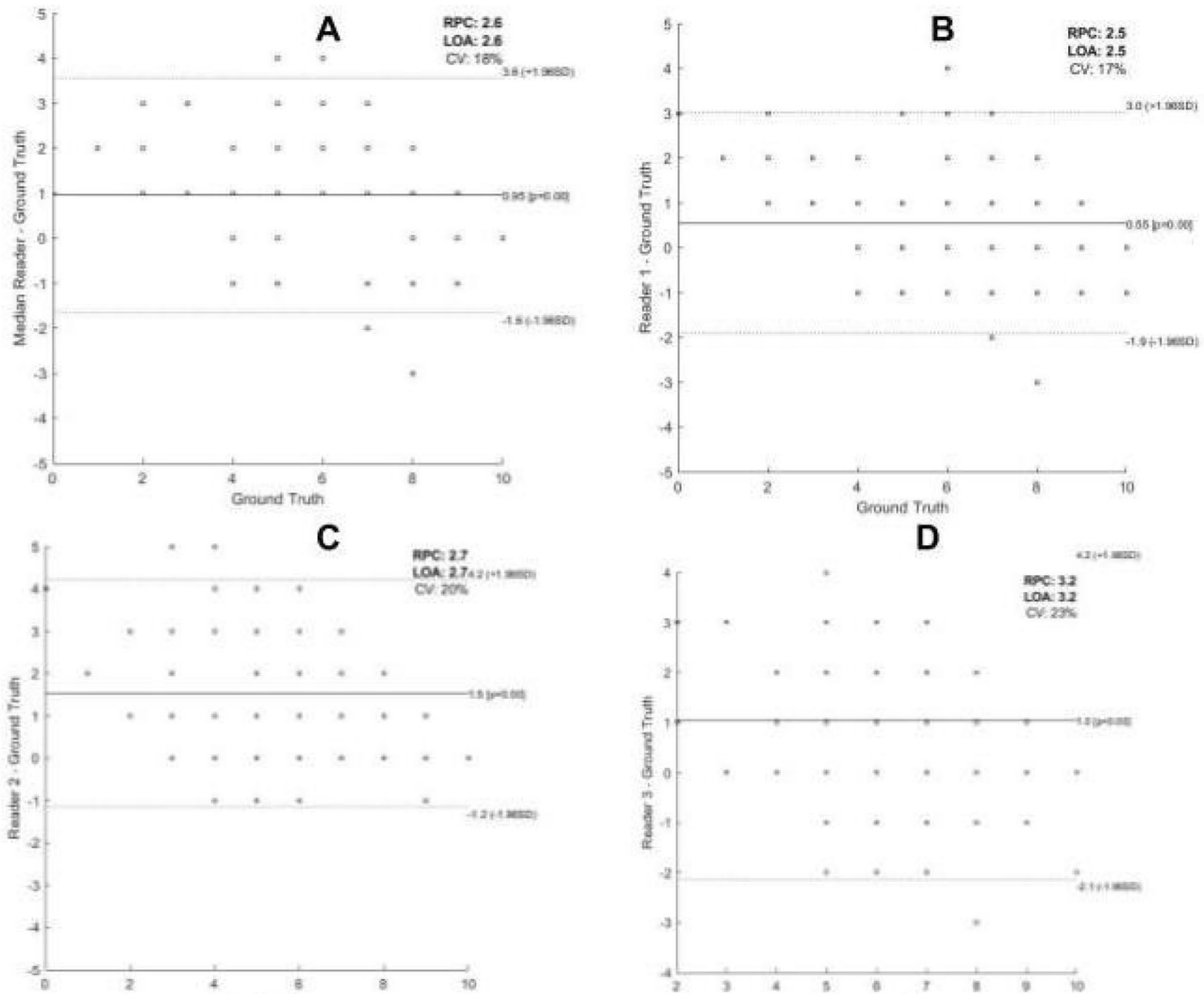


Fig. 4 Bland Altman plots of each reader vs ground-truth. **a** The median score of readers and ground truth vs the difference between the median score and ground truth. Similarly, **b**, **c**, **d** shows the

ground truth vs difference for each reader. Note that there was no systematic error trend in any of the plots



