# Application of Artificial Intelligence in Hip Ultrasound and its Performance in Detecting Developmental Dysplasia of the Hip

by

Siyavash Ghasseminia

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Medical Sciences - Radiology and Diagnostic Imaging

University of Alberta

# ABSTRACT

Developmental Dysplasia of Hip (DDH) which represents a wide range of abnormalities from acetabular dysplasia to fixed dislocation, is mainly defined by a loss of conformity between the femoral head and the acetabulum and it can lead to structural instability and osteoarthritis. The diagnosis of DDH on ultrasound is mainly based on Graf method. It is primarily focused on obtaining a single high-quality coronal 2D image containing elements such as acetabulum, ilium, and round femoral head. Graf method on 2DUS not only suffers from low inter and intra- rater agreement, but its reproducibility has also been shown to decline over time. Recording 'sweep' images allows more comprehensive hip assessment and introduces opportunities for automation by artificial intelligence (AI). In this thesis, agreement between readers with various background and expertise and an AI algorithm in detecting DDH is assessed. Additionally, this thesis evaluates the accuracy of AI in classification of DDH from 3DUS and its correlation with conventional clinical 2DUS.

# PREFACE

A version of Chapter 1 has been published in Indian Journal of Orthopaedics – Ghasseminia, S., Hareendranathan, A.R. & Jaremko, J.L. Narrative Review on the Role of Imaging in DDH. JOIO (2021). https://doi.org/10.1007/s43465-021-00511-5

A version of Chapter 2 has been accepted for publication in Journal of Pediatric Orthopaedics – Ghasseminia, S., Jaremko, J.L., et al. Inter-observer variability of hip dysplasia indices on sweep ultrasound for novices, experts, and artificial intelligence

# ACKNOWLEDGEMENTS

Firstly, I would like to thank Dr. Jacob Jaremko. I would have never been able to even start this journey without his generous support and could have not imagined concluding it without his continuous supervision and mentorship. I did not know what scientific research really meant and without doubt it is owed to his kind patience with my many random calls, weekend messages and midnight updates, that I could conduct one. I am so grateful for everything I learned from him. Thank you Jacob!

I want to also thank Dr. Abhilash Hareendranathan for his continuous support and kindness. He was always there to help with everything that seemed stuck. I appreciate every minute I spent collaborating with him and every advice he gave me.

I want to thank “Team MEDO” and especially Dr. Dornoosh Zonoobi for her irreplaceable help in conducting this research. Without her and without them this was never going to be possible. Dornoosh jaan thank you for everything!

I would like to thank Carol Rae for bearing with my questions/emails every now and then, for her comprehensive and swift responses, and especially for all the reminders she sent me for every deadline I was going to miss.

Finally, I want to thank Charlotte for always being there for me and putting up with all my nags, Yasi for all her encouragements and kindliness and my family for their beautiful smiles on our every-morning video calls.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| DDH | Developmental Dysplasia of Hip |
| CT | Computed Tomography |
| MRI | Magnetic Resonance Imaging |
| US | Ultrasound |
| 2D | Two Dimensional |
| 3D | Three Dimensional |
| CMT | congenital muscular torticollis |
| AI | Artificial Intelligence |
| ICC | Intraclass Correlation Coefficient |
| THA | total hip arthroplasty |
| FAVA | femoral anteversion angle |
| AAI | Anterior Acetabular Index |
| PAI | Posterior Acetabular Index |
| LCEA | Lateral Centre-Edge Angle |
| FHC | Femoral Head Coverage |
| CI | Confidence Interval |
| AVN | Avascular necrosis |
| CEUS | Contrast Enhanced Ultrasound |
| ACR | American College of Radiology |
| AAP | American Academy of Pediatrics |
| AAOS | American Academy of Orthopaedic Surgeons |

| | |
|---|---|
| AJR | American Journal of Roentgenology |
| FDA | Food And Drug Administration |
| SD | Standard Deviation |
| ACA | Acetabular Contact Angle |
| SN | Sensitivity |
| SP | Specificity |
| LR+ | Positive Likelihood ratios |
| LR- | Negative Likelihood ration |
| PPV | Positive Predictive Value |
| NPV | Negative Predictive Value |
| TN | True Negative |
| FN | False Negative |
| FP | False Positive |
| TP | True Positive |
| $R^2$ | R squared |
| IHDR | International Hip Dysplasia Registry |

# LIST OF SYMBOLS

| | |
|---|---|
| α | Alpha Angle |
| B | Beta Angle |
| ° | Degree |

# INTRODUCTION AND THESIS SCOPE

Developmental dysplasia of hip (DDH) is characterized by a non-conformity between the femoral head and acetabulum. DDH can be treated by non-invasive methods (Pavlik harness) if it is detected in infancy (<6 months). However, if it is missed, it can lead to osteoarthritis which may only be treated by hip replacement surgeries that can be hard, painful, and expensive.

Many names have been coined, several definitions have been suggested by experts, clinicians, and organizations, numerous risk factors have been studied, and various detection and diagnosis methods (physical examination, radiography, ultrasound screening etc.) have been introduced for DDH. Logically, there are also many classification paradigms that have been proposed for DDH.

In the first Chapter of this thesis, the developmental dysplasia of hip is first studied from a historical perspective. Some of the challenges in defining DDH are explained and consequently some of the more important associated risk factors are discussed. In the same Chapter DDH various classifications are reviewed and different medical imaging modalities (X-ray, CT, MR, and ultrasound) and their respective utilization in detection and management of DDH are covered.

Chapter One also discusses the opportunities for automation of DDH detection and classification using new technologies such as artificial intelligence. Finally, Chapter One concludes with some of the widely accepted standards for, as well as the concerns associated with DDH diagnosis and its treatment.

The focus of this thesis is on first understanding the role of imaging in DDH and then the classification of DDH by using two and three-dimensional ultrasound and the application of AI in detecting abnormal hips. This combines two innovations. first, acquiring images in a different way,

by use of cine sweep video / 3D ultrasound instead of single 2D image captures; and second, the use of artificial intelligence to interpret these images. Chapter Two presents the results of an inter-observer variability of hip dysplasia indices on sweep ultrasound. Two sets of ultrasound hip images (2D images and sweep images) were rated by 12 readers with different levels of (reading hip ultrasound) expertise and experience. The same sets of images were also rated by an artificial intelligence algorithm and the results were statistically analysed and compared. Chapter two studies the correlation between the ratings of human raters and AI as just another individual reader.

While Chapter Two evaluates reliability, the validity of the new ultrasound tools is assessed, on a large scale, in Chapter Three. This is a multi-center study on the performance of AI in reading 3D ultrasound images of the hip. Thousands of 2D images of infants' hips as well as their respective free-hand 3D ultrasound images, through an 8-year study were prospectively collected, stored, clinically diagnosed, and eventually read by an artificial intelligence algorithm. Chapter Three evaluates the accuracy of AI in differentiating abnormal hips from healthy ones and discusses the future opportunities for enhancing the clinical procedures by the power of AI without heavy reliance on the acquisition of perfect hip ultrasound images. It is beneficial to mention that apart from some extra non-graphical information captured from a 3D ultrasound probe throughout the acquisition of images (e.g., spacing between the slices), 2D sweeps and 3D images are very similar despite their huge differences in availability and cost.

In summary, this thesis outlines work in which we first assessed the current state of hip dysplasia imaging to identify a need for improved image acquisition and interpretation tools, then after developing these tools (a protocol for cine sweep / 3D ultrasound acquisition and AI image interpretation), evaluated their reliability and validity. We conclude that the new approach has key

benefits over existing imaging and may even make widespread population screening for hip dysplasia possible.

# CHAPTER 1

# A NARRATIVE REVIEW ON THE ROLE OF IMAGING IN DDH.

## Introduction

### *Definition, Risk Factor, Diagnosis, Classification*

Developmental Dysplasia of Hip (DDH) is considered essentially a condition of instability **[1]**. The term "congenital dislocation of the hip" was first coined by Dupuytren in 1847 **[2]**. He describes it as a displacement which appears due to a defect in the depth or completeness of the acetabulum. However, not all dysplastic hips are dislocated. PJ Klisic, in an article called "Congenital dislocation of the hip a misleading term" **[3]** noted that due to the pathologic variability of the disorder, and that it can emerge at various points throughout skeletal development, the term "congenital dislocation" should be changed to "developmental displacement". Since displacement is thought secondary to changes in anatomic shape, size, and orientation, the term "development dysplasia of the hip" has now been widely accepted to describe the misalignment between the femoral head and the acetabulum **[4]**

Viktor Bialik **[**5**]** defines three time periods in the history of modern medicine for the determination and diagnosis of DDH. In the first phase (1920s to 1950s), DDH prevalence was determined opportunistically when seen post-mortem or surgically almost randomly approximated (0% for Africans and 0.06% ~ 40% for other ethnicities). During the second phase (1950s to 1980s), the incidence was determined based on the detection of neonates' unstable hips by

conducting physical examinations and radiographs. (0.04% to 16.8%). During the third phase (1980s onwards), ultrasound became routinely available, resulting in a large increase in imaging of DDH. As is often the case when more imaging is done for a disease, an increase in the estimated incidence occurred (4.4% to 51.8%). This wide range is due to various definitions and classification of DDH. Clearly, diagnosing more than half the population with DDH is of doubtful real clinical value, and Bialik states that "clinical and sonographic neonatal screening, whether separately or in combination, seems to have introduced more confusion by eventually disclosing wide discrepancies between the clinical and sonographic findings."

DDH terminology and definitions were first introduced by Dupuytren **[2]** and Klisic **[3]**. However, it was Barlow who proposed a rigorous physical examination technique in 1963 **[6, 7, 8]**. His proposed examination approach was a continuation to his 1961 study on 7,742 children for congenital dislocation and other abnormalities in the first week of life. Classification systems introduced for DDH over the years vary based on the clinical and imaging methods of evaluation. We will take a closer look at some of them (X-ray, CT and Ultrasound) in the following sections.

## X-ray, CT and MR

Before ultrasound, the sole images available for DDH diagnosis were pelvis radiographs (*Figure 1.1*). These are necessarily limited in capability to evaluate hip alignment and stability due to their static two-dimensional nature. Additionally, for infants whose growth plates are open and axial growth is still expected, the femoral head and a great part of the acetabulum are cartilaginous and hence not visible **[1]**. Weinstein et al. classified DDH in a manner intended to be relevant to treatment **[9]**: (1) inclination of the acetabulum with centralized ossification center, Shenton's line intact (**dysplasia**), (2) subluxated ossification center, Shenton's line broken (**subluxation**), and (3) ossification center outside the acetabulum (**dislocation**). The Tönnis Classification (Grade 1~4)

quantifies the severity of DDH using the relative position of the ossific nucleus and the acetabulum on X-ray images of the hip joint **[10]**. The acetabular index, measured from Hilgenreiner's line through the triradiate cartilages **[11]** *(Figure 1.2*), has age-specific normal values and is often used in diagnosis and follow-up. Since Tönnis method required the ossification centre to be present, a new radiographic classification was proposed by the International Hip Dysplasia Institute (IHDI) that used the mid-point of the proximal femoral metaphysis as a reference landmark **[12]** solving the limitation of the Tönnis method.

***Figure 1.1*** X-Ray of a patient with a dysplastic, subluxed left hip and a normal right hip.

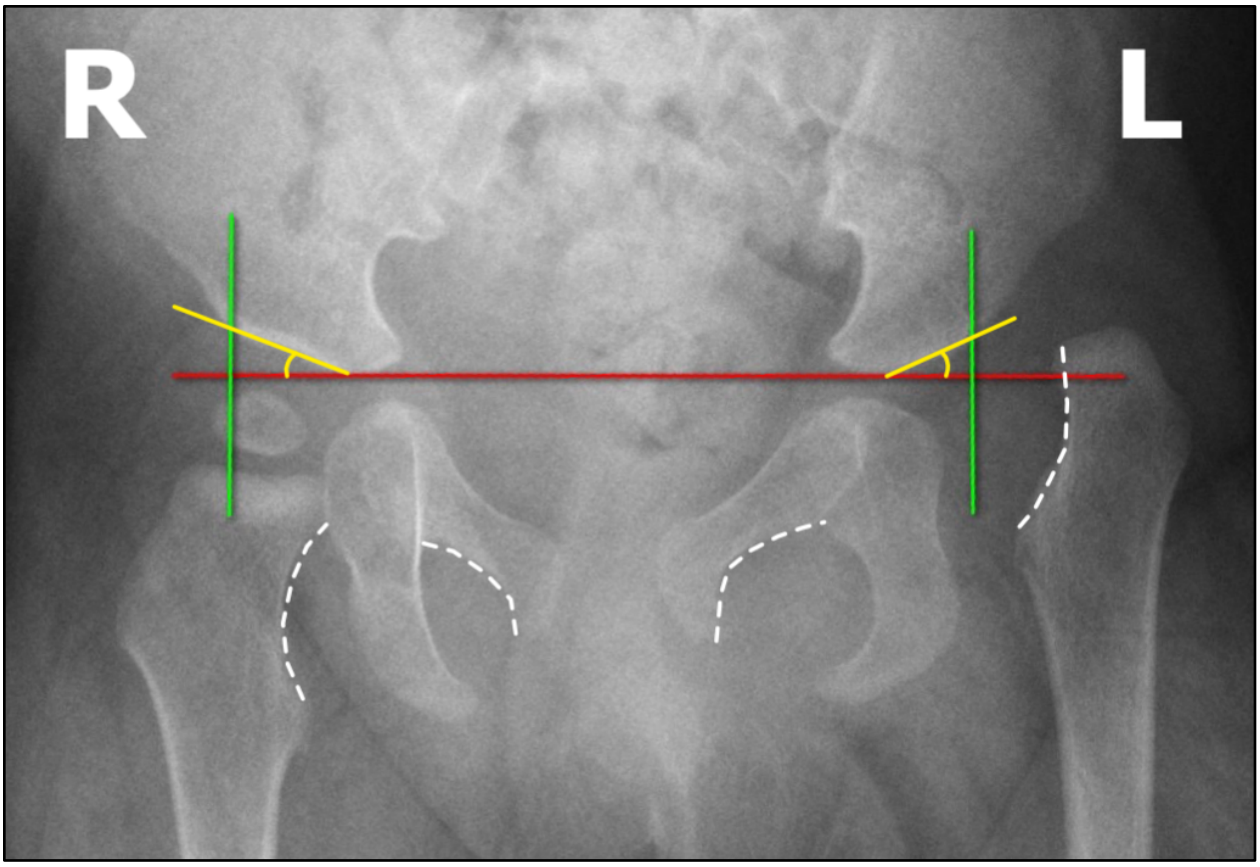


**Figure 1.2** Hilgenreiner (red) and Perkins lines (green). The left hip shows delayed development of the femoral head ossification centre, which is positioned lateral to Perkin's line, indicating dislocation. Also, although acetabular index (yellow) is only slightly increased at the left hip, the disruption of the normally continuous arc of Shenton's line on this side, together with superior/lateral position and delayed ossification of the femoral head, confirms a dysplastic, dislocated left hip. The normal right femoral head ossification centre is medial to Perkin's line, as expected with normal acetabular coverage.

Cross-sectional imaging (CT and MRI) offers the ability to more fully evaluate 3D hip deformity than radiographs (Figure 1.3 and 1.4). In 2012, Akiyama et al, using pelvic CT images of 79 hips, studied the correlation between acetabular version and coverage with three subgroups of hip dysplasia (anterior, global, and posterior deficiency) **[13]**. This was one of the first attempts to look at the DDH from a three-dimensional perspective. Later on in 2017, a similar but larger study by Nepple et al **[14]** was conducted to better understand the variability in 3D acetabular

deficiency and to define subtypes of acetabular dysplasia based on 3D morphology. They also considered the same three patterns (anterior, global, and posterior deficiency), which they concluded commonly occurred among young adult patients with mild, moderate, and severe acetabular dysplasia.

In another CT study Fujii et al **[15]** concluded that acetabular tilt angle was increased in dysplastic hips and reported a correlation between the rotational position of the acetabulum in the pelvis with acetabular version and coverage in hip dysplasia. Perhaps the first usage of CT in classification of DDH was by Hartofilakidis et al. in 1996. They used CT to investigate four parameters of acetabular anatomy, a continuation to their 1988 study which had classified DDH into three classes as (1) dysplastic (2) low dislocated and (3) high dislocated **[16, 17, 18]**.

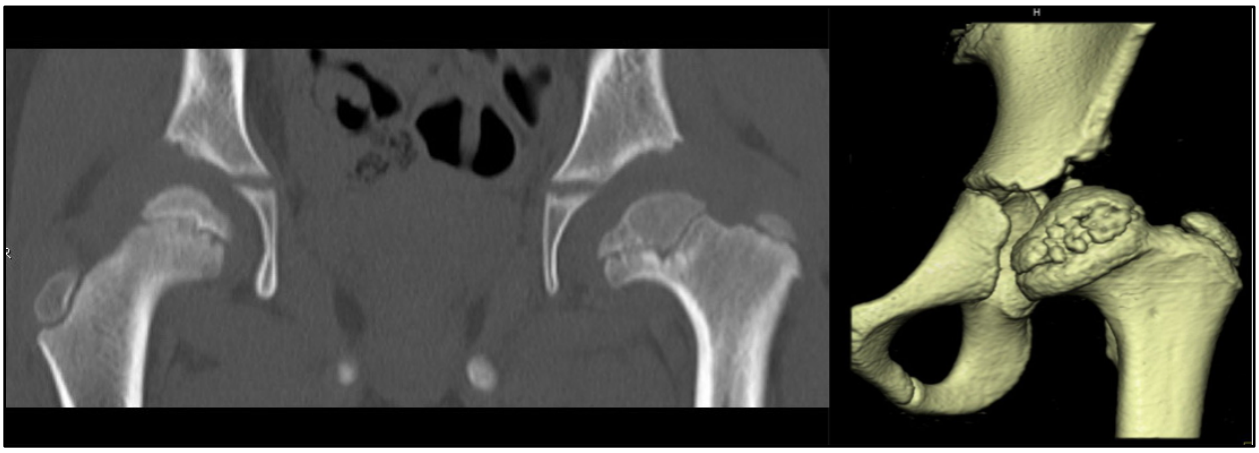

**Figure 1.3** CT scan in DDH. (a) Coronal and (b) 3D reformatted images showing the 3D shape of a chronically dysplastic left hip with flattened and fragmented femoral head due to avascular necrosis, in a 6-year-old with left sided DDH.

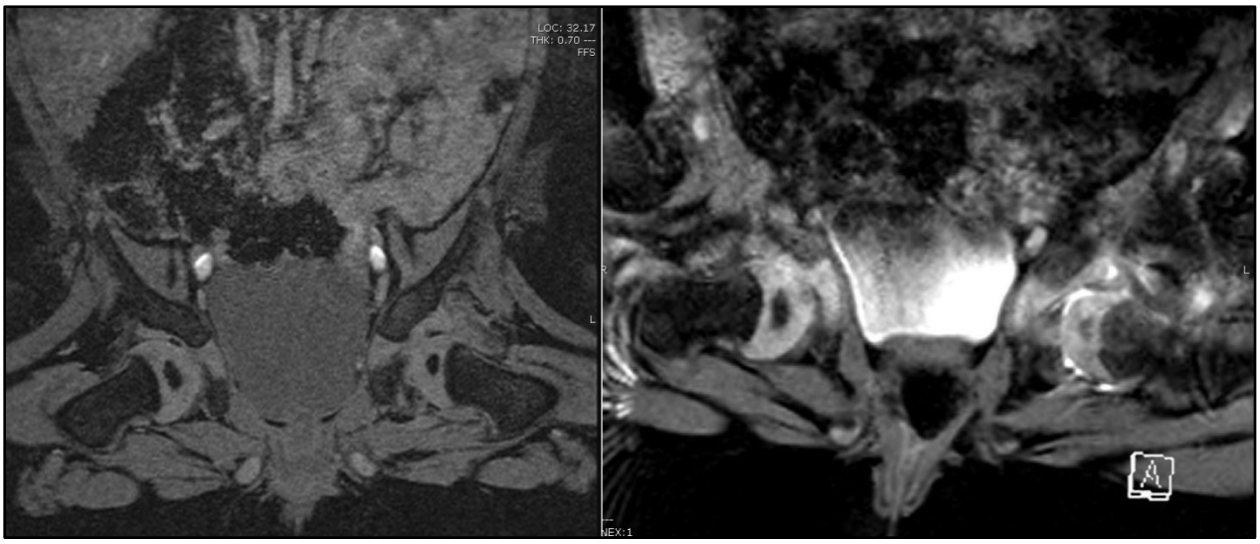


**Figure 1.4** MRI in a DDH patient performed while still sedated immediately after spica cast placement by a surgeon. The left hip is dysplastic, mildly subluxed, and has an inverted labrum (triangular low signal) which may be blocking reduction of the hip. These images demonstrate that MRI of infant hips is of low resolution and can be difficult to interpret reliably.

In 2017, Wilkin GP et al, in their article "A Contemporary Definition of Hip Dysplasia and Structural Instability" summarized that hip dysplasia is in fact a 3D deformity of the acetabulum and that multiple patterns of hip instability exist that may not be completely assessed on 2D imaging **[19]**. In the same year, Joel Wells et al worked on head and neck offset differences of the femora of the dysplastic hip, since according to them DDH represented a spectrum of deformities on both sides of the joint in contrast to the many studies that had been conducted only on the acetabular side **[20]**.