Fig. 5 Examples of hip scans of varying image quality. **a** A poor quality scan (score=2) in which the iliac roof is not straight, femoral head is not clearly visible, the scan also does not contain the os ischium. **b** Moderate quality scan (score=6) in which the ilium is

straight but not horizontal, femoral head is adequately visualized, and os ischium is faint. **c** Excellent quality scan (score=10) with horizontal ilium, clear femoral head, os ischium and labrum

Diagnostic assessment

All images, regardless of quality, were analysed using a commercial FDA-approved AI diagnostic system (MEDO-Hip) to automatically measure alpha angle and coverage and predict the probability of hip dysplasia. When this AI system cannot identify anatomy reliably, it generates an output of ‘uninterpretable’. Out of 107 images, 13 were uninterpretable by the AI system. The median value of quality scoring by our readers on the 10-point scale for each image was compared to AI interpretability. All images missed by the AI system had median quality scores of less than 8/10 (Fig. 6). The median quality score was significantly lower for the 13 cases uninterpretable by AI (score 3 ± 1.2) than for the 94 cases for which AI was able to identify the anatomy (quality score 7 ± 1.8 , $p < 0.05$).

We also analysed the scores provided by the expert for specific image features in the images that were uninterpretable by AI. All images that were uninterpretable had inadequate visualisation of the os ischium (not visible in 77% and faintly visible in 23%) and presence of motion artifacts (major distortions 77% and minor distortions in 23%) (Table 4).

The measurements made by AI on each image it could interpret were categorised as acceptable or not acceptable/ requiring adjustment by an expert (JJ). The proportion of acceptable AI measurements was significantly higher in images with a median quality score of 8/10 or higher (89%) than in images with a quality of 7/10 or less (57%, $p < 0.01$).

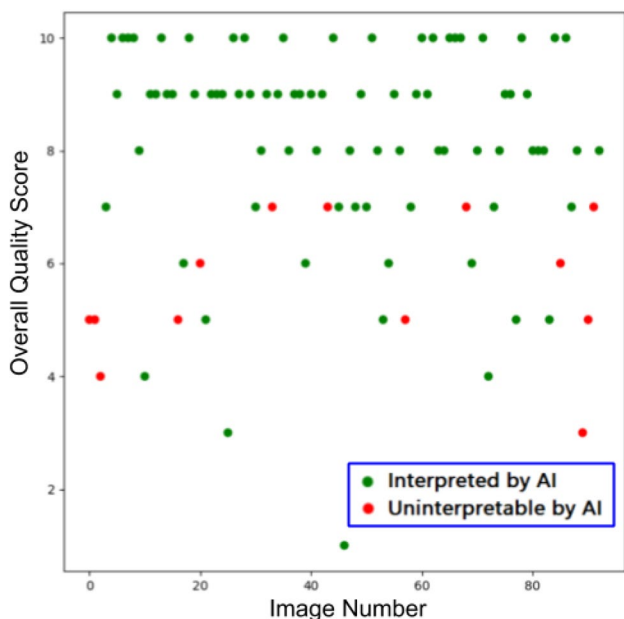


Fig. 6 Median quality scores from 4 readers on all images, and MEDO-Hip interpretability. Note that above a threshold of 7 the AI system was able to analyze all images

Table 4 Scores assigned by the expert for images that were uninterpretable

	Feature scores		
	0	1	2
Ilium	24%	38%	38%
Os ischium	77%	23%	0%
Labrum	7%	93%	–
Femoral head	0%	100%	–
Motion artifacts	23%	77%	0%
Other artifacts	7%	39%	54%

Note that motion artifacts were present in all the images (77% minor and 23% major)

Discussion

Diagnosis of hip dysplasia from ultrasound images depends on image quality, and automated AI analysis of these images may be more dependent on high-quality images than that of human experts. In this study, we proposed and evaluated a new scoring system to systematically assess the image quality of hip ultrasound scans, and used this to demonstrate the effect scan quality has on the functioning of a commercial AI system for hip dysplasia diagnosis. We showed that systematic quality scoring is significantly more reliable than holistic scoring, and that quality scores $\leq 7/10$ are associated with significantly higher risk that AI will either be unable to interpret the images or will produce unexpected results that require expert human adjustment.