Jaremko et al, in a retrospective study of infants and toddlers with DDH who had been treated with spica casting, reviewed multiple indices from different sources to determine which indices showed sufficient reliability to be potentially useful in assessment of acetabular geometry, degree of hip reduction and barriers to reduction **[21]** (Figure 1.5). Later, in 2017 Hesham et al **[22]**, reported a high inter- and intra-rater reliability (ICC>0.90) between CT and MR indices in children and adolescents with hip disorders in a much older range (mean age = 15.4±4.1 years).

These studies demonstrate that CT and MRI can be used to understand the 3D deformity of hip dysplasia, although due to logistical constraints (radiation dose in CT, the need for anaesthesia in MRI, and cost and availability for both modalities), this is only possible in a small subgroup of patients. Jia H. et al **[23]**, in a retrospective study concluded that MRI was well suited to detecting barriers to hip reduction in DDH. Rosenbaum, Daniel G. et. al **[24]** note that MRI provides excellent soft tissue contrast without use of ionizing radiation.

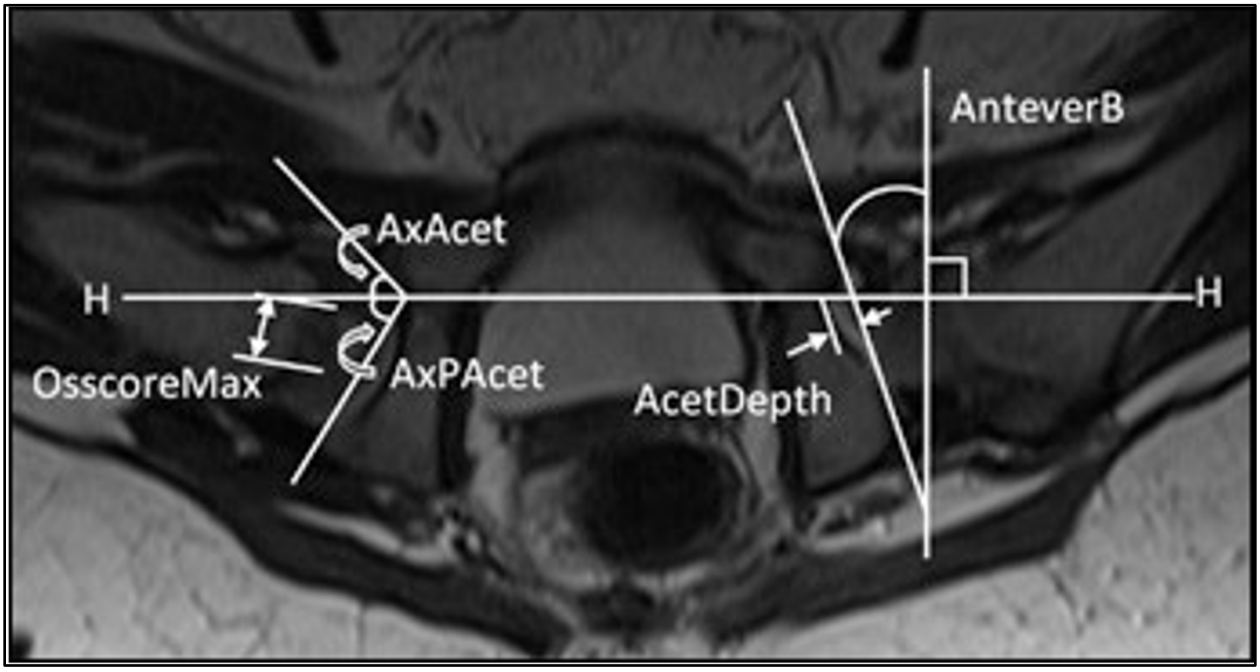


**Figure 1.5** Some of the measurements possible on a hip MRI scan. Indices include axial anterior and posterior acetabular angles (AxAcet, AxPAcet), acetabular bony anteversion (AnteverB) and depth (AcetDepth), and maximum size of the bony ossification centre (OssCoreMax). (source: [21]).

Multiple indices have been developed to assess acetabulum morphology and orientation on cross-sectional imaging, with many measurements possible (Figure 1.5) **[21]**. These can help interpret 3D geometry. Osman et al **[25]** showed that post-reduction MRI-based parameters including anterior acetabular index (AAI), posterior acetabular index (PAI) and abduction angle correlated with persistent acetabular dysplasia in patients who underwent open reduction.

CT is not a primary imaging modality in infantile hip dysplasia since the hip is not ossified yet, but it is primarily useful in operative and re-operative planning to assess angles and lines for optimized 3D correction. Shi et al **[26]** explains that CT scan-based 3-dimensional templating provides the best accuracy for total hip arthroplasty (THA) to treat the cases of neglected DDH. Albers et al **[27]** also discuss the role of CT in preoperative planning for osteotomies for treatment of DDH (Figure 1.6). Tallroth, Kaj, and Jyri Lepistö **[28]**, studied CT-scan of 70 hips from patients who had not been diagnosed with DDH and tried to define normative CT measurements some of which included AA-angle, CE-angle, ACE-angle and AcetAV-angle. Tallroth et al **[29]** compared the accuracy of the measurements of femoral anteversion angle (FAVA) for both 2D and 3D CT scan and concluded that 3D is more accurate than 2D.

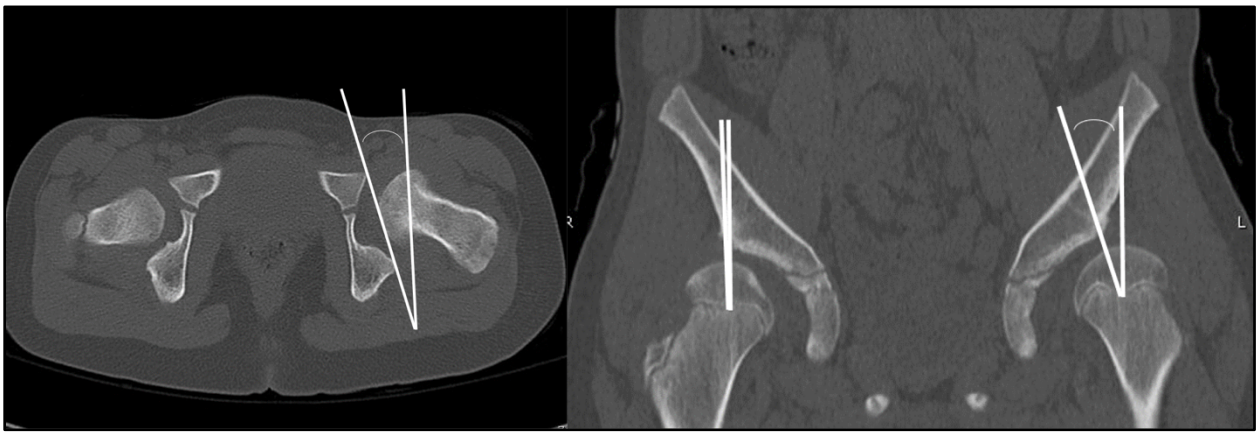

**Figure 1.6** Measurements such as those as in Fig. 1.5 can also be performed on CT, with higher spatial resolution and improved bony detail. CT plays a role particularly in older children to plan surgical correction. (a) Axial CT image from the same patient as in Fig. 1.3, showing the increased bony anteversion at the left hip during planning for a revision osteotomy. (b) Coronal CT image from a 10-year-old girl demonstrating lateral centre-edge angles (LCEA). A decreased LCEA is associated with DDH. In this patient the angle is negative at the left hip (i.e., centre of femoral head is lateral to the acetabular edge), and near zero at the right hip, suggesting dislocation and subluxation respectively.

Hip arthrography is also utilized for DDH as it facilitates the viewing of the cartilaginous part of femoral head and acetabulum. An arthrogram refers to images (X-ray, CT or MRI) of a joint after contrast material is injected into it. Ahmed et al **[30]** in their study of arthrogram in evaluation of closed reduction of DDH, conclude that the reliability of diagnosing hip concentricity in management of DDH by closed reduction is high. Similarly, Grissom et al **[31]** states that "the arthrogram helps to demonstrate the best position of the femur to obtain concentric reduction of the hip". Arthrogram is often done fluoroscopically in infants under anaesthesia, such as just before spica casting (Figure 1.7).

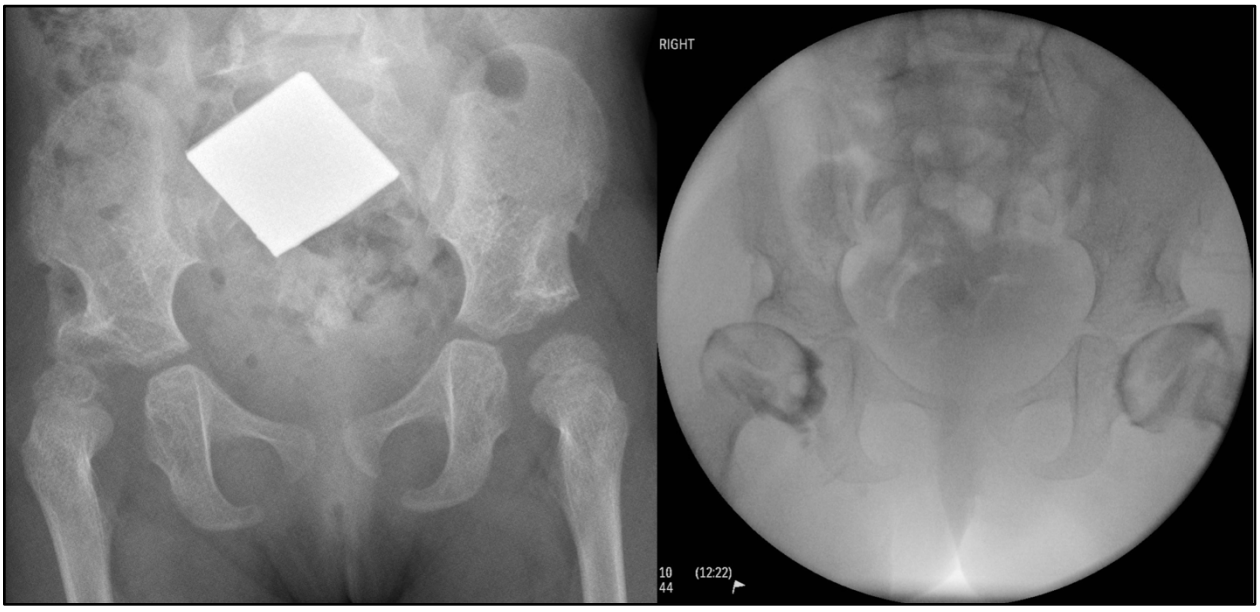


**Figure 1.7** (a) Xray and (b) arthrogram images in a 3-year-old girl with bilateral hip dysplasia. (a) On the X-ray, note irregularity of bilateral acetabular roofs related to prior osteotomies. The right hip appears laterally subluxed and articular surfaces appear irregular. (b) For arthrogram, a surgeon injected contrast material into both hip joints under general anaesthesia. The images outline smoother articular cartilage surfaces than might be expected from the bony contours on Xray, and also demonstrate that joints are better aligned than appreciated on radiographs.

## Ultrasound (2D, 3D)

Since the 1980's, the most commonly used modality in diagnosing DDH in infancy has become ultrasound, which assesses bone and soft tissues with high resolution, high contrast, and the potential for dynamic assessment of stability. The most common analysis of ultrasound images is based on the Graf measurement method, which utilizes measurements performed on a static 'standard plane' 2D coronal image of the mid-hip. Graf classified DDH into several categories **[32, 33, 34]** based on the value of a bone angle (α) and in some cases, a soft tissue angle (β) (Figure 1.8) as well as age to determine subtypes. Graf categorization includes four main classes. **(a)** normal, **(b)** delayed ossification (dysplasia), **(c)** partial dislocation (subluxation), and **(d)** dislocation (total luxation).

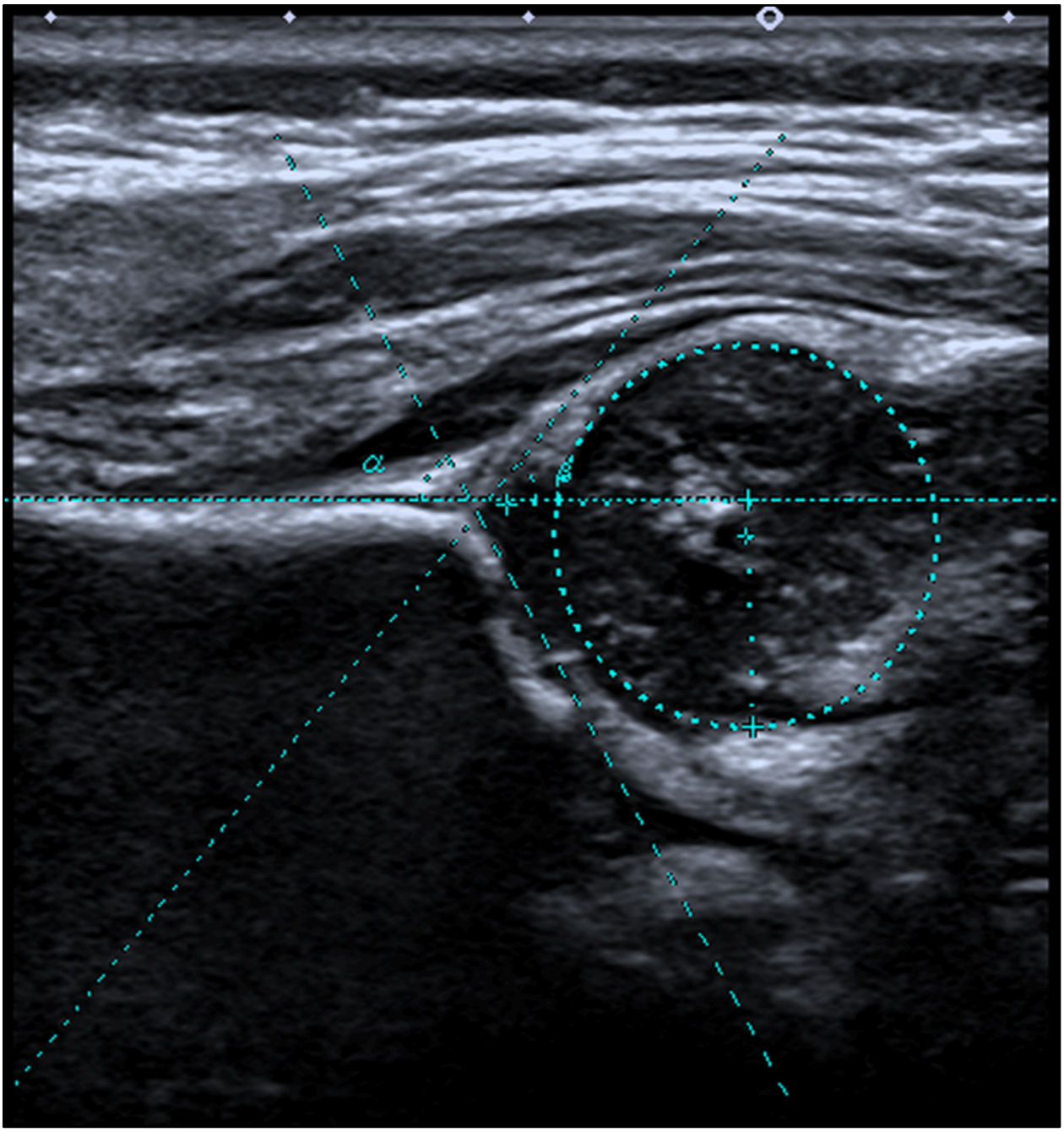


**Figure 1.8** Graf plane, alpha and beta angles.

Although Graf classification of DDH which heavily relies on 2D ultrasonic images is widely accepted and utilized, several studies have shown that it lacks reproducibility (unless acquired and read by experts who have been extensively trained). In 1995, Rosendahl et al **[35]**, reported a high intra-observer agreement (kappa=0.7) but only moderate inter-observer agreement

(kappa=0.5) in the early diagnosis of DDH. Simon et al in another inter-observer article **[36]** state that although no severe cases were missed in their study of 158 US images on classification of DDH based on Graf method, agreement for the classification of normal versus abnormal was only moderate (kappa=0.55). Similar results were reported by Roovers et al **[37]** on reproducibility of US screening examination when read by diagnostic radiographers. They reported a kappa score of 0.65 for differentiating a type I hip versus type IIa~IV but a poor to moderate score of 0.47 for the exact graf classification.

Many have studied the high variability of ultrasound indices of DDH, which can be summarized with two short statements from Orak MM et al **[38]** and Dias JJ et al **[39]**. They respectively concluded that "Sonographic evaluation of the hip appears to vary depending on the investigator" and "Our results showed poor reliability on both counts (inter and intra-observer agreements)". In 2014, Jaremko et al **[40]**, using 3D US (Figure 1.9), showed that alpha angles measured at routine 2D US can vary substantially between 2D scans solely because of changes in probe positioning, up to 19°, which is greater than the size of the Graf classification categories and risks misclassification in up to 50-75% of cases. The display of the full acetabular shape in 3D scans has potential to improve the accuracy of DDH assessment. Mostofi et al **[41]** showed in their reliability study of 2D and 3D ultrasound that novice users after only 1.5 hours of training could acquire hip scans almost as consistently as experts on 3D US (quality score for novice= 4.2±1.0 vs expert=4.9±0.3). The inter-rater reliability was reported poor for 2D US but moderate to high for 3D US.

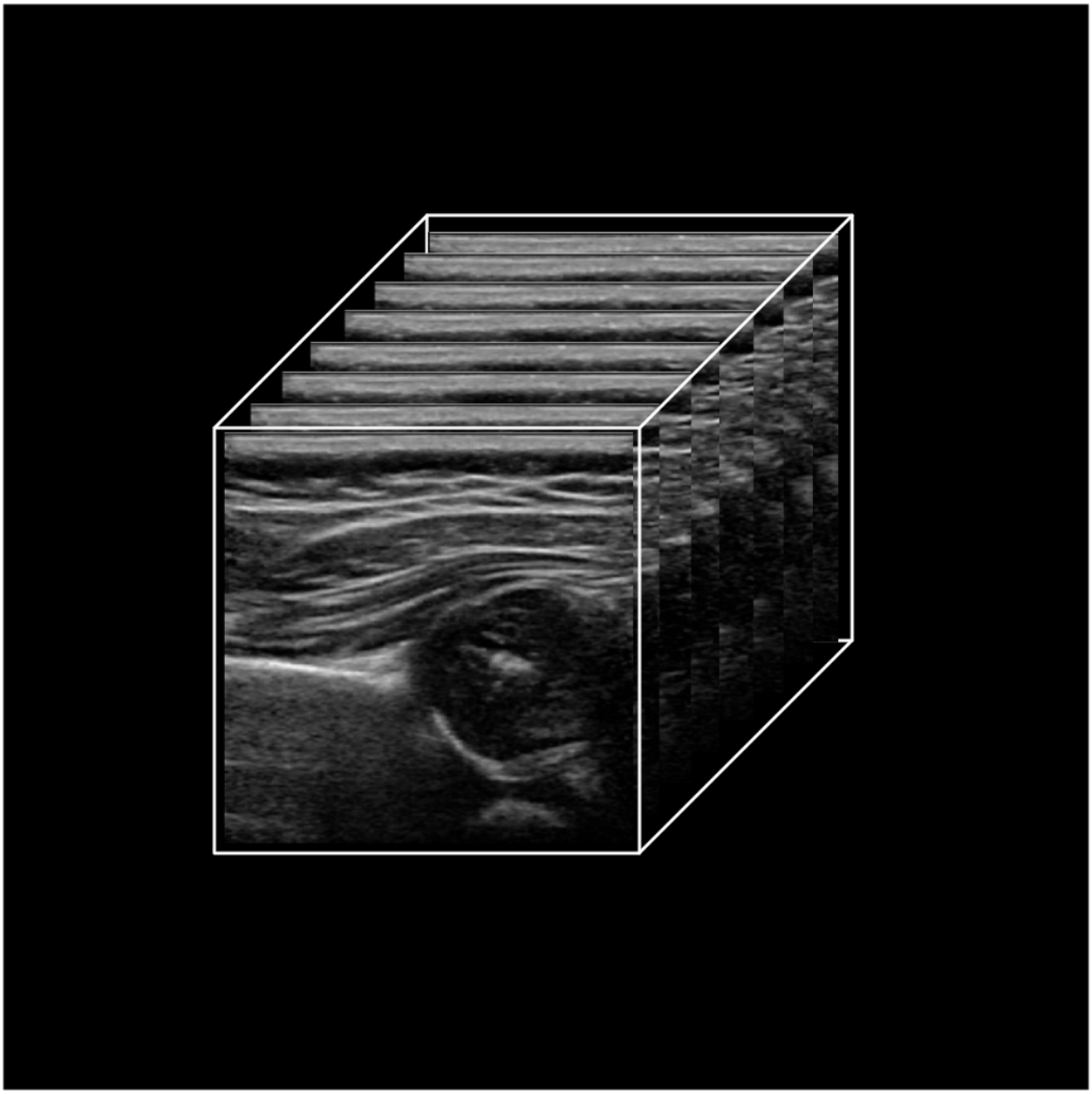

**Figure 1.9** 3D ultrasound: a block of image slices obtained by mechanical probe movement at consistent spacing from each other, allowing evaluation of the hip from anterior to posterior similar to CT or MRI image sequences.

Later, Zonoobi et al **[42]** conducted a multi-center study of DDH diagnosis using 3D ultrasound which confirmed its use could reduce the number of borderline cases which subsequently would have required follow-up imaging by over two-thirds compared to 2D

ultrasound. Quader et al **[43]** later suggested using a new 3D metric for femoral head coverage (FHC3D) based on a tomographic reconstruction of 2D cross-sections. This metric significantly reduced the variability of the 2D-based FHC metric (~20% $p$<0.05). Quader **[44]** also reported much lower test-retest standard deviation for the 3D alpha angle compared to 2D alpha angle.

Despite the improvements in reliability and more comprehensive assessment of the whole hip shape vs. 2D ultrasound, 3D ultrasound is limited in reach since high-resolution linear 3D probes are costly and not routinely available. Manual cine 'sweep' videos obtained with a 2D ultrasound probe may be a more readily achievable surrogate for true 3D ultrasound hip imaging, but this has not been well studied to date.

## Imaging evaluation of Avascular necrosis of the hip (AVN)

AVN of the femoral head is a frequent cause of musculoskeletal disability, causes major diagnostic and therapeutic challenges and is imaged in different ways **[45]**. X-ray is insensitive in depicting AVN **[46]**. It is only effective when the structural damages have already occurred. AVN could be detected intraoperatively using contrast ultrasound. Ntoulia et. al **[47]** explain that ultrasound intraoperative detection of decreased femoral head perfusion aids the surgeon to relocate the hip to less abduction, which prevents irreversible necrosis. Back SJ et al **[48]** noted that CEUS (contrast-enhanced ultrasound) studies in their research all successfully showed blood flow in the femoral epiphysis before and after reduction. Gornitzky et al **[49]** concluded that a perfusion MRI performed immediately after closed reduction of DDH could identify reduced blood flow, potentially reducing the incidence of avascular necrosis after such treatment. Tiderius C. et al **[50]** in a retrospective study agreed that gadolinium-enhanced MRI provides information about femoral head perfusion that may be predictive for future AVN. Contrast-enhanced ultrasound and perfusion MRI are somewhat specialized tests not available in all centres, however.

## Automation and Artificial Intelligence

The wealth of data available in 3D ultrasound also provides enhanced opportunities for automation of image analysis. A semi-automatic method for segmenting (modeling) the acetabulum bone in infant hips was introduced by Hareendranathan et al **[51]** in 2016. Their method was accurate within 1 voxel. Houssam El-Hariri **[52]**, trained a 3D-U-Net neural network **[53]** to automatically segment the pelvis bone surfaces in neonatal hip 3D-US. In 2016, Golan et al **[54]** reported promising results comparing their novel usage of convolutional networks to segment a 3D hip US image to classify them using Graf metrics. Zhang et al **[55]** added a region of interest (ROI) layer to a Fully Convolutional Network (FCN) as a new pipeline to segment the acetabulum from 3DUS images. Tang et al **[56]** evaluated segmentation-by-detection for the same task, improving on the previous 3D U-Net. In 2020, the United States Food and Drug Administration approved MEDO Hip, an AI-powered commercial application which processes 2D or 3D US images and suggests Graf DDH diagnostic categorization. Very recently, El-Hariri et al **[57]** trained and tested the performance of a 3D-U-Net on a dataset of 136 volumes (3D US) and achieved a Dice score of 85% segmenting the pelvis bone surface. They discuss that their model outperformed other methods of segmentation for both pelvis bone surface and femoral head. These tools allow computers to automatically detect the acetabulum much the way our smartphone cameras detect faces. They have not yet been tested in large-scale clinical trials.

## Concerns, standards, and recommendations

Although many different methods of imaging by various modalities for DDH are introduced, appropriateness of imaging for deciding on the treatment is significantly important due to valid concerns on overtreatment or the radiation harms.

American College of Radiology (ACR) in their appropriateness criteria for DDH-Child **[58]** states that "the potential benefits of early diagnosis and treatment must be weighed against the risk of overtreatment and potential for iatrogenic complications.". For similar reasons, the American Academy of Pediatrics (AAP) recommends usage of US between 4 to 6 weeks of age **[59]**, and the American Academy of Orthopaedic Surgeons (AAOS) recommends pediatric orthopedic referral before 4 weeks of age **[60]**. ACR states that after the ossification, pelvic radiography is the preferred imaging modality (4 to 6 months). American Journal of Roentgenology (AJR) in their general imaging review **[61]** mentions that CT is primarily used for management, typically in the postoperative period and is currently used infrequently due to ionizing radiation harms. AJR continues that MRI is more and more utilized for treatment planning and monitoring.

## Discussion

In this review we have taken a historical perspective to assess the role of imaging in assessment of developmental dysplasia of the hip. Unlike cardiovascular disease, where death or myocardial infarction are indisputable endpoints, hip dysplasia is notoriously difficult to reliably define, and there are seemingly as many diagnostic criteria and examination systems as there are imaging modalities and investigators. The key risk factors clearly include ethnicity, female sex and breech presentation during pregnancy, but underlying mechanisms for development of dysplasia are incompletely understood. There is a fairly consistent historical base rate of frankly dislocated hips from severe dysplasia, but the overall incidence of DDH depends strongly on how it is diagnosed, with ultrasound tending to identify a higher proportion of hips as dysplastic than clinical examination. In regions where ultrasound is routinely used, rates of surgery for late-

presenting hip dysplasia are lower than elsewhere **[62]**, implying that there is some benefit to identifying DDH by imaging in infancy.

X-ray is a traditional modality to evaluate DDH but its role is quite limited in the infant period. CT and MRI more fully assess 3D anatomy but are impractical for routine assessment of large numbers of infants. The Graf method in diagnosing and classifying DDH from ultrasound images is well studied but lacks reliability, with the limitations and drawbacks of 2D ultrasound highlighted by more recent work using 3D ultrasound. The most recent development in DDH imaging is the use of computer science technologies to automate DDH diagnosis.

The effect of advancement of technology on the field of hip dysplasia is strong and complex. Early ultrasound techniques have likely led to over-diagnosis, but recent advances in 3D imaging and automated computer image interpretation could allow hip imaging to be more easily, cost-effectively, and reliably acquired and assessed. Eventually, these advancements could facilitate large multi-center studies to enhance our understanding of the 3D deformity and prognosis of hip dysplasia, and to allow cost-effective broad population screening to reduce the burden of disability and pain from osteoarthritis due to hip dysplasia.

## References.

**[1]** Hip Dysplasia, Understanding and Treating Instability of the Native Hip, Paul E. Beaulé Editor

**[2]** Dupuytren G. Original or congenital displacement of the heads OF THIGH-bones. Clin Orthop Relat Res. 1964;33:3–8.

**[3]** Klisic PJ. Congenital dislocation of the hip--a misleading term: brief report. J Bone Joint Surg Br. 1989 Jan;71(1):136. doi: 10.1302/0301-620X.71B1.2914985. PMID: 2914985.

**[4]** Seringe R, Bonnet J-C, Katti E. Pathogeny and natu- ral history of congenital dislocation of the hip. Orthop Traumatol Surg Res. 2014;100(1):59–67. https://doi. org/10.1016/j.otsr.2013.12.006.

**[5]** Bialik V, Bialik GM, Blazer S, Sujov P, Wiener F, Berant M. Developmental dysplasia of the hip: a new approach to incidence. Pediatrics. 1999 Jan;103(1):93-9. doi: 10.1542/peds.103.1.93. PMID: 9917445.

**[6]** BARLOW TG. EARLY DIAGNOSIS AND TREATMENT OF CONGENITAL DISLOCATION OF THE HIP. Proc R Soc Med. 1963 Sep;56(9):804-6. PMID: 14080075; PMCID: PMC1897214.

**[7]** Barlow TG. Congenital dislocation of the hip. Early diagnosis and treatment. Lond Clin Med J. 1964;5:47–58.

**[8]** Ortolani M. Congenital hip dysplasia in the light of early and very early diagnosis. Clin Orthop Relat Res. 1976;119:6–10.

**[9]** Weinstein SL, Mubarak SJ, Wenger DR. Fundamental concepts of developmental dysplasia of the hip. Instr Course Lect. 2014;63:299–305.

**[10]** Tönnis D. Normal values of the hip joint for the evaluation of X-rays in children and adults. Clin Orthop Relat Res. 1976 Sep;(119):39-47. PMID: 954321.

**[11]** Noordin S, Umer M, Hafeez K, Nawaz H. Developmental dysplasia of the hip. Orthop Rev (Pavia). 2010 Sep 23;2(2):e19. doi: 10.4081/or.2010.e19. PMID: 21808709; PMCID: PMC3143976.

**[12]** Narayanan, Unni MBBS, MSc, FRCS(S)*; Mulpuri, Kishore MBBS, MS (Ortho), MHSc(Epi)†; Sankar, Wudbhav N. MD‡; Clarke, Nicholas M.P. ChM, FRCS, FRCS.Ed§; Hosalkar, Harish MBBS, MD‖; Price, Charles T. MD, FAAP¶ International Hip Dysplasia Institute Reliability of a New Radiographic Classification for Developmental Dysplasia of the Hip, Journal of Pediatric Orthopaedics: July/August 2015 - Volume 35 - Issue 5 - p 478-484 doi: 10.1097/BPO.0000000000000318

**[13]** Akiyama M, Nakashima Y, Fujii M, Sato T, Yamamoto T, Mawatari T, Motomura G, Matsuda S, Iwamoto Y. Femoral anteversion is correlated with acetabular version and coverage in Asian women with anterior and global deficient subgroups of hip dysplasia: a CT study. Skeletal Radiol. 2012 Nov;41(11):1411-8. doi: 10.1007/s00256-012-1368-7. Epub 2012 Feb 13. PMID: 22327395.

**[14]** Nepple, Jeffrey J. MD1,a; Wells, Joel MD, MPH1; Ross, James R. MD2; Bedi, Asheesh MD3; Schoenecker, Perry L. MD1; Clohisy, John C. MD1 Three Patterns of Acetabular Deficiency Are Common in Young Adult Patients With Acetabular Dysplasia, Clinical Orthopaedics and Related Research: April 2017 - Volume 475 - Issue 4 - p 1037-1044 doi: 10.1007/s11999-016-5150-3

**[15]** Fujii M, Nakashima Y, Sato T, Akiyama M, Iwamoto Y. Acetabular tilt correlates with acetabular version and coverage in hip dysplasia. Clin Orthop Relat Res. 2012 Oct;470(10):2827-35. doi: 10.1007/s11999-012-2370-z. Epub 2012 Apr 28. PMID: 22544668; PMCID: PMC3441999.

**[16]** Hartofilakidis G, Stamos K, Ioannidis TT. Low friction arthroplasty for old untreated congenital dislocation of the hip. J Bone Joint Surg Br. 1988;70(2):182–6.

**[17]** Hartofilakidis G, Yiannakopoulos CK, Babis GC. The morphologic variations of low and high hip disloca- tion. Clin Orthop Relat Res. 2008;466(4):820–4. Published online 2008 Feb 21. https://doi.org/10.1007/ s11999-008-0131-9.

**[18]** Hartofilakidis G, Stamos K, Karachalios T, Ioannidis TT, Zacharakis N. Congenital hip disease in adults. Classification of acetabular deficiencies and opera- tive treatment with acetabuloplasty combined with total hip arthroplasty. J Bone Joint Surg Am. 1996;78(5):683–92.

**[19]** Wilkin GP, Ibrahim MM, Smit KM, Beaulé PE. A Contemporary Definition of Hip Dysplasia and Structural Instability: Toward a Comprehensive Classification for Acetabular Dysplasia. J Arthroplasty. 2017 Sep;32(9S):S20-S27. doi: 10.1016/j.arth.2017.02.067. Epub 2017 Mar 3. PMID: 28389135.

**[20]** Wells J, Nepple JJ, Crook K, Ross JR, Bedi A, Schoenecker P, Clohisy JC. Femoral Morphology in the Dysplastic Hip: Three-dimensional Characterizations With CT. Clin Orthop Relat Res. 2017 Apr;475(4):1045-1054. doi: 10.1007/s11999-016-5119-2. PMID: 27752989; PMCID: PMC5339134.

**[21]** Jaremko JL, Wang CC, Dulai S. Reliability of indices measured on infant hip MRI at time of spica cast application for dysplasia. Hip Int. 2014 Jul-Aug;24(4):405-16. doi: 10.5301/hipint.5000143. Epub 2014 May 30. PMID: 24970320.

**[22]** Hesham K, Carry PM, Freese K, Kestel L, Stewart JR, Delavan JA, Novais EN. Measurement of Femoral Version by MRI is as Reliable and Reproducible as CT in Children and Adolescents With Hip Disorders. J Pediatr Orthop. 2017 Dec;37(8):557-562. doi: 10.1097/BPO.0000000000000712. PMID: 28323254; PMCID: PMC5368029.

**[23]** Jia, H., Wang, L., Chang, Y. et al. Assessment of irreducible aspects in developmental hip dysplasia by magnetic resonance imaging. BMC Pediatr 20, 550 (2020). https://doi.org/10.1186/s12887-020-02420-2

**[24]** Rosenbaum, Daniel G., et al. "MR Imaging in Postreduction Assessment of Developmental Dysplasia of the Hip: Goals and Obstacles." RadioGraphics, no. 3, Radiological Society of North America (RSNA), May 2016, pp. 840–54. Crossref, doi:10.1148/rg.2016150159.

**[25]** Onaç O, Alpay Y, Yapıcı F, Bayhan Aİ. Correlation of postoperative magnetic resonance image measurements with persisting acetabular dysplasia in open reduction of developmental hip dysplasia. Jt Dis Relat Surg. 2021;32(2):461-467. doi: 10.52312/jdrs.2021.48. Epub 2021 Jun 11. PMID: 34145825; PMCID: PMC8343841.

**[26]** Shi XT, Li CF, Cheng CM, Feng CY, Li SX, Liu JG. Preoperative Planning for Total Hip Arthroplasty for Neglected Developmental Dysplasia of the Hip. Orthop Surg. 2019 Jun;11(3):348-355. doi: 10.1111/os.12472. Epub 2019 Jun 13. PMID: 31197911; PMCID: PMC6595139.

**[27]** Albers CE, Rogers P, Wambeek N, Ahmad SS, Yates PJ, Prosser GH. Preoperative planning for redirective, periacetabular osteotomies. J Hip Preserv Surg. 2017 Sep 14;4(4):276-288. doi: 10.1093/jhps/hnx030. PMID: 29250336; PMCID: PMC5721378.

**[28]** Tallroth, Kaj, and Jyri Lepistö. "Computed Tomography Measurement of Acetabular Dimensions: Normal Values for Correction of Dysplasia." Acta Orthopaedica, no. 4, Informa UK Limited, Jan. 2006, pp. 598–602. Crossref, doi:10.1080/17453670610012665.

**[29]** Shalaby, Mennatallah Hatem, et al. "CT Measurement of Femoral Anteversion Angle in Patients with Unilateral Developmental Hip Dysplasia: A Comparative Study between 2D and 3D Techniques." The Egyptian Journal of Radiology and Nuclear Medicine, no. 3, Springer Science and Business Media LLC, Sept. 2017, pp. 639–43. Crossref, doi:10.1016/j.ejrnm.2017.02.007.

**[30]** A. Ahmed, Amin, and Mohie El Din Fadel. "Role of Intraoperative Arthrogram in Decision Making of Closed versus Medial Open Reduction of Developmental Hip Dysplasia." International Journal of Research in Orthopaedics, no. 6, Medip Academy, Oct. 2019, p. 1037. Crossref, doi:10.18203/issn.2455-4510.intjresorthop20194200.

**[31]** Grissom L, Harcke HT, Thacker M. Imaging in the surgical management of developmental dislocation of the hip. Clin Orthop Relat Res. 2008 Apr;466(4):791-801. doi: 10.1007/s11999-008-0161-3. Epub 2008 Feb 21. PMID: 18288547; PMCID: PMC2504666.

**[32]** Graf R. Classification of hip joint dysplasia by means of sonography. Arch Orthop Trauma Surg. 1984;102(4):248–55.

**[33]** Graf R. Fundamentals of sonographic diagno- sis of infant hip dysplasia. J Pediatr Orthop. 1984;4(6):735–40.

**[34]** Graf R. Ultrasonography-guided therapy. Orthopade. 1997;26(1):33–42.

**[35]** Rosendahl K, Aslaksen A, Lie RT, Markestad T. Reliability of ultrasound in the early diagnosis of developmental dysplasia of the hip. Pediatr Radiol. 1995;25(3):219-24. doi: 10.1007/BF02021541. PMID: 7644309.

**[36]** 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 Med Wkly. 2004 Nov 13;134(45-46):671-7. PMID: 15611889.

**[37]** Roovers EA, Boere-Boonekamp MM, Geertsma TS, Zielhuis GA, Kerkhoff AH. Ultrasonographic screening for developmental dysplasia of the hip in infants. Reproducibility of assessments made by radiographers. J Bone Joint Surg Br. 2003 Jul;85(5):726-30. PMID: 12892198.

**[38]** 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 J Orthop. 2015 Nov-Dec;49(6):610-4. doi: 10.4103/0019-5413.168753. PMID: 26806967; PMCID: PMC4705726.

**[39]** Dias JJ, Thomas IH, Lamont AC, Mody BS, Thompson JR. The reliability of ultrasonographic assessment of neonatal hips. J Bone Joint Surg Br. 1993 May;75(3):479-82. doi: 10.1302/0301-620X.75B3.8496227. PMID: 8496227.

**[40]** Jaremko JL, Mabee M, Swami VG, Jamieson L, Chow K, Thompson RB. 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. 2014 Dec;273(3):870-8. doi: 10.1148/radiol.14140451. Epub 2014 Jun 25. PMID: 24964047.

**[41]** Mostofi, E., Chahal, B., Zonoobi, D. et al. Reliability of 2D and 3D ultrasound for infant hip dysplasia in the hands of novice users. Eur Radiol 29, 1489–1495 (2019). https://doi.org/10.1007/s00330-018-5699-1

**[42]** Zonoobi D, Hareendranathan A, Mostofi E, Mabee M, Pasha S, Cobzas D, Rao P, Dulai SK, Kapur J, Jaremko JL. Developmental Hip Dysplasia Diagnosis at Three-dimensional US: A Multicenter Study. Radiology. 2018 Jun;287(3):1003-1015. doi: 10.1148/radiol.2018172592. Epub 2018 Apr 24. PMID: 29688160.

**[43]** Quader N., Hodgson A.J., Mulpuri K., Cooper A., Abugharbieh R. (2017) A 3D Femoral Head Coverage Metric for Enhanced Reliability in Diagnosing Hip Dysplasia. In: Descoteaux M., Maier-Hein L., Franz A., Jannin P., Collins D., Duchesne S. (eds) Medical Image Computing and Computer Assisted Intervention − MICCAI 2017. MICCAI 2017. Lecture Notes in Computer Science, vol 10433. Springer, Cham. https://doi.org/10.1007/978-3-319-66182-7_12

**[44]** Quader, Niamul. "Automatic Characterization of Developmental Dysplasia of the Hip in Infants Using Ultrasound Imaging." *University of British Columbia*, University of British Columbia, 2018, doi:10.14288/1.0364129.

**[45]** Stoica Z, Dumitrescu D, Popescu M, Gheonea I, Gabor M, Bogdan N. Imaging of avascular necrosis of femoral head: familiar methods and newer trends. Curr Health Sci J. 2009 Jan;35(1):23-8. Epub 2009 Mar 21. PMID: 24778812; PMCID: PMC3945237.

**[46]** Resnick D, Niwayama G. Osteonecrosis: diagnostic techniques, special situations and complications. In: Resnick D, editor. Diagnosis of Bone and Joint Disorders. 3. Philadelphia: WB Saunders Co; 1995. pp. 3495–3558.

**[47]** Ntoulia A, Barnewolt CE, Doria AS, Ho-Fung VM, Lorenz N, Mentzel HJ, Back SJ. Contrast-enhanced ultrasound for musculoskeletal indications in children. Pediatr Radiol. 2021 Mar 30. doi: 10.1007/s00247-021-04964-6. Epub ahead of print. PMID: 33783575.

**[48]** Back SJ, Chauvin NA, Ntoulia A, Ho-Fung VM, Calle Toro JS, Sridharan A, Morgan TA, Kozak B, Darge K, Sankar WN. Intraoperative Contrast-Enhanced Ultrasound Imaging of Femoral Head Perfusion in Developmental Dysplasia of the Hip: A Feasibility Study. J Ultrasound Med. 2020 Feb;39(2):247-257. doi: 10.1002/jum.15097. Epub 2019 Jul 23. PMID: 31334874.

**[49]** Gornitzky AL, Georgiadis AG, Seeley MA, Horn BD, Sankar WN. Does Perfusion MRI After Closed Reduction of Developmental Dysplasia of the Hip Reduce the Incidence of Avascular Necrosis? Clin Orthop Relat Res. 2016 May;474(5):1153-65. doi: 10.1007/s11999-015-4387-6. PMID: 26092677; PMCID: PMC4814438.

**[50]** Tiderius C, Jaramillo D, Connolly S, Griffey M, Rodriguez DP, Kasser JR, Millis MB, Zurakowski D, Kim YJ. Post-closed reduction perfusion magnetic resonance imaging as a

predictor of avascular necrosis in developmental hip dysplasia: a preliminary report. J Pediatr Orthop. 2009 Jan-Feb;29(1):14-20. doi: 10.1097/BPO.0b013e3181926c40. PMID: 19098638.

**[51]** Hareendranathan AR, Mabee M, Punithakumar K, Noga M, Jaremko JL. A technique for semiautomatic segmentation of echogenic structures in 3D ultrasound, applied to infant hip dysplasia. Int J Comput Assist Radiol Surg. 2016 Jan;11(1):31-42. doi: 10.1007/s11548-015-1239-5. Epub 2015 Jun 20. PMID: 26092660.