Using our image quality scoring system, the reader provides a score based on 6 features. Four of these are based on the visibility of landmarks (ilium, labrum, femoral head, os ischium) necessary for Graf assessment. Acquiring an image with straight and horizontal ilium is crucial for accurate alpha angle measurement, so we gave this a high weight of 2 points. Graf criteria also recommend performing measurement at the deepest part of the joint, where the diameter of the femoral head is highest and the os ischium is clearly visible, so this was also weighted at 2 points. Visibility of the labrum is important for measuring the beta angle, which is used in some centres. Similarly, complete visualisation of the femoral head is important for measurement of acetabular coverage, which is commonly used in clinical practice.

In our study, more than 95% of images had clearly visible labrum and femoral heads. This was expected, since scanning was performed in a rigorous research environment by experienced sonographers in dedicated sessions, representing an idealised situation. However, in routine clinical practice at scanning centres with less experienced sonographers, we would expect to see more examples of inadequate images in which the labrum and femoral head were not clearly visible.

We included two features that related to the presence of imaging artifacts, since these are commonly seen in 3DUS and handheld sweeps. Artifacts could either be due to movement by the infant, accidental hand movement or imaging artifacts intrinsic to ultrasound (such as shadowing, reverberation or mirror artifacts).

This study was performed on 3DUS, in which a stack of slices are obtained at a consistent spacing governed by the mechanical movement of a transducer. The same quality-scoring system can be applied to 2DUS sweeps, short videos taken as a sonographer manually ‘sweeps’ across the hip, which might have more prominent motion artifacts. The scoring system can also apply to conventional single 2DUS images, with the motion artifact score generally expected to be 0.

Agreement

The overall agreement of readers was significantly and substantially higher for the new scoring system than for holistic scoring of images. This is likely because unlike holistic scoring, the new scoring system asks the user for a set of simpler (mostly binary) decisions. Non-expert readers also showed high agreement in assigning an overall score per scan ($ICC = 0.93$). Agreement between readers ranged between moderate (for labrum) to high (for ilium, os ischium and femoral head) for categorisation of anatomical landmarks. The relatively lower agreement for labrum scores was likely because this structure is more difficult to visually identify than bony edges. This is a well-known problem that also likely accounts for the high inter-observer variability of the beta angle, which is measured based on the position of the labrum [25]. Low agreement may also be due in part to the high prevalence ($> 90\%$) of visible labrum; there was disagreement on which cases did not show a labrum. Agreement was low for scoring of motion artifacts and moderate for other image artifacts. This is likely because these scores were more subjective (‘minor’ vs. ‘major’) than the other features. There would be room for improvement by developing more specific definitions of the amount of each artifact to score as a 1 or 2.

Non-expert and expert readers showed close agreement on quality scores without bias on Bland–Altman plots. We noted a nonsignificant trend toward poorer agreement on quality scores in abnormal hips (dysplastic and borderline). This is not surprising, as dysplastic hips can have various shapes that are difficult for readers to interpret. For example, the iliac roof in dysplastic hips is generally rounded, which could result in challenges in identifying whether a scan has a flat and horizontal iliac roof.

Impact of scan quality on AI systems

Several approaches [17–22] have been proposed for automatic interpretation of hip ultrasound. These packages are designed to function best with high-quality images. We gave one such system—the FDA-approved MEDO-Hip—a difficult test of performance by supplying it with our set of images of deliberately widely varying quality. As shown in Fig. 6, 100% of images with quality $> 7/10$ and 95% of images scored $> 5/10$ were successfully analysed. The accuracy of the AI system correlated with scan quality scores, as significantly larger proportions of high-quality images (89% vs 57% in lower quality images) gave acceptable AI interpretations, and all 13 images that MEDO-Hip was unable to interpret had quality scores $\leq 7/10$. Specifically, all images that were uninterpretable had poor visibility of the os ischium and significant motion artifacts.