**[52]** El-Hariri, Houssam. 2020. "Reliable and Robust Hip Dysplasia Measurement with Three-Dimensional Ultrasound and Convolutional Neural Networks." Electronic Theses and Dissertations (ETDs) 2008+. T, University of British Columbia. doi:http://dx.doi.org/10.14288/1.0389533.

**[53]** Çiçek, Özgün et al. "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation." Lecture Notes in Computer Science (2016): 424–432. Crossref. Web.

**[54]** Golan D., Donner Y., Mansi C., Jaremko J., Ramachandran M., on behalf of CUDL (2016) Fully Automating Graf's Method for DDH Diagnosis Using Deep Convolutional Neural Networks. In: Carneiro G. et al. (eds) Deep Learning and Data Labeling for Medical Applications. DLMIA 2016, LABELS 2016. Lecture Notes in Computer Science, vol 10008. Springer, Cham. https://doi.org/10.1007/978-3-319-46976-8_14

**[55]** Z. Zhang, M. Tang, D. Cobzas, D. Zonoobi, M. Jagersand and J. L. Jaremko, "End-to-end detection-segmentation network with ROI convolution," 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), 2018, pp. 1509-1512, doi: 10.1109/ISBI.2018.8363859.

**[56]** 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, pp. 1356-1359, doi: 10.1109/ISBI.2018.8363823.

**[57]** Houssam El-Hariri, Antony J. Hodgson, Kishore Mulpuri, Rafeef Garbi, Automatically Delineating Key Anatomy in 3-D Ultrasound Volumes for Hip Dysplasia Screening, Ultrasound in Medicine & Biology, 2021, ISSN 0301-5629, https://doi.org/10.1016/j.ultrasmedbio.2021.05.011.

**[58]** Expert Panel on Pediatric Imaging: Jie C. Nguyen, MD, MSa ; Scott R. Dorfman, MDb ; Cynthia K. Rigsby, MDc ; Ramesh S. Iyer, MDd ; Adina L. Alazraki, MDe ; Sudha A. Anupindi, MDf ; Dianna M. E. Bardo, MDg ; Brandon P. Brown, MDh ; Sherwin S. Chan, MD, PhDi ; Tushar Chandra, MDj ; Matthew D. Garber, MDk ; Michael M. Moore, MDl ; Nirav K. Pandya, MDm; Narendra S. Shet, MDn ; Alan Siegel, MD, MSo ; Boaz Karmazyn, MD.p, Developmental Dysplasia of the Hip (DDH)–Child. Available at https://acsearch.acr.org/docs/69437/Narrative/. American College of Radiology. Accessed Aug 19, 2021

**[59]** 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.

**[60]** Mulpuri K, Song KM, Goldberg MJ, Sevarino K. Detection and Nonoperative Management of Pediatric Developmental Dysplasia of the Hip in Infants up to Six Months of Age. J Am Acad Orthop Surg 2015;23:202-5.

**[61]** American Journal of Roentgenology. 2014;203: 1324-1335. 10.2214/AJR.13.12449

**[62]** 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 Dec 6;362(9399):1883-7. doi: 10.1016/S0140-6736(03)14957-4. PMID: 14667743.

# CHAPTER 2

# INTER-OBSERVER VARIABILITY OF HIP DYSPLASIA INDICES ON SWEEP ULTRASOUND FOR NOVICES, EXPERTS, AND ARTIFICIAL INTELLIGENCE.

## Introduction

Developmental dysplasia of the hip (DDH) is characterized by a lack of conformity between the femoral head and acetabulum, associated with structural instability and predisposing to osteoarthritis [1]. If diagnosed in early infancy (<6 months) DDH can be treated by non-invasive methods such as Pavlik harness. Ultrasound (US) imaging of the hip, first proposed in 1980 [2], is widely used to diagnose DDH, typically using the bone angle (alpha) and acetabular coverage (d/D) (Figure 2.1) [3]. A complementary index, the beta angle, may be helpful in severely dysplastic hips [4]. Harke et al. also proposed a dynamic technique of hip sonography incorporated motion and stress maneuvers [5]

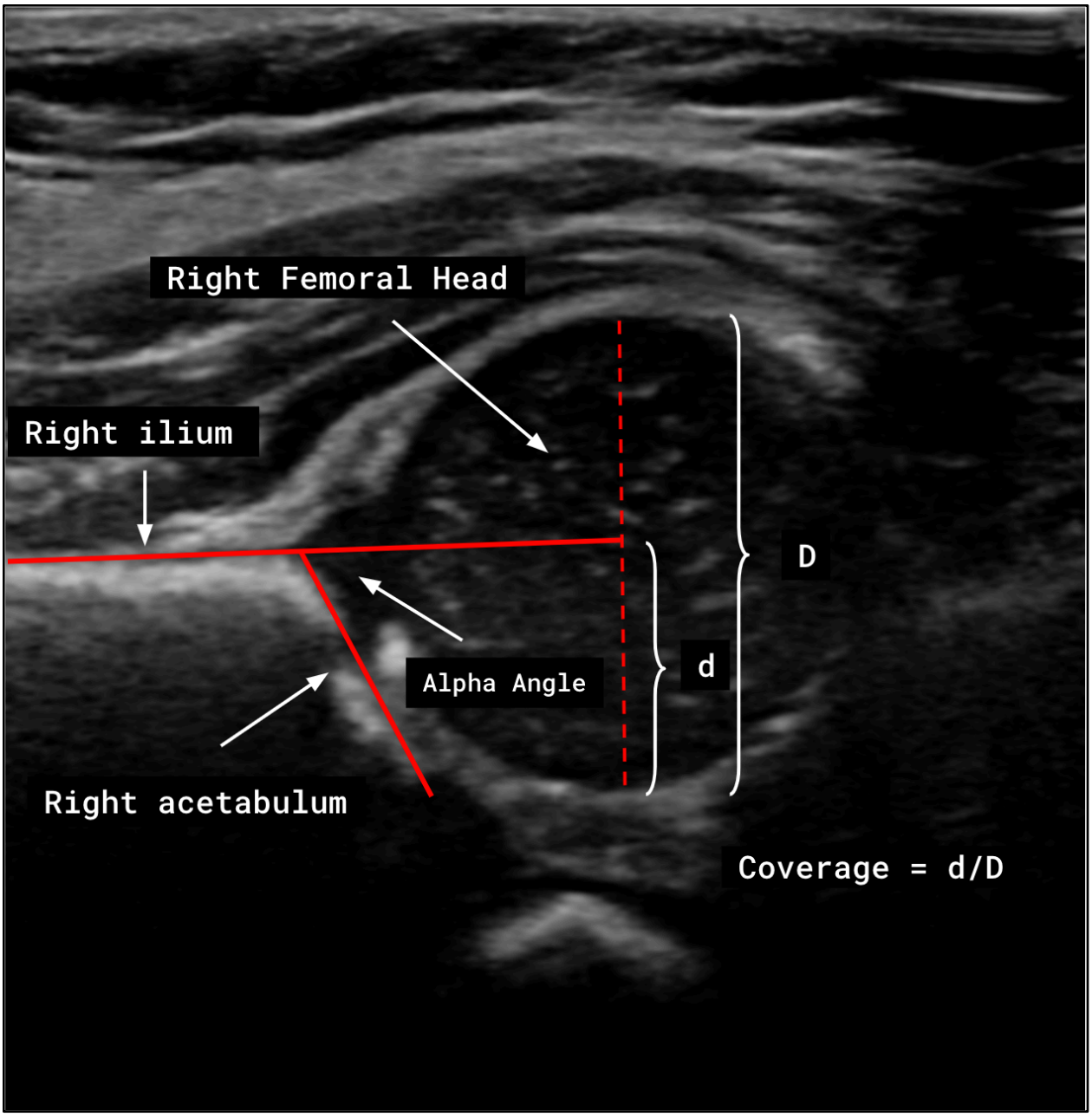


**Figure 2.1** Illustration of the Graf plane used for calculating the alpha angle and coverage for detecting DDH measured on a coronal ultrasound image of the right hip

For the Graf method, a high-quality single 2D ultrasound image is required but can be difficult to acquire reproducibly, risking misdiagnosis, especially for inexperienced users [6]. With advances in technology, sweeps (videos recording the view on the scanner as the user sweeps the

probe through the entire hip) are increasingly being stored. The additional data in sweeps is well-suited to automatic evaluation by artificial intelligence (AI) algorithms.

AI has been used in DDH US in several ways. It can estimate conventional indices or novel 3D indices (promising improved diagnostic accuracy) and minimizes inter-observer variability [7]. A semi-automated segmentation to generate 3D acetabular surface models showed effectiveness in diagnosis of DDH via Graf indices [8]. Several automated approaches to diagnose DDH from hip US have been described, including derivation of a contour alpha angle [9], use of an object localization unit to improve semantic segmentation accuracy [10], automatic identification of suitable 2D US images and extraction of dysplasia metrics [11], and a technique allowing real-time inference for clinical workflow [12].

Although the use of sweeps for DDH US shows promise, there is little data on inter-observer variability for human and AI assessment of sweeps. This study analyzes inter-observer variability of readers with varying experience levels and an AI algorithm in the assessment of infant hip US sweeps for dysplasia. We hypothesized that reliability would be equivalent for analysis of single 2D images vs. sweeps and that AI would perform equivalently to a human reader.

## Methods and Materials

### Images

This study was approved by the University of Alberta Health Research Ethics Board. Imaging was performed at Stollery Children's Hospital, Edmonton, Alberta, Canada, from 2012-2020. All subjects were referred for hip ultrasound based on clinical suspicion of DDH. At the first routine clinical hip 2DUS, written informed consent was obtained from the subject's parents. Images were acquired by people with diverse levels of experience ranging from expert

sonographers to researchers who had been trained to take a hip scan, intended to represent the range seen in a typical clinical practice, from junior to expert sonographers.

We studied one hip in each of 70 male and 170 female patients. The 240 cases were a non-random sample deliberately including the widest possible range of image quality. Using known clinical diagnoses, we included a wide range of hip morphology (normal to severely dysplastic). For each hip we randomly chose whether to use a single coronal plane US image (n=120), or a sweep US (n=120).

Patients underwent US at mean age 61 days (females: 7-267 days, mean=59; males: 4-140 days, mean=66, median=51). We observed clinical care for at least 6 months to classify each imaged hip as Normal (Category 0; 2D n=56, sweep n=59), Borderline, meaning questionably abnormal initially (generally Graf IIa) but with findings that resolved spontaneously at follow-up imaging and clinical examination (Category 1; 2D n=22, sweep n=14), or Dysplastic (requiring treatment by using a Pavlik harness and/or surgery; Category 2; 2D n=38, sweep n=44).

**Image Processing**

Images were analyzed using US FDA cleared third-party software (MEDO Hip) allowing viewing and interpreting of US images (Figure 2.2). This software also facilitated de-identification and transfer of large amounts of data (240 scans made up of 15,912 images) to readers.

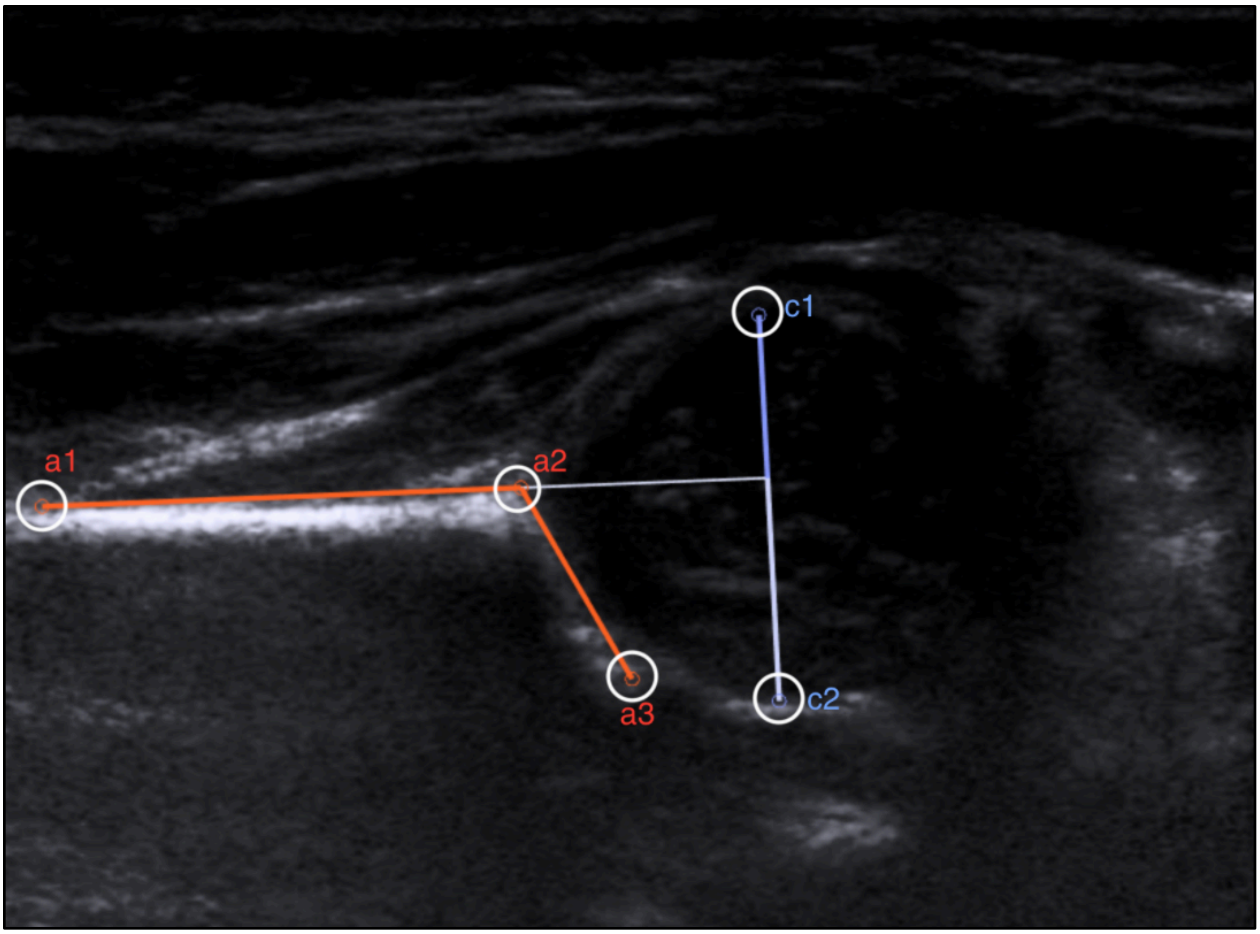


**Figure 2.2** Each reader picked 5 landmarks (circles) to allow the algorithm to determine the alpha angle and the coverage.

## Readers

We recruited 12 readers with different backgrounds and experience levels in reading hip US scans. (Table 2.1).

| Name | Designation | Years of experience | DDH sub specialist |
|---|---|---|---|
| CB | Sonographer | 23 | Yes |
| SK | Sonographer | 19 | Yes |
| JJ | Radiologist | 17 | Yes |
| ND | Radiologist | 15 | No |
| AS | Clinician | 12 | No |
| PS | Radiologist | 12 | No |
| SM | Radiologist | 10 | No |
| DK | Clinician | 9 | No |
| AR | Researcher | 5 | No |
| MM | Researcher | 5 | No |
| EO | Researcher | 4 | No |
| SG | Researcher | 2 | No |

**Table 2.1** Readers list and their years of experience interpreting ultrasound sorted by years of experience.

Readers examined each 2D image using the Graf technique by identifying the 5 landmarks needed to calculate the alpha angle and coverage (Figure 2.2). For sweeps, readers first chose the best slice (Figure 2.3) matching Graf standard plane criteria, before identifying the 5 landmarks. Readers were blinded to the diagnosis of the selected pool of images.

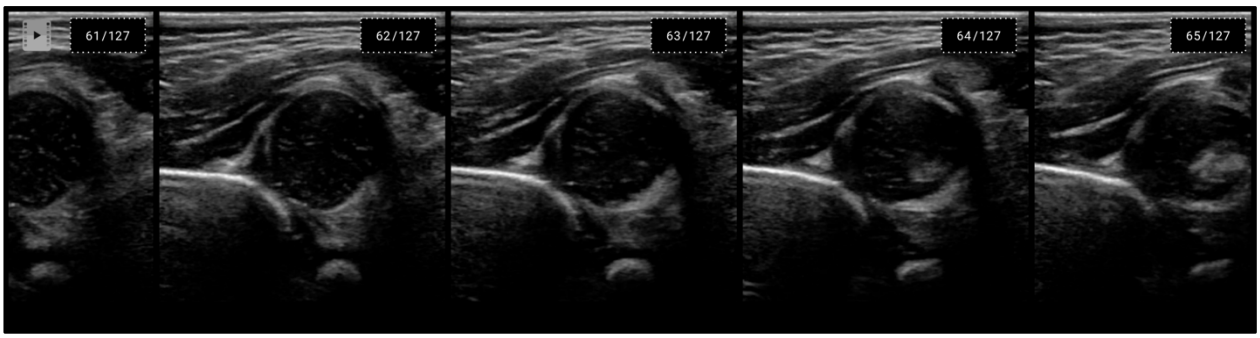


**Figure 2.3** Readers had to choose their preferred slice within the sweep and then make the measurements.

**AI Methods**

An overview of the method is shown in Figure 2.4. We used MEDO Hip to generate an AI-estimated alpha angle and coverage, as though this tool was an additional reader. The AI initially segments the acetabulum and femoral head (Figure 2.5) and then uses it to pick the preferred slice within a sweep. Then it (1) fits Line 1 over the horizontal portion of the acetabular mask; (2) fits Line 2 over the angular portion of the acetabular mask; (3) finds Line 3, the maximal vertical diameter line passing through the femoral head mask (4) uses basic geometry to identify the 5 landmarks.

As shown in Figure 2.4, A CNN is used to obtain segmentation masks of the acetabulum and femoral head. Based on these the alpha angle and coverage for each slice is then calculated. These values were used to train a probabilistic graphical model (PGM) that estimates the probability of dysplasia and classifies each hip as either normal or dysplastic.

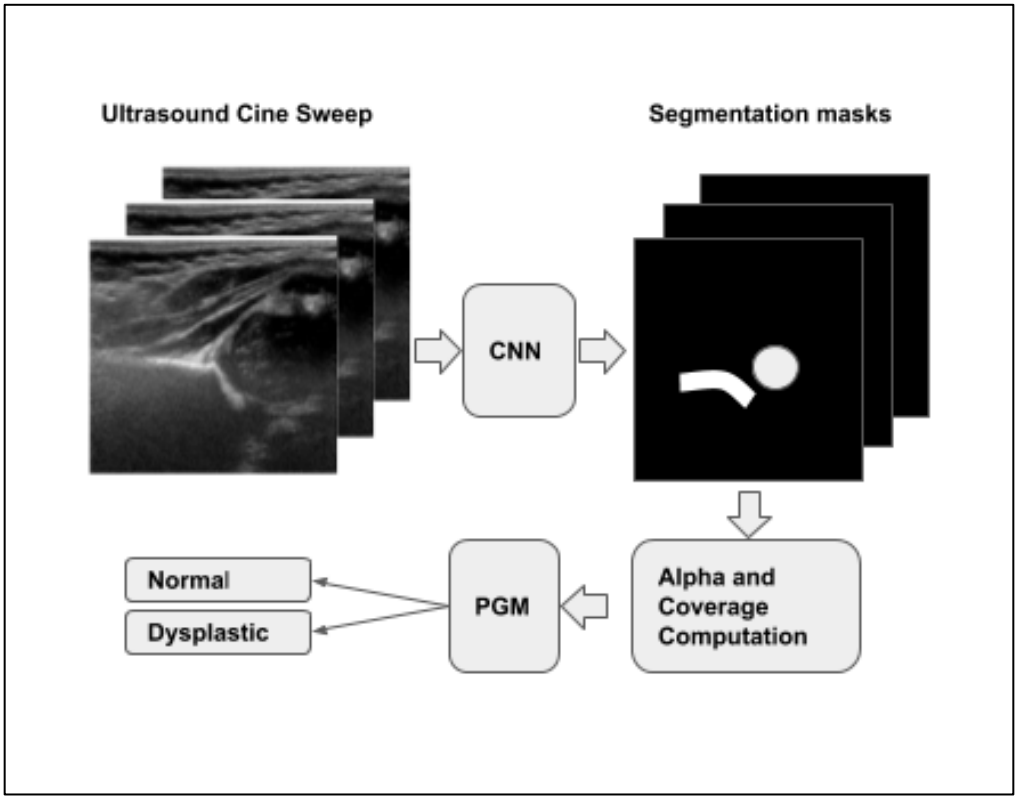

**Figure 2.4** Overview of the proposed approach for AI-augmented DDH diagnosis. Instead of using end-to-end deep learning we propose a more modular approach which computes the alpha angle and coverage to arrive at a diagnosis

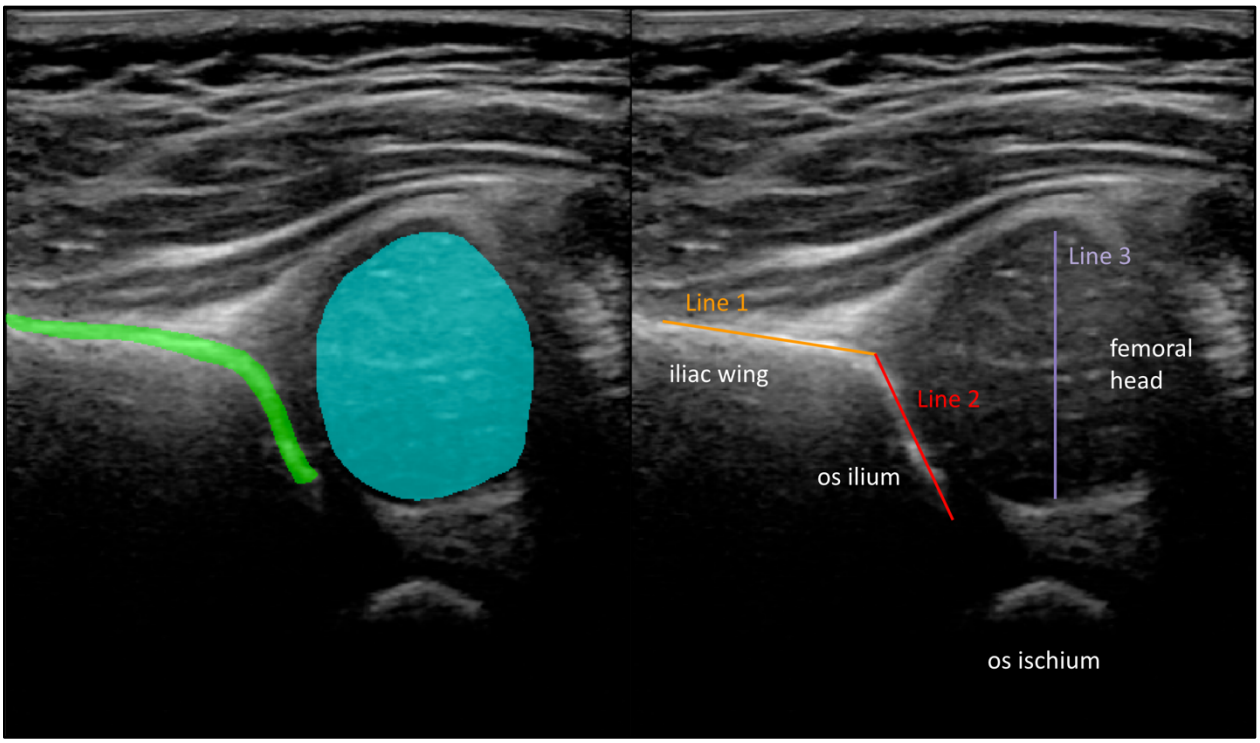


**Figure 2.5** Example of an AI-segmented ultrasound hip (left) and corresponding highlight of the acetabulum and femoral head (right), showing the three lines generated by automated segmentation analysis. Note the slight tilt of the inferior aspect of the iliac wing; the AI computes indices for each image slice in which anatomical landmarks are identifiable, even if these do not quite form a perfect Graf coronal plane image.

The CNN used for segmenting the acetabulum and femoral head was based on a modification of the U-Net architecture. The modified U-Net was trained on a set of 20000 ultrasound image slices. In each of the images the contours of the acetabulum and femoral head were traced by an expert radiologist. The output of the trained U-Net is the mask of the acetabulum

and femoral head which was then used to calculate the alpha angle and coverage. As a post-processing step, a morphological closing operation with a circular structuring element of size 10 pixels was performed to remove smaller objects.

The skeleton S from the post-processed acetabular mask as shown in Figure 2.6 was extracted. The steps involved in computing the alpha angle are described below:

1. Define a baseline lb also shown in Figure 3 connecting leftmost and rightmost pixels in the mask
2. Compute the distance d between each pixel and the baseline lb and determine the farthest point from the line as the apex point a
3. Fit two straight lines la and li, through the acetabular roof and illum passing through the apex a using least square fit
4. Compute the alpha angle (A) based on the slope of lines la and li

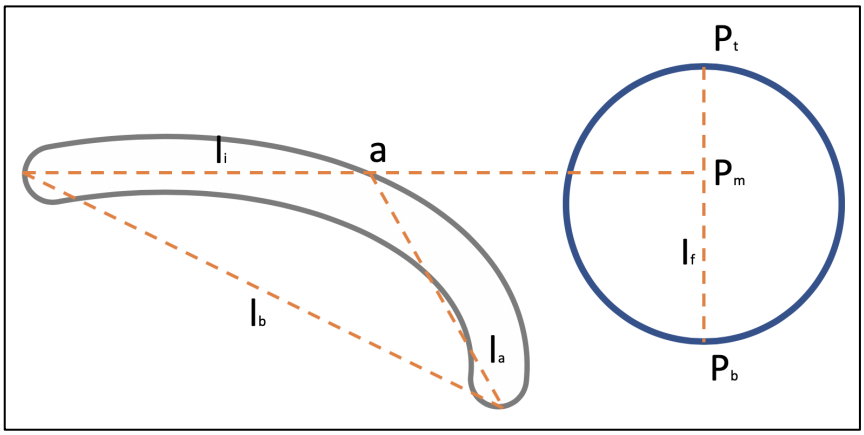


**Figure 2.6** Schematic illustration of computation of alpha angle and coverage

Similarly, the computation of coverage involves the following steps:

1. Define a line lf through the highest point pt and lowest point pb of the femoral head mask
2. Compute the intersection point pm of straight lines lf and li
3. Compute coverage C = |pb-pm|/|pt-pm|

The values of A and C for each slice is analyzed by the PGM to estimate the probability of dysplasia.

The PGM used the values of alpha angle and coverage as inputs to a non-linear logistic regression model. The output of the logistic regression is defined as the conditional probability of not having dysplasia (D) given the angle (A) and coverage(C).

**Statistics**

We compared alpha angle and coverage measurements from all reader pairs, within groups of readers with similar backgrounds and also within groups of similar levels of experience. While computing index measurements, we took the median of the 3 DDH sub-specialist readers' measurements for the alpha angles as the widely accepted standard. Randolph's Kappa [13][14][15] was used for categorical data and intraclass correlation coefficient (ICC(2,1)) for continuously-valued data. The performance of each reader to measure the alpha angle was evaluated against the widely accepted standard using the ICC score. ICC score of each group was also calculated to be compared against other readers groups to analyze the impact of expertise in reading DDH US. DDH sub-specialists were compared with non-DDH-specialist medical imaging experts in diagnosing the DDH US images (kappa). We also compared these two groups of medical imaging experts' performances against the widely accepted standard. The performance of AI was eventually evaluated against widely accepted standard.

Significant differences were identified by the presence of non-overlapping 95% confidence intervals. We illustrated this on a diagram and also calculated the mean difference ± standard deviation for the DDH-sub-specialist imaging experts. We repeated this analysis by adding AI as though it was an additional reader. All statistics were computed using R (version 4.0.2 (2020-06-22))

## Results

Interobserver reliability for alpha angle and coverage, was highest on 2D images (Table 2.2). Interobserver reliability decreased slightly for sweeps. For both 2D and sweeps, sonographers had the highest inter-observer reliability (ICC=0.95~0.98 for alpha angles and 0.90~0.93 for coverage) (Table 2.2).

| | | Single Images | | Sweep Images | |
|---|---|---|---|---|---|
| | Readers | Alpha | Coverage | Alpha | Coverage |
| Researchers | 4 | **0.85**<br>0.81 ~ 0.89 | **0.94**<br>0.92 ~ 0.96 | **0.84**<br>0.80 ~ 0.87 | **0.91**<br>0.88 ~ 0.93 |
| Sonographers | 2 | **0.95**<br>0.93 ~ 0.96 | **0.98**<br>0.97 ~ 0.98 | **0.90**<br>0.87 ~ 0.92 | **0.93**<br>0.91 ~ 0.95 |
| Radiologists | 4 | **0.91**<br>0.85 ~ 0.94 | **0.94**<br>0.92 ~ 0.96 | **0.80**<br>0.73 ~ 0.86 | **0.88**<br>0.85 ~ 0.91 |
| Clinicians | 2 | **0.88**<br>0.85 ~ 0.91 | **0.89**<br>0.85 ~ 0.92 | **0.77**<br>0.69 ~ 0.83 | **0.85**<br>0.81 ~ 0.89 |

**Table 2.2** Inter-observer reliability for indices of hip dysplasia, expressed as ICC(2,1) and its 95% confidence interval, for different groups of readers.

We compared each individual reader rating (including AI) with the clinical diagnosis (Table 2.3) Agreement ranged from fair to good by criteria of Landis and Koch [16]; AI showed moderate agreement with clinical diagnosis (Kappa 0.47-0.49).

| | Single Images | Sweep Images |
|---|---|---|
| Researcher 1 | **0.67**<br>0.57 ~ 0.78 | **0.43**<br>0.30 ~ 0.56 |
| Researcher 2 | **0.5**<br>0.37 ~ 0.63 | **0.39**<br>0.26 ~ 0.52 |
| Researcher 3 | **0.59**<br>0.46 ~ 0.70 | **0.56**<br>0.45 ~ 0.68 |
| Researcher 4 | **0.44**<br>0.31 ~ 0.56 | **0.26**<br>0.13 ~ 0.41 |
| Clinician 1 | **0.42**<br>0.29 ~ 0.55 | **0.21**<br>0.08 ~ 0.35 |
| Clinician 2 | **0.61**<br>0.49 ~ 0.72 | **0.28**<br>0.15 ~ 0.41 |
| Radiologist 1 | **0.55**<br>0.42 ~ 0.66 | **0.45**<br>0.32 ~ 0.59 |
| Radiologist 2 | **0.63**<br>0.51 ~ 0.74 | **0.35**<br>0.21 ~ 0.47 |
| Radiologist 3 | **0.56**<br>0.43 ~ 0.69 | **0.41**<br>0.28 ~ 0.53 |
| Radiologist 4 | **0.55**<br>0.42 ~ 0.66 | **0.44**<br>0.29 ~ 0.56 |
| Sonographer 1 | **0.55**<br>0.43 ~ 0.66 | **0.54**<br>0.41 ~ 0.67 |
| Sonographer 2 | **0.56**<br>0.44 ~ 0.68 | **0.37**<br>0.24 ~ 0.51 |
| AI | **0.49**<br>0.36 ~ 0.62 | **0.47**<br>0.36 ~ 0.60 |

**Table 2.3** Kappa scores and 95% CI of individual readers (including AI) vs clinical diagnosis for 2D and sweeps.

We divided the 6 imaging specialists (radiologists and sonographers) into two groups, (1) readers sub-specializing in DDH diagnosis (JJ, SK, CB) and (2) readers less frequently assessing DDH (ND, SM, PS). We considered the median of the 3 DDH sub-specialist raters' readings as the widely accepted standard for each index. We then calculated the ICC scores of each group of raters vs the widely accepted standard (table 2.4). The agreement of radiologists and clinicians with the widely accepted standard, while still high, was significantly poorer for sweeps than 2D images ($p<0.05$). AI performance was not significantly different from radiologists, although AI was more consistent between 2D and sweep images (ICC=0.90 vs. 0.87 for alpha, and 0.91 vs. 0.91 for coverage).

| | | Single Images | | Sweep Images | |
|---|---|---|---|---|---|
| | #Readers | Alpha | Coverage | Alpha | Coverage |
| Researchers | 4 | **0.88**<br>0.85 ~ 0.91 | **0.95**<br>0.93 ~ 0.96 | **0.86**<br>0.83 ~ 0.89 | **0.92**<br>0.90 ~ 0.94 |
| Sonographers | 2 | **0.97**<br>0.96 ~ 0.98 | **0.99**<br>0.98 ~ 0.99 | **0.94**<br>0.92 ~ 0.95 | **0.96**<br>0.95 ~ 0.97 |
| Radiologists | 4 | **0.93**<br>0.88 ~ 0.95 | **0.95**<br>0.94 ~ 0.97 | **0.84**<br>0.78 ~ 0.88 | **0.90**<br>0.87 ~ 0.92 |
| Clinicians | 2 | **0.91**<br>0.89 ~ 0.93 | **0.92**<br>0.90 ~ 0.94 | **0.82**<br>0.77 ~ 0.86 | **0.87**<br>0.84 ~ 0.90 |
| AI | 1 | **0.90**<br>0.85 ~ 0.93 | **0.91**<br>0.68 ~ 0.96 | **0.87**<br>0.83 ~ 0.90 | **0.91**<br>0.88 ~ 0.93 |

**Table 2.4** ICC scores of each group of readers vs human gold standard.

We examined reliability in more detail for the 6 imaging specialists. Considering all images, the group of DDH sub-specialist readers had near-perfect inter-observer agreement for 2D (Kappa=0.80), and substantial agreement for sweeps (Kappa=0.72). Non-sub-specialist imaging

experts showed substantial agreement for 2D (Kappa=0.66) and significantly lower agreement for sweeps (Kappa=0.54) (Figure 2.7).

Looking only at category 0 or 1 (normal or borderline), all 6 imaging readers showed high agreement with clinical diagnosis on 2D images (Kappa=0.76-0.82), but non-sub-specialist imaging experts struggled to interpret the sweeps (Kappa=0.50-0.74). In contrast, when considering only the dysplastic hips (category 2) the sub-specialists' performance remained similar but non-sub-specialists demonstrated poorer agreement for categorization of 2D images than for sweeps (Kappa=0.47 vs. 0.61). (Figure 2.7).

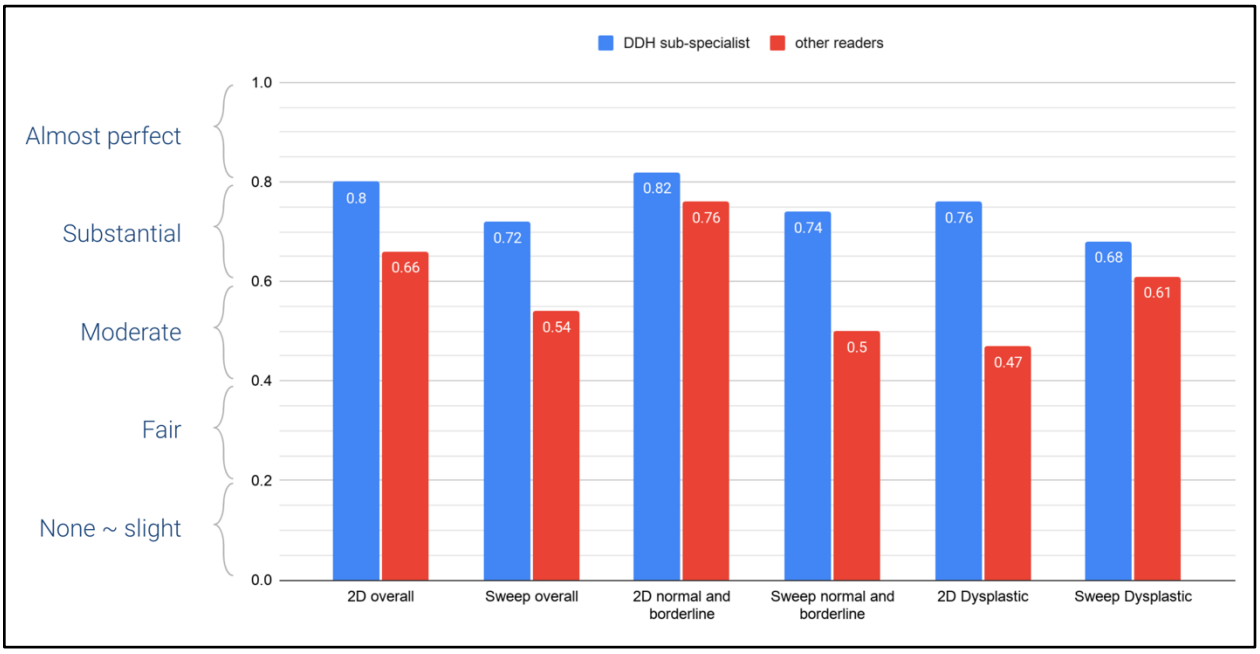


**Figure 2.7** Inter-observer agreement by Randolph's Kappa between the DDH sub-specialists and between the non-sub-specialist medical imaging experts, all images (categories 0, 1, 2)

We observed similar patterns calculating inter-observer variability as ICC vs. the widely accepted standard (Figure 2.8). In the full data set, sub-specialists had slightly higher ICC than the other readers (0.97[0.96~0.98] vs 0.95 [0.94~0.97]) for 2D images; and higher ICC for sweeps (0.94[0.92~0.95] ($p<0.05$) vs. 0.87[0.82~0.90] ($p<0.05$)).

Considering only category 0 (normal) or 1 (borderline), the sub-specialists again showed a slightly higher ICC for alpha angle measurements (0.95[0.93~0.96] vs 0.91[0.88~0.94] for 2D images, and significantly higher for sweeps (0.90[0.86~0.93] vs 0.77[0.70~0.83], $p<0.05$).

For Category 2 (dysplastic), interobserver variability decreased for the sub-specialists, but increased, for the other readers on 2D images (ICC=0.94 [0.91~0.96] vs. 0.92 [0.88~0.95]). On sweeps, sub-specialists had a slightly lower ICC score (0.87 [0.81~0.92]) while non-sub-specialists' score dropped more sharply (0.74 [0.57~0.84], $p<0.05$).

We also compared the reliability of AI as a 'reader' vs the human widely accepted standard (Figure 2.8). AI had substantial to perfect agreement with the widely accepted standard for both 2D and sweeps, performing slightly inferiorly to non-subspecialist human readers on single images but nearly identically to these readers on sweeps.

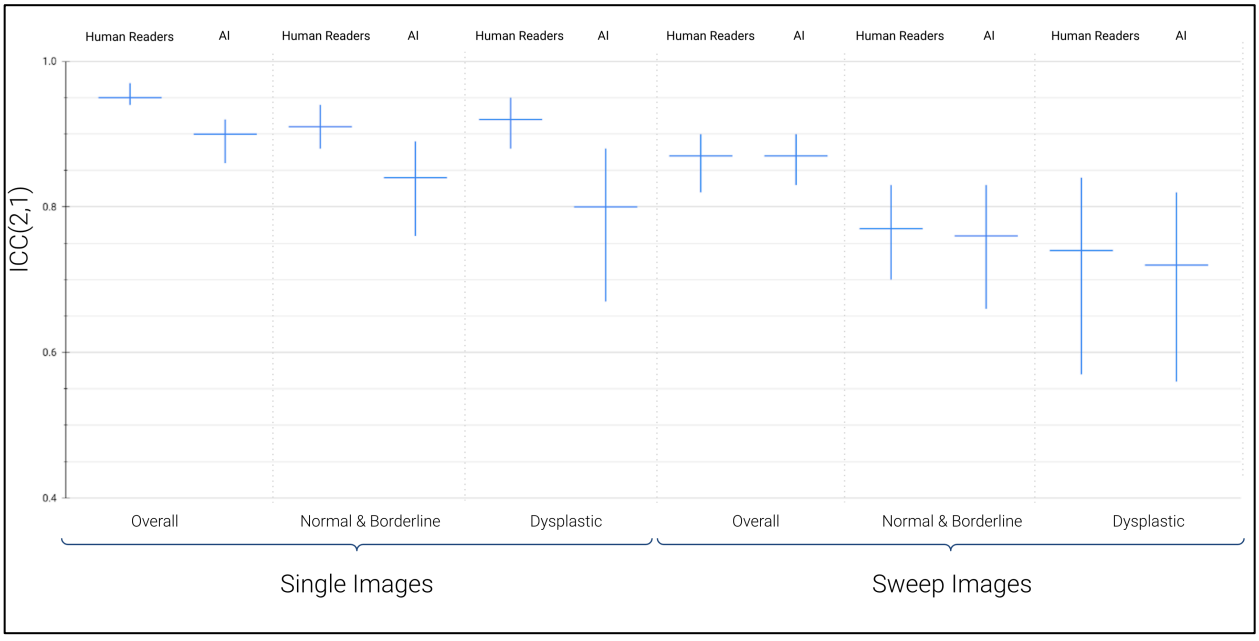


**Figure 2.8** Agreement between the widely accepted standard alpha angle and non DDH sub-specialist imaging experts for different patient groups, as ICC with vertical bars representing the 95% CI. AI had substantial to perfect agreement with the widely accepted standard for both 2D and sweeps, performing

slightly inferiorly to non-subspecialist human readers on single images but nearly identically to these readers on sweeps.

AI showed high and statistically equivalent intraclass correlations to the human expert widely accepted standard for alpha angle measurements across all hips (ICC=0.90 for 2D, 0.87 for sweeps). Variability between AI and widely accepted standard alpha angle measurements trended higher for dysplastic hips than for all hips (Table 2.5).

| | Overall | | Category 1 & 0 | | Category 2 | |
|---|---|---|---|---|---|---|
| | 2D | Sweep | 2D | Sweep | 2D | Sweep |
| ICC(2,1) | **0.90**<br>0.85 ~ 0.93 | **0.87**<br>0.83 ~ 0.90 | **0.84**<br>0.76 ~ 0.89 | **0.76**<br>0.66 ~ 0.83 | **0.80**<br>0.66 ~ 0.88 | **0.72**<br>0.56 ~ 0.82 |

**Table 2.5** Agreement between AI and human gold standard for alpha angle.

We also compared actual case-by-case differences (mean absolute difference) in readings. We calculated the signed differences in alpha angle measurements vs. the widely accepted standard, for each human reader and AI. For 2D images, the largest differences were 1.1° (SD = 2.4°). For AI the average difference was -1.9° (SD=5.7°). For sweeps, the largest average difference for the measurement of alpha angles between human readers was 0.29° (SD=4.2°) while AI average difference was -1.3° (SD=7.2°); (Figures 2.9 and 2.10 and Table 2.6).