This highlights that an AI system for image analysis is, as one would expect of any human expert, less accurate when analyzing lower-quality images. Our quality-scoring system could potentially be used as a preprocessing step. An initial assessment of quality could be performed up front and any low-quality scan flagged to be repeated prior to presenting the images to the diagnostic AI system, which would improve the overall clinical utility and acceptance.

A key advantage of our quality-scoring system is that it explains why the image is uninterpretable and gives feedback to the sonographer as to how to improve it. An AI network could be used to calculate the quality score, potentially even in real time, providing clear feedback to the user as to exactly what needs to be improved for the images to be of adequate quality. Whether calculated by AI automatically or scored manually, this new scoring system can be used by sonographers at the time of scanning or as part of interpretation.

A limitation of our scoring system is that the individual features used are still rated subjectively. However, these are based on simple and clearly identifiable features, which gave high accuracy and agreement between readers. One approach to completely eliminate inter-observer variation in scoring would be to develop computational techniques like machine learning to estimate each of the individual features. More extensive studies on larger datasets are planned as a follow-up to this pilot study. Data from these studies will be used to develop techniques to fully automate the scoring.

Conclusion

We proposed a new scoring system for assessing the quality of ultrasound hip scans, showing that the new scoring system is reliable for users of varying experience and produces a useful threshold: scans with quality scores of 7/10 or less

significantly increase the risk of inaccurate or incomplete diagnostic image interpretation by an AI system.

Our quality-scoring system can be used as an aid for a sonographer and radiologist in assessing scan adequacy and improving sonographer training, and as a preprocessing step in AI systems for hip ultrasound diagnosis.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s40477-021-00560-4>.
Acknowledgements The authors thank Women and Children's Health Research Institute (WCHRI) for the research funding that supported this work. We also thank Dr Sukdeep Dulai for her insights on surgical management of DDH.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical statement This study was performed in line with the principles of the Declaration of Helsinki Declaration of 1975, as revised in 2000.

Informed consent Not applicable.