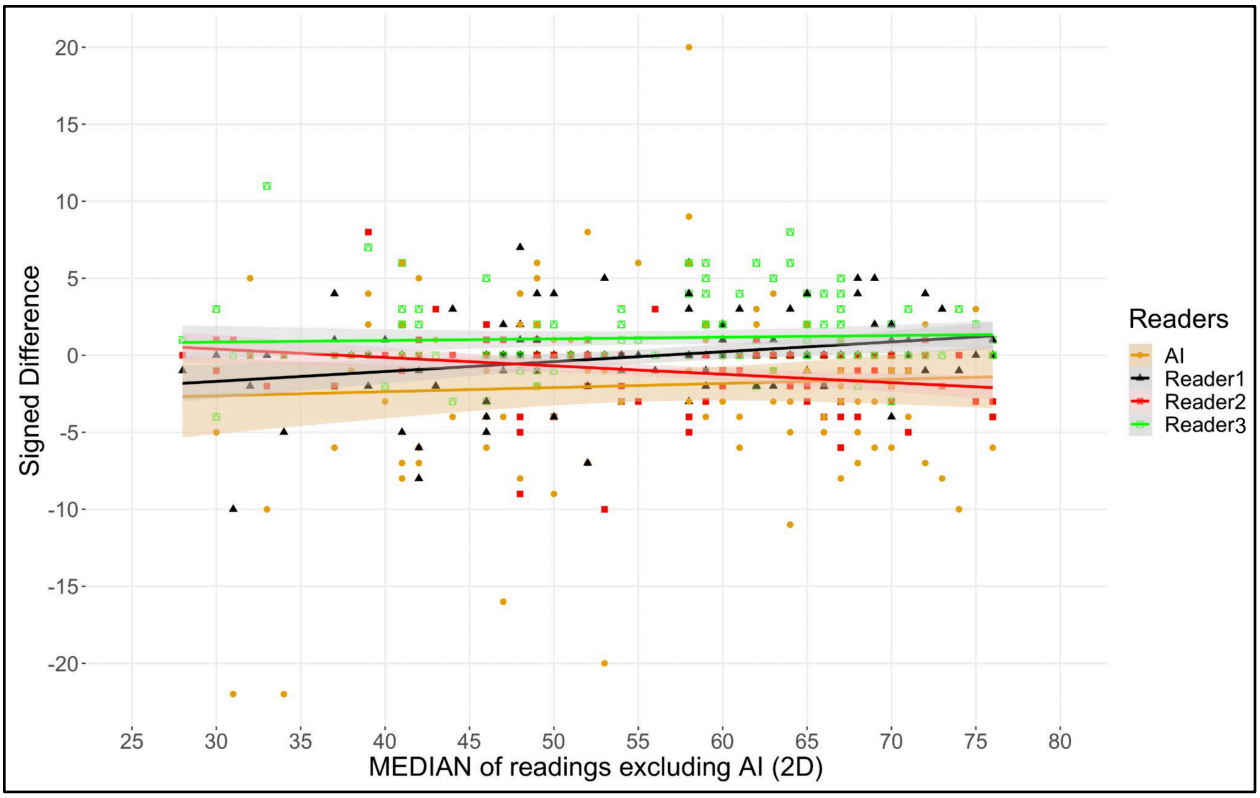


**Figure 2.9** Signed difference between each reader and widely accepted standard for alpha angle for 2D images and their respective 95% confidence intervals

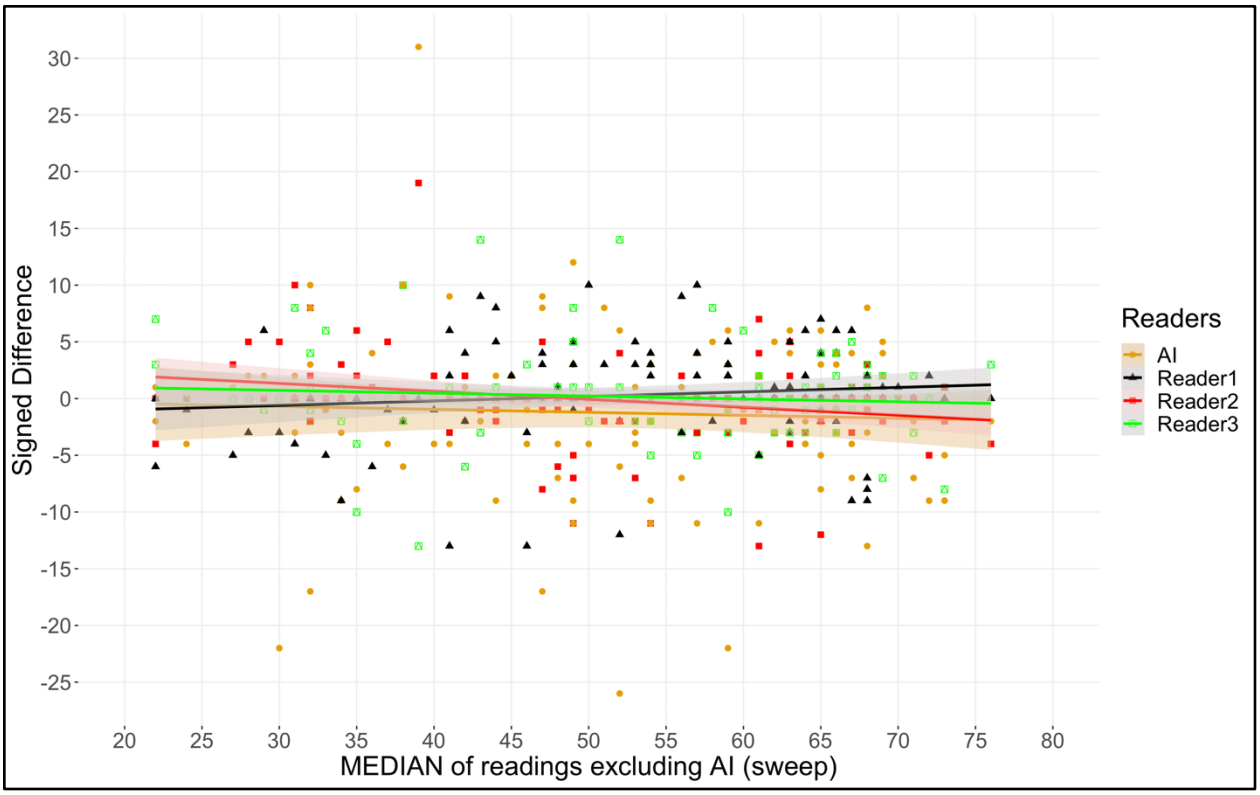


**Figure 2.10** Signed difference between each reader and widely accepted standard for alpha angle for sweeps and their respective 95% confidence intervals

| Difference from gold standard (mean ± SD) | 2D | Sweep |
|---|---|---|
| Reader 1 | -0.035 ± 2.7 | 0.29 ± 4.2 |
| Reader 2 | -1 ± 2.2 | -0.25 ± 3.9 |
| Reader 3 | 1.1 ± 2.4 | 0.15 ± 3.8 |
| AI | -1.9 ± 5.7 | -1.3 ± 7.2 |

**Table 2.6** Signed difference between selected reader values and gold standard for alpha angle (degrees).

When comparing non-subspecialist to subspecialist human readers pairwise, sensitivity for category 2 hips requiring treatment was high (0.81~1.0) but specificity was poor (0.05~0.22 for 2 of the 3 readers, 0.45~0.55 for the other reader). Comparing AI with sub-specialists, sensitivity

was also high (0.83~1), and specificity low but within the range of human non-specialist readers (0.16~0.21).

On a case-by-case review, all outliers where AI readings differed substantially from human observers were observed to be in the lowest-quality images (Figure 2.11).

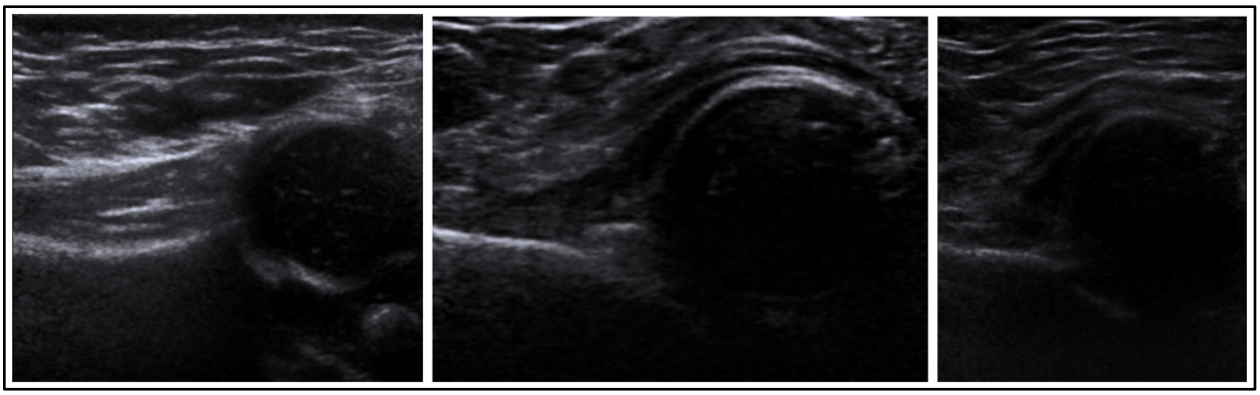

**Figure 2.11** examples of images where AI prediction either failed or was substantially different from the widely accepted standard. These are poor quality images (generally obtained at the beginning of our data collection in 2012-13) which do not meet Graf standard plane criteria and are difficult for human observers and AI to assess.

## Discussion

In this paper we assessed interobserver variability of DDH index measurement and diagnostic classification for 12 human readers and AI on 120 2D US images and 120 sweep US images. We found that variability between readers was wide and was influenced more by the level of reader experience specifically with DDH ultrasound than by the reader's professional specialty. We also found that AI performed equivalently to our human non-subspecialist readers.

Performing conventional 2D hip ultrasound involves the user finding the hip joint, then making fine changes in probe position to optimally approximate the Graf standard plane, saving an optimized 2D image and measuring indices including alpha angle on that image. This is time-consuming with a squirming infant and requires a highly trained and experienced user. With

modern probes it is possible to simply save all the images from an US sweep through the entire hip joint and select the best image for analysis afterward. Such a protocol is attractive for large-scale screening because the images could be acquired by a less-experienced user, more complete hip anatomy is recorded, and image selection and index measurement can be automated using AI. However, the reliability of this approach has not been evaluated previously.

Numerous prior studies have focused on the reliability of making measurements on 2D images. Our variability for 2D images was higher than in a recent study on 798 infants, in which all interobserver agreements were classified as excellent [17]. In another small study 3 experienced readers rated the same set of images twice, pre-, and post-training in Graf classification, finding ICC scores from 0.91~0.94 to 0.97~0.98 after training [18]. These very high ICC are likely the maximum achievable, on high-quality images read by experienced observers. More concordant with our results, a larger study using single images in 210 infants showed only moderate interobserver reliability for the alpha angle (ICC 0.62 overall; 0.47 between two residents, 0.65 between two specialists and 0.68 between two professors) [19]. An earlier study with more observers (22) than images (20) found intraobserver and interobserver agreement ratios on the Graf classification to be 0.65 and 0.51 respectively [20], slightly poorer than in our study. We observed higher reliability than Quader et al, whose systematic review and meta-analysis of 497 articles showed inter-examination and interrater ICC for alpha angle to vary from 0.03 to 0.445 [21]. Inter-examination variability is the most demanding metric because different images are obtained at different times by each observer, rather than simply having different observers interpret the same image or sweep. A retrospective study comparing 12 readers in 15 DDH radiographs using the Tonnis and IHDI classifications found Fleiss Kappa scores to rise with increased user experience [22]. Our study mirrors these findings in a different imaging modality.

Our study is the first we are aware of to assess interobserver variability on ultrasound sweeps. As expected, since readers had an additional task of image selection, reliability was lower on sweeps than 2DUS images, for all readers and AI but by different amounts. Sonographers were not significantly less reliable on sweeps, while radiologists and clinicians were. This may be because sonographers are deeply familiar with the process of selecting the best image as their daily task. Reliability was also more affected by direct user experience with DDH scans than user background; even among imaging professionals there were significant differences in reliability between experienced and less-experienced users. Fortunately, there was still at least moderate reliability of diagnostic classification in all types of images.

Another novel aspect of our study is the comparison of AI with human readers for DDH indices and diagnosis. We found high AI accuracy vs. widely accepted standard, ICC 0.87 for sweeps and 0.90 for single images. AI agreement with clinical diagnosis was moderate (Kappa 0.47 for sweeps, 0.49 for single images), but this represented better performance than achieved by 10/12 human readers on sweeps.

AI performance was intermediate between expert readers and less-experienced human readers on this challenging data set. AI performance would likely be improved by an initial automated pre-screen of images to reject the lowest-quality images from being analyzed [23].

Overall, we observed only slightly lower inter-observer reliability on sweeps than on conventional single 2D ultrasound images for hip dysplasia index measurement and diagnosis and found that an AI package interpreted sweeps similarly to expert non-subspecialist human readers, performing only slightly lower compared to the widely accepted standard. Since sweeps provide more data than single images, are more easily obtained by non-experts, and are amenable to post-scan analysis by human experts or artificial intelligence, these results motivate further study of hip

dysplasia ultrasound using sweeps. Since AI performed intermediate between non-expert and expert readers, one use of AI could be to help readers improve closer to the expert level.

## References

**[1]** Hip Dysplasia, Understanding and Treating Instability of the Native Hip, Paul E. Beaulé Editor

**[2]** Graf, R. "The Diagnosis of Congenital Hip-Joint Dislocation by the Ultrasonic Combound Treatment." Archives of Orthopaedic and Traumatic Surgery, no. 2, Springer Science and Business Media LLC, Sept. 1980, pp. 117–33. Crossref, doi:10.1007/bf00450934.

**[3]** Jacobino, Bruno de Castro Paixão, et al. "Using the Graf Method of Ultrasound Examination to Classify Hip Dysplasia in Neonates." Autopsy and Case Reports, no. 2, Editora Cubo, 2012, pp. 5–10. Crossref, doi:10.4322/acr.2012.018.

**[4]** Falliner, A., et al. "Comparing Ultrasound Measurements of Neonatal Hips Using the Methods of Graf and Terjesen." The Journal of Bone and Joint Surgery. British Volume, no. 1, British Editorial Society of Bone & Joint Surgery, Jan. 2006, pp. 104–06. Crossref, doi:10.1302/0301-620x.88b1.16419.

**[5]** Harke HT, Grissom LE: Performing dynamic sonography of the infant hip. AJR Am J Roentgenol 1990;155:837–844.

**[6]** Jaremko, Jacob L., et al. "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, no. 3, Radiological Society of North America (RSNA), Dec. 2014, pp. 870–78. Crossref, doi:10.1148/radiol.14140451.

**[7]** Shin, YiRang, et al. "Artificial Intelligence in Musculoskeletal Ultrasound Imaging." Ultrasonography, no. 1, Korean Society of Ultrasound in Medicine, Jan. 2021, pp. 30–44. Crossref, doi:10.14366/usg.20080.

**[8]** Golan D., Donner Y., Mansi C., Jaremko J., Ramachandran M., on behalf of CUDL (2016) Fully Automating Graf's Method for DDH Diagnosis Using Deep Convolutional Neural Networks. In: Carneiro G. et al. (eds) Deep Learning and Data Labeling for Medical Applications. DLMIA 2016, LABELS 2016. Lecture Notes in Computer Science, vol 10008. Springer, Cham. https://doi.org/10.1007/978-3-319-46976-8_14

**[9]** Hareendranathan, A. R., D. Zonoobi, M. Mabee, D. Cobzas, K. Punithakumar, M. Noga, and J. L. Jaremko. "Toward Automatic Diagnosis of Hip Dysplasia from 2D Ultrasound." In 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), 982–85, 2017. https://doi.org/10.1109/ISBI.2017.7950680.

**[10]** Zhang, Z., M. Tang, D. Cobzas, D. Zonoobi, M. Jagersand, and J. L. Jaremko. "End-to-End Detection-Segmentation Network with ROI Convolution." In 2018 IEEE 15th International

Symposium on Biomedical Imaging (ISBI 2018), 1509–12, 2018. https://doi.org/10.1109/ISBI.2018.8363859.

**[11]** Quader N, Hodgson AJ, Mulpuri K, Schaeffer E, Abugharbieh R. Automatic Evaluation of Scan Adequacy and Dysplasia Metrics in 2-D Ultrasound Images of the Neonatal Hip. Ultrasound Med Biol. 2017 Jun;43(6):1252-1262. doi: 10.1016/j.ultrasmedbio.2017.01.012. Epub 2017 Mar 22. PMID: 28341489.

**[12]** Paserin O., Mulpuri K., Cooper A., Hodgson A.J., Abugharbieh R. (2017) Automatic Near Real-Time Evaluation of 3D Ultrasound Scan Adequacy for Developmental Dysplasia of the Hip. In: Cardoso M. et al. (eds) Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures. CARE 2017, CLIP 2017. Lecture Notes in Computer Science, vol 10550. Springer, Cham. https://doi-org.login.ezproxy.library.ualberta.ca/10.1007/978-3-319-67543-5_12

**[13]** Randolph, J. J. (2005). Free-marginal multirater kappa: An alternative to Fleiss´ fixed-marginal multirater kappa. Paper presented at the Joensuu University Learning and Instruction Symposium 2005, Joensuu, Finland, October 14-15th, 2005. (ERIC Document Reproduction Service No. ED490661)

**[14]** Warrens, M. J. (2010). Inequalities between multi-rater kappas. Advances in Data Analysis and Classification, 4(4), 271-286. doi:10.1007/s11634-010-0073-4

**[15]** Brennan, R. L., & Prediger, D. J. (1981). Coefficient Kappa: Some uses, misuses, and alternatives. Educational and Psychological Measurement (41)3, 687-699.

**[16]** Landis, J. Richard, and Gary G. Koch. "The Measurement of Observer Agreement for Categorical Data." Biometrics 33, no. 1 (1977): 159-74. Accessed December 18, 2020. doi:10.2307/2529310.

**[17]** Pedrotti, Luisella, Ilaria Crivellari, Alessandro Degrate, Federica De Rosa, Francesca Ruggiero, and Mario Mosconi. “Interpreting Neonatal Hip Sonography: Intraobserver and Interobserver Variability.” Journal of Pediatric Orthopaedics B 29, no. 3 (May 2020): 214–218. https://doi.org/10.1097/BPB.0000000000000670.

**[18]** Yildiz, Kadri, Hayrunnisa BEKİS BOZKURT Bekis, Türkhun Çetin, and Vahit Yildiz. “Interobserver Reliability in the Ultrasonic Evaluation with Graf Method of Developmental Dysplasia of the Hip: The Importance of Education for Ultrasonography Classification.” Journal of Health Sciences and Medicine 3, no. 2 (March 19, 2020): 11–124. https://doi.org/10.32322/jhsm.676820.

**[19]** Karakus, Ozgun, Ozgur Karaman, Ahmet Sinan Sari, Mehmet Mufit Orak, and Hasan Hilmi Muratli. “Is It Difficult to Obtain Inter-Observer Agreement in the Measurement of the Beta Angle in Ultrasound Evaluation of the Paediatric Hip?” Journal of Orthopaedic Surgery and Research 14, no. 1 (July 17, 2019): 221. https://doi.org/10.1186/s13018-019-1263-1.

**[20]** Ömeroglu, Hakan, Ali Biçmoglu, Süha Koparal, and Sinan Seber. “Assessment of Variations in the Measurement of Hip Ultrasonography by the Graf Method in Developmental Dysplasia of the Hip§.” Journal of Pediatric Orthopaedics B 10, no. 2 (April 2001): 89–95.

**[21]** Quader, Niamul, Emily K. Schaeffer, Antony J. Hodgson, Rafeef Abugharbieh, and Kishore Mulpuri. “A Systematic Review and Meta-Analysis on the Reproducibility of Ultrasound-Based Metrics for Assessing Developmental Dysplasia of the Hip.” Journal of Pediatric Orthopaedics 38, no. 6 (July 2018): e305. https://doi.org/10.1097/BPO.0000000000001179.

**[22]** Ismiarto, YD, P Agradi, and ZN Helmi. “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 13, no. 3 (November 2019): 60–65. https://doi.org/10.5704/MOJ.1911.010.

**[23]** Hareendranathan AR, Chahal B, Ghasseminia S, Zonoobi D, Jaremko JL. Impact of scan quality on AI assessment of hip dysplasia ultrasound. J Ultrasound. 2021 Mar 5. doi: 10.1007/s40477-021-00560-4. Epub ahead of print. PMID: 33675031.

# CHAPTER 3 AUTOMATED DIAGNOSIS OF HIP DYSPLASIA FROM 3D ULTRASOUND USING ARTIFICIAL INTELLIGENCE: A TWO-CENTRE MULTI-YEAR STUDY.

## Introduction

Developmental dysplasia of the hip (DDH), characterised by a lack of conformity between the femoral head and acetabulum [1], includes a wide spectrum of abnormalities of the acetabulum and the proximal femur [2]. DDH prevalence varies depending on its definition and is in the range of 28.5 cases per 1000 [3]. If detected before acetabular ossification (<6 months), DDH can very often be corrected noninvasively (e.g. with a Pavlik harness). Delayed or missed diagnosis is associated with, more invasive interventions often including multiple surgeries in childhood and markedly higher rates of poor outcome (ie. early onset end stage osteoarthritis necessitating hip arthroplasty in young adulthood.