References

- Furnes O, Lie SA, Espehaug B, Vollset SE, Engesaeter LB, Havelin LI (2001) Hip disease and the prognosis of total hip replacements. *J Bone Jt Surg Br* 83-B(4):579–579. <https://doi.org/10.1302/0301-620x.83b4.0830579>
- Price CT, Ramo BA (2012) Prevention of hip dysplasia in children and adults. *Orthop Clin N Am* 43(3):269–279
- Atalar H, Sayli U, Yavuz OY, Uraş I, Dogruel H (2007) Indicators of successful use of the Pavlik harness in infants with developmental dysplasia of the hip. *Int Orthop* 31(2):145–150
- Buonsenso D et al (2020) Developmental dysplasia of the hip: real world data from a retrospective analysis to evaluate the effectiveness of universal screening. *J Ultrasound*. <https://doi.org/10.1007/s40477-020-00463-w>
- Buonsenso D, Menzella N, Morello R, Valentini P (2020) Indirect effects of COVID-19 on child health care: delayed diagnosis of developmental dysplasia of the hip. *J Ultrasound* 23(3):443–444
- Shorter D, Hong T, Osborn DA (2013) Cochrane Review: screening programmes for developmental dysplasia of the hip in newborn infants. *Evid Based Child Health* 8(1):11–54
- Dezateux C, Rosendahl K (2007) Developmental dysplasia of the hip. *Lancet* 369(9572):1541–1552
- Bache CE, Clegg J, Herron M (2002) Risk factors for developmental dysplasia of the hip: ultrasonographic findings in the neonatal period. *J Pediatr Orthopaed B* 11(3):212–218. <https://doi.org/10.1097/01202412-200207000-00004>
- Clarke NM, Clegg J, Al-Chalabi AN (1989) Ultrasound screening of hips at risk for CDH. Failure to reduce the incidence of late cases. *J Bone Joint Surg Br* 71(1):9–12
- Mehdizadeh M, Dehnavi M, Tahmasebi A, Mahlisha Kazemi Shishvan SA, Babakhan Kondori N, Shahnazari R (2020) Transgluteal ultrasonography in spica cast in postreduction assessment of developmental dysplasia of the hip. *J Ultrasound* 23(4):509–514
- Barbuto L et al (2019) Pediatric musculoskeletal ultrasound: a pictorial essay. *J Ultrasound* 22(4):491–502
- Graf R (1984) Fundamentals of sonographic diagnosis of infant hip dysplasia. *J Pediatr Orthop* 4(6):735–740
- Jaremko JL, Mabee M, Swami VG, Jamieson L, Chow K, Thompson RB (2014) Potential for change in US diagnosis of hip dysplasia solely caused by changes in probe orientation: patterns of alpha-angle variation revealed by using three-dimensional US. *Radiology* 273(3):870–878
- Geng C et al (2020) Using 3-dimensional ultrasound iSlice technology for the diagnosis of developmental dysplasia of the hip. *J Ultrasound Med* 39(6):1117–1123
- Mostofi E et al (2019) Reliability of 2D and 3D ultrasound for infant hip dysplasia in the hands of novice users. *Eur Radiol* 29(3):1489–1495
- Zonoobi D et al (2018) Developmental hip dysplasia diagnosis at three-dimensional US: a multicenter study. *Radiology* 288(3):912–912. <https://doi.org/10.1148/radiol.2018184016>
- Hareendranathan AR, Mabee M, Punithakumar K, Noga M, Jaremko JL (2016) A technique for semiautomatic segmentation of echogenic structures in 3D ultrasound, applied to infant hip dysplasia. *Int J Comput Assist Radiol Surg* 11(1):31–42. <https://doi.org/10.1007/s11548-015-1239-5>
- Hareendranathan AR, Mabee M, Punithakumar K, Noga M, Jaremko JL (2016) Toward automated classification of acetabular shape in ultrasound for diagnosis of DDH: contour alpha angle and the rounding index. *Comput Methods Programs Biomed* 129:89–98
- Hareendranathan AR et al (2017) Semiautomatic classification of acetabular shape from three-dimensional ultrasound for diagnosis of infant hip dysplasia using geometric features. *Int J Comput Assist Radiol Surg* 12(3):439–447
- M. Tang, Z. Zhang, D. Cobzas, M. Jagersand, and J. L. Jaremko, "Segmentation-by-detection: A cascade network for volumetric medical image segmentation," *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. 2018, doi: <https://doi.org/10.1109/isbi.2018.8363823>.
- Zhang Z, Tang M, Cobzas D, Zonoobi D, Jagersand M, Jaremko JL (2018) End-to-end detection-segmentation network with ROI convolution. In: *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. <https://doi.org/10.1109/isbi.2018.8363859>.
- Quader N, Hodgson A, Abugharbieh R (2014) Confidence weighted local phase features for robust bone surface segmentation in ultrasound. *Workshop on Clinical Image*. https://doi.org/10.1007/978-3-319-13909-8_10.
- Paserin O, Mulpuri K, Cooper A, Hodgson AJ, Garbi R (2018) Real time RNN based 3D ultrasound scan adequacy for developmental dysplasia of the Hip. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, 2018, pp 365–373
- Harcke HT et al (2009) AIUM practice guideline for the performance of an ultrasound examination for detection and assessment of developmental dysplasia of the hip. *J Ultrasound Med* 28(1):114–119
- Simon EA, Saur F, Buerge M, Glaab R, Roos M, Kohler G (2004) Inter-observer agreement of ultrasonographic measurement of alpha and beta angles and the final type classification based on the Graf method. *Swiss Med Wkly* 134(45–46):671–677

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.