Physical examination by Ortolani and Barlow tests can detect DDH in neonates but requires a skilled examiner and has limited sensitivity in older infants and for milder disease [4]. Ultrasonic imaging (US) diagnosis of the hip, originally introduced in 1980 by Graf et al. [5], classifies DDH by measuring the acetabular angle (alpha) and the position of the femoral head (coverage). The Graf method utilizes a single 2D ultrasound image (Graf plane) for classification (Figure 3.1). However, the diagnosis is susceptible to errors due to inaccurate orientation of the probe, seen more often with less experienced users [6]

To improve reliability and minimize overdiagnosis [3], three-dimensional (3D) US has been used [7]. 3D US provides a more comprehensive view of the hip by imaging the whole joint from anterior to posterior. This not only aids human readers to more reliably assess the scans [8], it also facilitates automation of diagnosis by artificial intelligence (AI) algorithms.

AI has been used in DDH diagnosis in several ways, including automatic segmentation of the bony acetabulum and derivation of geometric indices of hip [9]; computing the alpha angle by convolutional neural network [10] and enhancement of the segmentation accuracy of fully convolutional networks by incorporating a localization unit [11]. In 2020, a commercial AI application (MEDO Hip) was approved by the U.S. Food and Drug Administration to detect DDH in ultrasound images.

This study evaluated the diagnostic accuracy of this commercial AI application. To assess AI generalizability, data sets from two centres: Edmonton, the Canadian centre that provided the original training data for the app, and Melbourne, a large centre in Australia with a much higher prevalence of dysplastic hips due to referral patterns were evaluated. Focus was on the inter-rater agreement between AI and human expert diagnosis and the sensitivity and specificity of the AI diagnostic classification versus reference-standard clinical diagnosis at both centres. It was hypothesized that AI measurements would vary from human expert readings by no more than published ranges of variability between human readers, and that diagnostic classification as normal versus dysplastic would have near-100% sensitivity for severe DDH (Graf III+) at both centres.

## Methods and Materials

### Images and Inclusion Criteria

This retrospective study of prospectively collected data was approved by the health research ethics boards of the two participating centers. All patients had been referred for hip US

to assess for presence of hip dysplasia (from January 2012 to December 2020). With parental written consent, 3D US of each hip was acquired at the same visit as conventional 2D US. The scans were performed at Stollery Children's Hospital, Edmonton, Alberta, Canada, using Philips equipment (typically IU22 13 MHz VL13-5 probe) and the Royal Children's Hospital, Melbourne, Australia using Toshiba/Canon equipment (14LV7 probe) between 2012 and 2020.

Ultrasound scanning was performed at both centers as per the American College of Radiology recommendations. Coronal 3D US images of each hip were captured with the probe positioned with the head resting near the greater trochanter of the infant. Sonographers acquired an automated sweep through each hip image such that the central slice approximated the Graf standard plane.

**AI Methods**

The commercial MEDO Hip package was used, a US-FDA cleared AI analysis package previously trained using 20,000 prior hip images from several centers, including other patients in Edmonton who were not included in this study, and none from Melbourne. When a study is supplied to the app, AI first automatically detects the acetabulum and femoral head on each slice of every 3D image and segments these structures when present. It uses these features to choose the slice which best matches the Graf Standard Plane criteria. AI then measures the alpha angle and acetabular coverage on this slice (Figure 1). Using these indices and other image features, it predicts the hip classification as a Boolean value which indicates whether the patient requires treatment (dx=1) or not (dx=0)). The neural network used for segmentation is a U-Net-like architecture [12]. It consists of an encoding path and a decoding path. Features extracted from the encoder are combined through a sequence of up-convolutions and concatenations to generate the final segmentation. More elaboration has been provided in Chapter 2 under AI methods section.

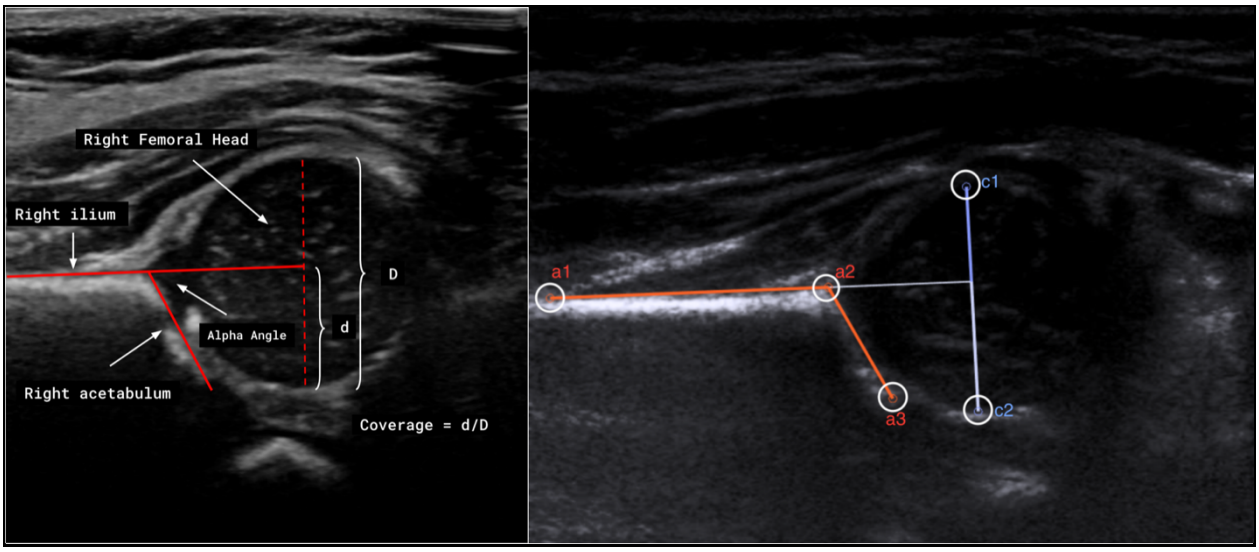


**Figure 3.1** (a) Illustration of alpha angle and coverage calculation in the standard Graf plane as performed in most clinics and (b) 5 Landmarks to measure alpha angle and acetabular coverage used in the MEDO Hip package.

## Statistics

The reference-standard diagnosis against which AI performance was judged was that used in routine clinical care, based on chart review with at least 6 months follow-up for hip that were not normal. The 3D ultrasound models and AI analysis were not available to clinicians or radiologists during their assessment. When hips were considered normal at initial radiologist/clinician review of 2D US images and clinical examination, and no further imaging or treatment was performed, they were classified as 0 (normal). Hips which were recalled for follow-up ultrasound imaging (typically due to immaturity, Graf IIa), but which were considered normal at follow-up and not treated, were classified as 1 (Borderline). Hips that went on to treatment, usually by Pavlik harness, were classified as 2 (Dysplastic). AI classification versus clinical diagnosis was compared via confusion matrix (sensitivity SN, specificity SP, positive and negative predictive values PPV, NPV), for a binary classification in which classes 0 and 1 (normal and borderline) were combined as nondysplastic, versus class 2 (dysplastic).

AI and reference-standard diagnosis were also evaluated as though they were competing raters. Percentage agreement (the proportion of cases for which AI and clinical diagnosis were in agreement) and Randolph's Kappa score [13-15] were used to measure the reliability of categorical data. Agreement was evaluated by criteria of Landis and Koch, where a kappa score of 0.8~1.0 is near-perfect and 0.6~0.8 is substantial [16]. Since hip dysplasia classification is heavily based on the acetabular alpha angle (Figure 3.1), AI versus clinical alpha angle measurements were assessed. The clinical alpha angle was measured from a 2D Graf standard plane image and originally reported clinically. The AI alpha angle was selected by AI on the best slice of a 3DUS sweep image of the same hip (i.e., measured on a different image of a different scan than the clinical alpha angle). Bland-Altman plots were used to evaluate for bias between the two sets of measurements. Scatterplots and Pearson correlation coefficients from linear regression were generated across the entire dataset and also for each center individually.

## Results

From 2012-2020 there were a total of 3633 imaging studies at the two centres (Melbourne=3903, Edmonton=2351), that included concurrently obtained 2D and 3D US with at least two 3DUS image sets (for optimal AI performance). Only the first visit of each patient was included (to avoid biasing the results with any treatments performed before follow-up studies), leaving 2649 studies. Of these, 157 scans were excluded due to either incomplete clinical data or obvious data set errors (e.g., errors in hip labelling). This left 2492 scans (Edmonton=1294, Melbourne=1198) of 1563 unique patients. The mean patient age was 87 days (range 4-267 days). As expected, given DDH prevalence, nearly 70% of patients were girls. Age and gender distribution are shown in Table 3.1 Figure 3.2.

| | Both | Edmonton | Melbourne |
|---|---|---|---|
| **Number of Hips** | 2492 | 1294 | 1198 |
| **Normal or Borderline Hips** | 2327 | 1256 | 1071 |
| **Dysplastic Hips** | 165 | 38 | 127 |
| **Boys %** | 0.30 | 0.35 | 0.26 |
| **Girls %** | 0.68 | 0.65 | 0.71 |
| **Age Mean (Range)** | 87 (4 ~ 267) | 76 (4 ~ 267) | 100 (4 ~ 234) |
| **Boys Age Mean (Range)** | 85 (4 ~ 223) | 79 (4 ~ 223) | 94 (6 ~ 203) |
| **Girls Age Mean (Range)** | 88 (4 ~ 267) | 75 (6 ~ 267) | 102 (4 ~ 234) |

**Table 3.1** Age and Sex distribution of the dataset

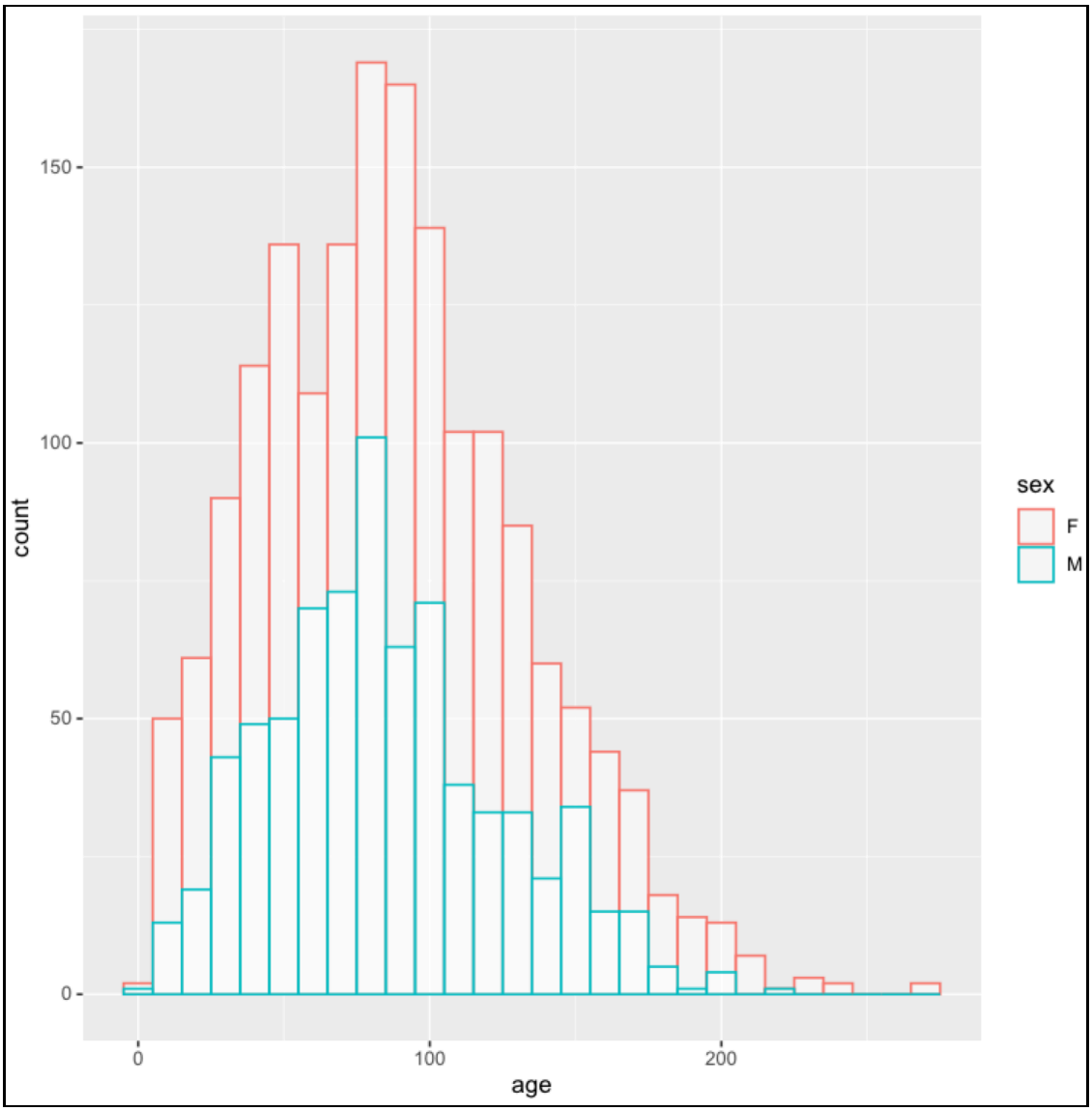


**Figure 3.2** Population distribution by age and sex.

Agreement between AI and clinical diagnosis for DDH diagnosis was high across the entire dataset (86%, p<0.001) and highest for Melbourne data (91%, p<0.001). According to the confusion matrix of positive and negative predictive values for DDH diagnosis by AI versus clinical diagnosis (Table 3.2), with DDH prevalence 6.6%, with respect to human expert reference-standard the AI had sensitivity 0.90 (17 false-negative cases, of which 16 were Graf II subtypes and only one was Graf III), specificity 0.86, negative predictive value NPV=99.2%, positive

predictive value PPV=32%. For the image that was clinically categorised Graf III hip dysplasia, the clinical alpha angle measurement was 42 degrees and AI 44 degrees.

| | | AI diagnosis | |
|---|---|---|---|
| | | **Positive** | **Negative** |
| **Clinical diagnosis** | **Positive** | TP = 148 | FN = 17 |
| | **Negative** | FP = 321 | TN = 2006 |

| | |
|---|---|
| **SN = TP / ( TP + FN )** | 0.897 |
| **SP = TN / ( TN + FP )** | 0.862 |
| **PPV = TP / ( TP + FP )** | 0.316 |
| **NPV = TN / (TN + FN )** | 0.992 |
| **Prevalence = TP + FN / all** | 0.066 |
| **LR+ = SN / ( 1 - SP )** | 6.502 |
| **LR- = ( 1 - SN ) / SP** | 0.12 |

**Table 3.2** Confusion Matrix comparing DDH diagnosis by AI analysis of 3D hip ultrasound images versus reference-standard expert clinical diagnosis from 2D ultrasound images, in Edmonton and Melbourne.

Comparing AI diagnosis from 3DUS and clinical diagnosis from 2DUS as two raters of the same underlying pathology, inter-rater reliability was high for the full dataset and also for each center. While AI and human diagnosis (normal vs. dysplastic) had substantial agreement in Edmonton (Randolph's kappa=0.64, $p<0.0001$) and across both centers ($k=0.73$ , $p<0.001$),

agreement was rated as near-perfect for Melbourne data ($k=0.82$ , $p<0.001$), which contained more dysplastic hips.

AI diagnosis was also tested against a purely image-based reference-standard diagnosis based on the Graf category (I, IIa, IIb, IIc, III) determined from the alpha angle measured clinically on 2DUS and the patient age [4]. As class I and IIa are considered normal and any classes above them need further investigation, three subclasses were formed as I, IIa and (IIb or higher). The kappa score was again calculated and it was observed that agreement was high and followed similar patterns (Table 3.3).

| | **Both** | **Edmonton** | **Melbourne** |
|---|---|---|---|
| **% Agreement** | 0.86 | 0.82 | 0.91 |
| **kappa**<br>**(AI vs. clinical diagnosis)** | 0.73<br>0.70 ~ 0.76 | 0.64<br>0.60 ~ 0.69 | 0.82<br>0.79 ~ 0.86 |
| **kappa**<br>**(AI vs. 2DUS Graf I / IIa / IIb+)** | 0.79<br>0.77 ~ 0.82 | 0.86<br>0.83 ~ 0.89 | 0.72<br>0.68 ~ 0.75 |

**Table 3.3** Inter-rater agreement between AI analysis of 3D hip ultrasound versus human expert clinical diagnosis of 2D ultrasound. “Clinical diagnosis” refers to the radiologist's/surgeon’s clinical decision at time of clinical management. “2DUS Graf I/IIa/IIb+” refers to the simplified Graf image classification on 2DUS.

Bland-Altman plots (Figure 3.3) did not demonstrate any clear systematic bias overall, and it was noted that the few cases seen outside 95% confidence interval boundaries were dysplastic scans from Melbourne.

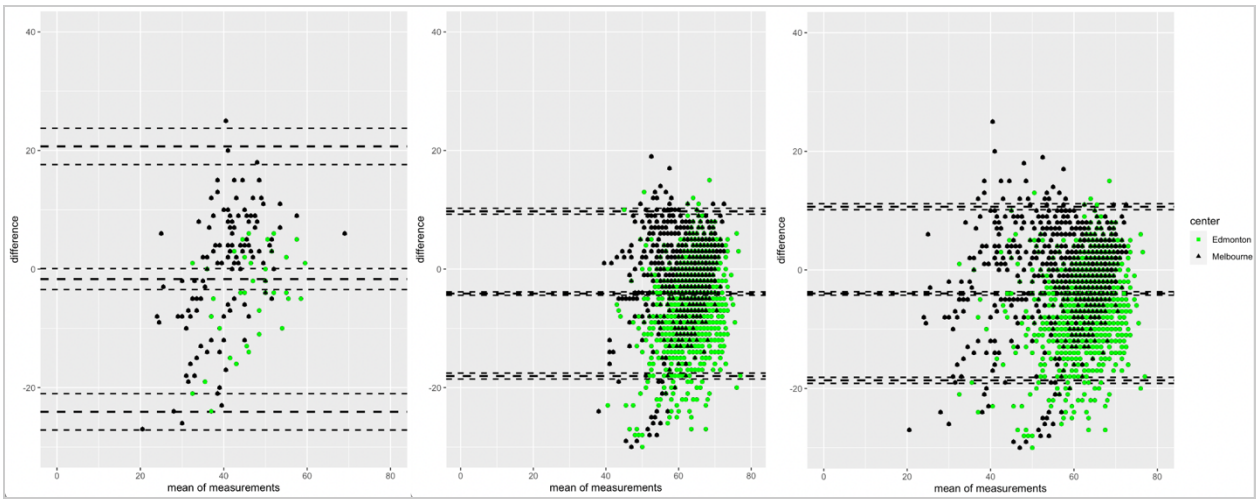


**Figure 3.3** (a) Relation of variability to mean values of alpha angle: Bland-Altman plots for alpha angle measurements at 3DUS/AI vs. 2DUS in the full two-centre dataset, (b) normal and borderline cases only; (c) dysplastic cases only.

A scatterplot (Figure 3.4) confirms high correlation of the AI and clinical reference-standard alpha angle measurements (r2=0.99 for both centres), although with a relatively wide range of variation and a few distant outliers. The ICC for AI versus clinical alpha angles were 0.56 (0.40-0.66, p<0.001) across the two centres, 0.36 (95% CI 0.06-0.56, p<0.001) in Edmonton, and 0.72 (0.69-0.74, p<0.001) in Melbourne.

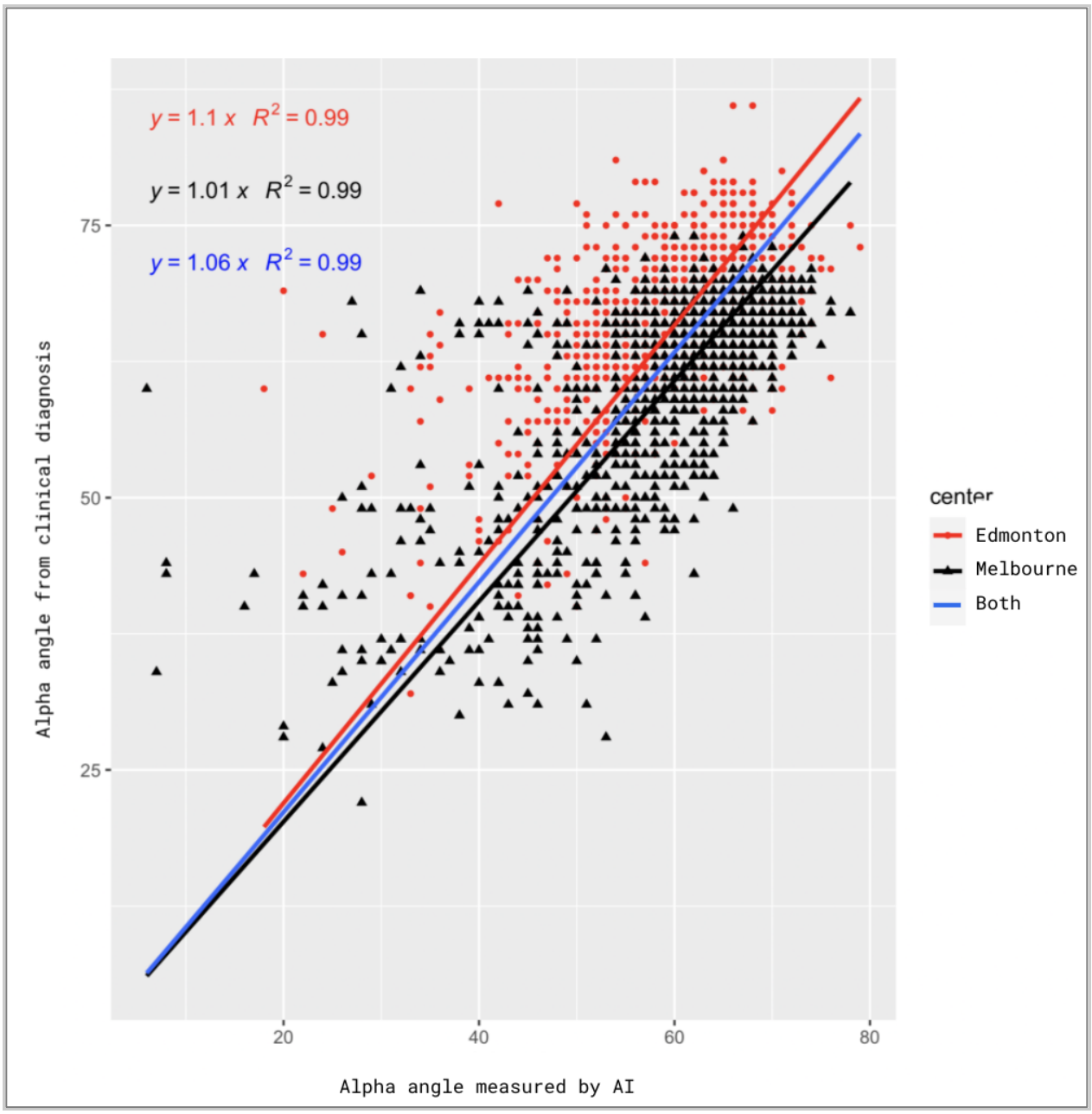


**Figure 3.4** 3DUS/AI versus 2DUS/human expert alpha angle measurements. AI computed alpha angles from the AI-selected best slice within one of two 3DUS data sets per hip. Human expert alpha angle measurements were obtained by the reporting radiologist at time of original clinical interpretation using conventional 2D ultrasound images obtained at the same visit, i.e different images.

When the obvious outliers from the Bland-Altman and scatter-plot diagrams were individually evaluated, three sources of error were noted: (1) dissimilarity of the 3D and the 2D

images to an extent suggesting a different hip had been scanned, likely due to unrecognized errors in data labeling (i.e., left or right hip or wrong patient), (2) human error in reading the 2D image (e.g., a large difference between the original reported clinical value and study team member reassessment of the saved images from the study), and (3) poor quality of the 3D image (e.g., poorly visualized acetabulum or femoral head) (Figure 3.5).

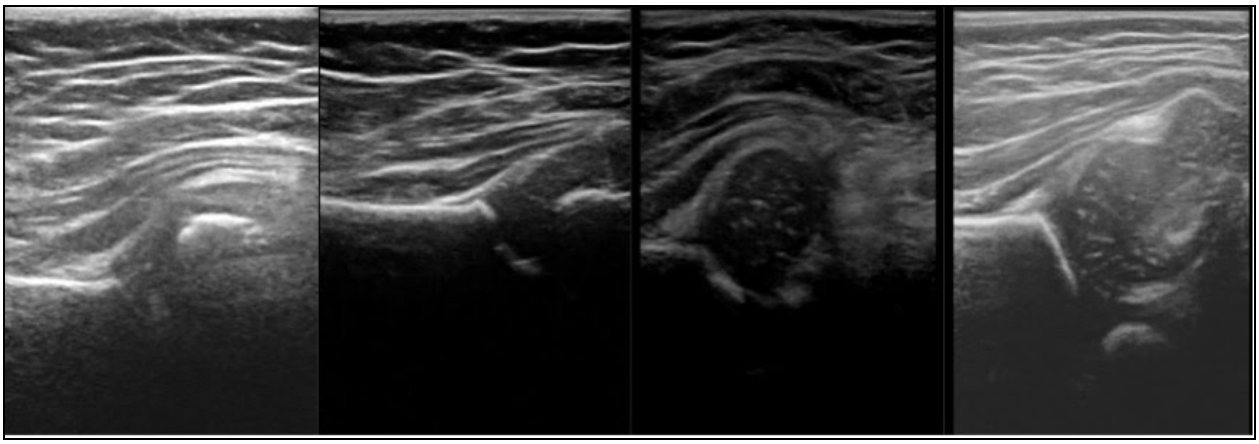

**Figure 3.5** Frames from 3DUS image stacks in 4 hips, in which there were large differences (a, b) and minimal differences (c,d) between 3DUS/AI and 2DUS/human expert alpha angle measurements. The two images on the left (a,b) had large differences in alpha angle measurement and are from 3DUS sweeps of poor quality, while the two images on the right (c,d), of higher quality, gave 3DUS/AI measurements similar to human 2DUS measurements.

The differences between alpha angle measurements (AI versus Clinical) by study centre were also stratified (Figure 3.6). On average the AI alpha angle measurements were slightly lower than clinical measurements at both centres, with a difference of 6.3±5.7° (mean ± standard deviation SD).

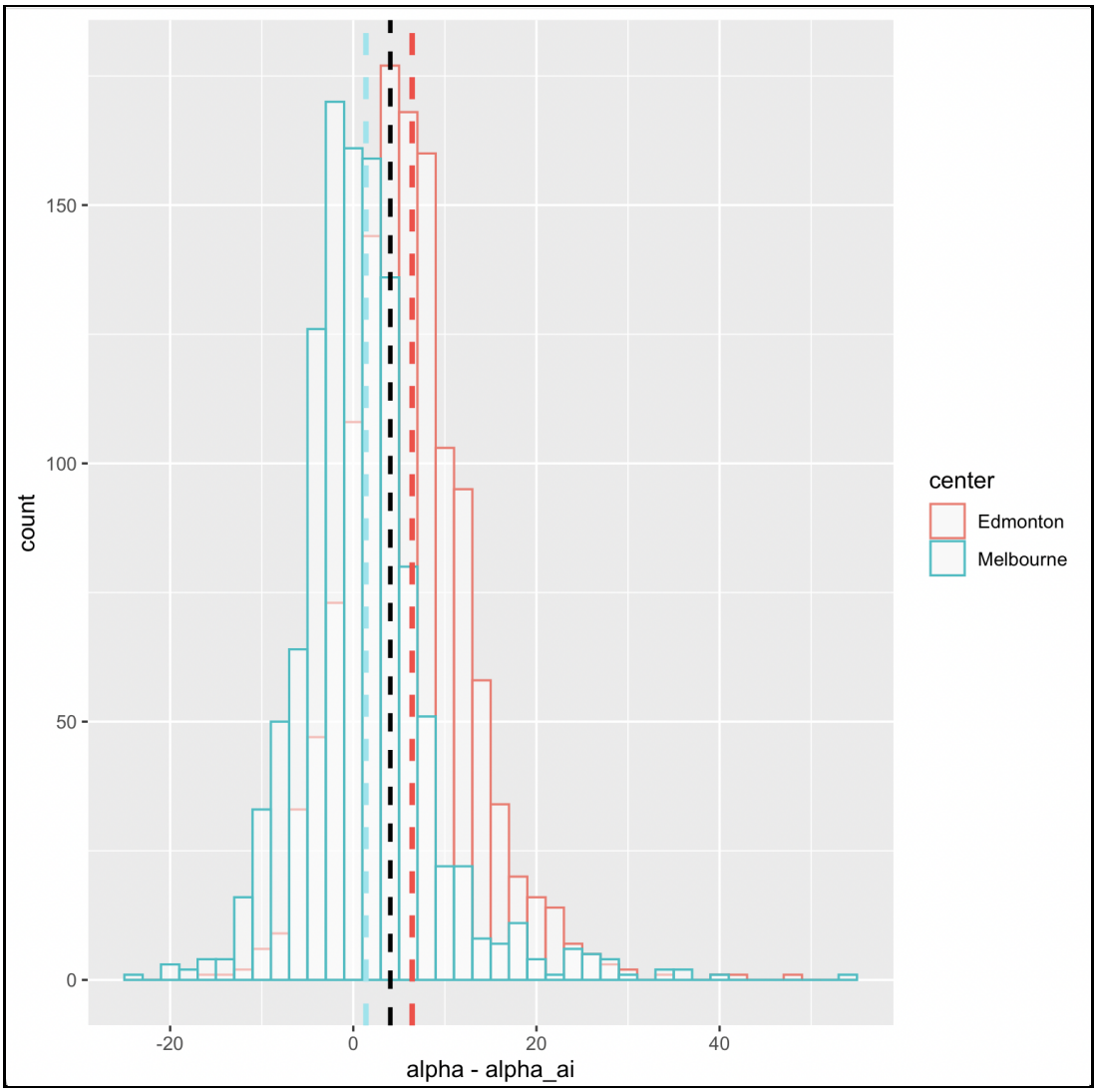


**Figure 3.6** Histogram of differences between alpha angle measurements (AI vs Clinical)

## Discussion

This study compared artificial intelligence analysis of 3D ultrasound images with original clinical interpretation of concurrently obtained conventional 2D ultrasound images for detection of hip dysplasia, in a large prospectively collected dataset from two different tertiary pediatric medical centers. This was a challenging test for AI. In this 8-year study, ultrasound equipment from two manufacturers was used by a diverse collection of sonographers and trainees of varying levels of experience to scan infants at centres in two continents with different distributions of

normal and dysplastic hips and patient demographics, comparing results to diagnosis made by clinicians with varying background and practice patterns. Despite all these factors, we found that 3DUS/AI showed high to near-perfect agreement with clinical diagnostic classification, and sensitivity of 90% for hip dysplasia, across the two centres.

This is the first cross-region multi-center study evaluating automated detection of hip dysplasia at ultrasound versus human observers. Many studies have measured the reliability of measuring and diagnosing DDH on 2DUS, concluding in general that despite high inter-observer variability, reliable diagnosis can be possible. In a recent study on 798 infants, kappa and intraclass correlation coefficients were reported to be very high between expert observers (>0.8) [17]. However, observers of different backgrounds approach images differently, with varying results: inter-rater reliability for measurement of the key sonographic index of DDH, the alpha angle, on 2DUS between two DDH orthopaedic specialists and two paediatric orthopaedic professors had ICC only 0.65 and 0.68 [18]. In a large systematic review on reproducibility of ultrasound-based DDH metrics including 28 studies, ICC scores for the alpha angle were poor to moderate, varying from 0.03 to 0.445 [19]. Because 3DUS provides more comprehensive visualization of the hip than saved single 2D images, it may improve reliability. 3DUS images are easier to acquire for novice sonographers since the probe only needs to be placed roughly near the greater trochanter. In a recent study, intra-rater reliability for the alpha angle was unchanged when expert readers used 3DUS (ICC for 2D=0.80, 3D=0.77), but 3DUS significantly improved reliability for novices (2D=0.54, 3D=0.74), and 3DUS markedly improved inter-rater agreement between expert and novice readers (2D=0.10, 3D=0.83) [7].

In the current study, alpha angles measured on 2DUS in clinical practice and alpha angles measured by AI on 3DUS scans obtained at the same visit showed ICC=0.56. It is expected that

this correlation would be only moderate because it is an inter-scan reliability measurement: the human observers and AI were measuring alpha angles on different images of the same hips, not providing competing interpretations of the identical images. Also, there was a significantly higher ICC in Melbourne (a surgical referral population), than Edmonton (primarily a screening population), likely due to the higher prevalence of dysplastic hips in Melbourne. Since ICC compares within-subject variability to total variability, the fact that Edmonton hips were more frequently normal decreased the total variability in Edmonton data, causing within-subject variability to have a proportionately higher effect reducing ICC in Edmonton.

Another way to look at the reliability of AI alpha angle measurement is the magnitude of differences. In this study, AI versus clinical alpha angle measurements (on different scans) differed by an average of 6.3±5.7° (mean ± standard deviation SD), similar to the SD=6.4-6.7° observed in a 2006 study when the alpha angle was re-measured in the same images by 3 observers with different levels of training [20]. This confirms the findings of a recent study showing AI inter-observer variability for alpha angle measurement to be similar to other human observers [21]

AI showed a robust classification of infants' hips as normal vs. dysplastic requiring treatment. Overall sensitivity for dysplastic hips was SN=90%, with specificity SP=86%. Negative predictive value exceeded 99% while positive predictive value was PPV=32%. This compares favorably to the PPV=14% for conventional DDH ultrasound reported in a recent UK analysis [22]. Additionally, of the 17 false-negative cases in which AI did not detect DDH in patients who went on to clinical treatment for hip dysplasia, 16 were relatively mild (mainly Graf IIb) and only one was severe (Graf III). In the single Graf III hip dysplasia case, the clinical alpha angle measurement was 42 degrees and AI 44 degrees, just on opposite sides of the 43 degree Graf threshold, leading AI to categorize this hip as Graf IIc versus human classification as Graf III.

This type of variability near a threshold is inevitable whenever a threshold exists. The hip was not considered normal by AI, just one grade lower dysplasia than by the human assessment. As expected, our accuracy was somewhat lower than that observed in a 2017 study which involved more laborious semi-automated assessment of 3DUS images in a data set with fewer dysplastic hips, SN=0.98 and PPV=0.93 in a setting which likely represented the optimal performance of 3DUS [3a]. It was observed, 3DUS/AI PPV=32%, indicating that approximately one in three AI-detected positive scans are dysplastic requiring treatment, may seem low at first glance, but is in the optimal range for mammography [23], an imaging-based population screening task with some similarities to DDH screening. Formal economic analysis of suitability for screening is outside the scope of the current study.

In this 8-year study with multiple personnel involved at two sites, the rates of incomplete or erroneous data entry were still relatively low (total 2.16%). The AI had a 100% technical success rate converging on a result in all cases, but this did require use of two 3DUS scans per hip, and there were occasional unusual/outlier AI results which on manual image review occurred mainly in cases with poor 3DUS image quality. Reviewing these images suggests that overall diagnostic accuracy of AI could be improved by adding an initial quality check on the 3D images. Hareendranathan et al. devised an AI-powered scoring system to measure the quality of a hip scan [24]. By adding this, the current low-quality scans which would be rejected by a DDH expert could be removed first-up. Ideally, this quality check could occur prospectively in real time, flagging scans which the user would repeat to improve diagnostic quality.

AI network accuracy at alpha angle measurement was lower on the most dysplastic hips from Melbourne, possibly because the AI network had been exposed to relatively few cases of severe hip dysplasia during training. This measurement is challenging for human observers in

dysplastic hips and the difference was likely of little clinical impact, since it led to misclassification of only one Graf III hip. Some errors in classification were likely also attributable to inevitable imperfections in translating textual clinical diagnosis from imaging reports and clinic charts into categories of normal, borderline and dysplastic. Accordingly, agreement of diagnostic classification was higher between 3DUS/AI and categories derived directly from applying Graf thresholds to clinically measured alpha angles at 2DUS (kappa=0.72-0.86) than between 3DUS/AI and the diagnostic categories derived from clinical practice (i.e treatment performed or not; kappa=0.64-0.82). The decision to apply treatment may also be based on factors not evident in ultrasound images, such as laxity, clinical concern or individual surgeon practice, which almost certainly accounted for some cases of treatment of sonographically normal appearing hips in this data, reducing AI sensitivity.

Strengths of this study include its large size and use of multi-year, multi-center data to provide a clear estimate of real-world AI performance in DDH detection. The study had limitations, most importantly (as is common in many DDH studies) the lack of an external reference standard beyond ultrasound and clinical management. There is a critical lack of long-term data linking infant ultrasound with outcomes beyond skeletal maturity in DDH, a concern motivating initiatives such as the International Hip Dysplasia Registry (IHDR) [25], of which both study sites are members. In this study it would have been desirable to add a reference standard comparison with indices measured on radiographs obtained in all patients at age 6-12 months, as per IHDR protocol, but unfortunately the ethical approval during the study term did not include obtaining radiographs of healthy infants. Results of this study demonstrate that AI analysis of 3DUS is able to closely mimic conventional clinical diagnosis in DDH. The future hope is that

the comprehensive hip shape data provided in 3DUS combined with outcomes data will eventually help improve the accuracy of the reference standard diagnosis of DDH.

A feature of this study that could be seen as a limitation was that AI and human observers were interpreting different scans: 3D image stacks and conventional 2D images, respectively. However, this was intentional, because we ultimately envision a screening test performed by lightly trained users via cine sweep imaging simulating 3D ultrasound, so it is more important to evaluate AI analysis of sweep/3D image stacks than AI performance on single images obtained by expert users. Since index values are known to vary with changes in probe orientation between scans [6], this inevitably decreased the human-AI correlations. However, the fact that AI and clinical diagnostic classifications were so similar despite being based on different images is also a strength, in that it confirms the robustness of the 3DUS/AI approach. A final practical limitation is that the high-resolution 3DUS linear probes used in this study are not widely available. Mass population screening would be more realistically achievable if handheld portable 2DUS probes could be used instead of 3DUS probes. To test this, planned future work will assess whether the observed high diagnostic accuracy is maintained when AI analyzes 2DUS ‘sweeps’ as pseudo-3D hip image stacks.

In conclusion, automated AI analysis of 3DUS had high diagnostic accuracy for classification of infant hips as normal or dysplastic compared to reference-standard clinical diagnosis made using 2DUS. These advances in technology could ultimately be the foundation for cost-effective mass population infant screening for hip dysplasia to reduce the burden of hip osteoarthritis worldwide.

## References

**[1]** Hip Dysplasia, Understanding and Treating Instability of the Native Hip, Paul E. Beaulé Editor

**[2]** Moraleda, L., J. Albiñana, M. Salcedo, and G. Gonzaléz-Morán. "Dysplasia in the Development of the Hip." Revista Española de Cirugía Ortopédica y Traumatología (English Edition) 57, no. 1 (January 1, 2013): 67–77. https://doi.org/10.1016/j.recote.2013.01.009.

**[3]** Shorter D, Hong T, Osborn DA. Cochrane Re- view: screening programmes for developmental dysplasia of the hip in newborn infants. Evid Based Child Health 2013;8(1):11–54.

**[4]** Dezateux C, Rosendahl K. Developmental dysplasia of the hip. Lancet. 2007 May 5;369(9572):1541-1552. doi: 10.1016/S0140-6736(07)60710-7. PMID: 17482986.

**[5]** Graf, R. & (1984). Fundamentals of Sonographic Diagnosis of Infant Hip Dysplasia. Journal of Pediatric Orthopaedics, 4 (6), 735-740.

**[6]** Jaremko JL, et al. 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. 2014;273:870–878. doi: 10.1148/radiol.14140451.

**[7]** Mostofi, E., Chahal, B., Zonoobi, D. et al. Reliability of 2D and 3D ultrasound for infant hip dysplasia in the hands of novice users. Eur Radiol 29, 1489–1495 (2019). https://doi.org/10.1007/s00330-018-5699-1

**[8]** Zonoobi, Dornoosh, Abhilash Hareendranathan, Emanuel Mostofi, Myles Mabee, Saba Pasha, Dana Cobzas, Padma Rao, Sukhdeep K. Dulai, Jeevesh Kapur, and Jacob L. Jaremko. "Developmental Hip Dysplasia Diagnosis at Three-Dimensional US: A Multicenter Study." Radiology 287, no. 3 (June 1, 2018): 1003–15. https://doi.org/10.1148/radiol.2018172592.

**[9]** Hareendranathan, A. R., D. Zonoobi, M. Mabee, D. Cobzas, K. Punithakumar, M. Noga, and J. L. Jaremko. "Toward Automatic Diagnosis of Hip Dysplasia from 2D Ultrasound." In 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), 982–85, 2017. https://doi.org/10.1109/ISBI.2017.7950680.

**[10]** Golan D., Donner Y., Mansi C., Jaremko J., Ramachandran M., on behalf of CUDL (2016) Fully Automating Graf's Method for DDH Diagnosis Using Deep Convolutional Neural Networks. In: Carneiro G. et al. (eds) Deep Learning and Data Labeling for Medical Applications. DLMIA 2016, LABELS 2016. Lecture Notes in Computer Science, vol 10008. Springer, Cham. https://doi.org/10.1007/978-3-319-46976-8_14

**[11]** Zhang, Z., M. Tang, D. Cobzas, D. Zonoobi, M. Jagersand, and J. L. Jaremko. "End-to-End Detection-Segmentation Network with ROI Convolution." In 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), 1509–12, 2018. https://doi.org/10.1109/ISBI.2018.8363859.

**[12]** Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

**[13]** Randolph, J. J. (2005). Free-marginal multirater kappa: An alternative to Fleiss´ fixed-marginal multirater kappa. Paper presented at the Joensuu University Learning and Instruction Symposium 2005, Joensuu, Finland, October 14-15th, 2005. (ERIC Document Reproduction Service No. ED490661)

**[14]** Warrens, M. J. (2010). Inequalities between multi-rater kappas. Advances in Data Analysis and Classification, 4(4), 271-286. doi:10.1007/s11634-010-0073-4

**[15]** Brennan, R. L., & Prediger, D. J. (1981). Coefficient Kappa: Some uses, misuses, and alternatives. Educational and Psychological Measurement (41)3, 687-699.

**[16]** Landis, J. Richard, and Gary G. Koch. "The Measurement of Observer Agreement for Categorical Data." Biometrics 33, no. 1 (1977): 159-74. Accessed December 18, 2020. doi:10.2307/2529310.

**[17]** Pedrotti, Luisella, Ilaria Crivellari, Alessandro Degrate, Federica De Rosa, Francesca Ruggiero, and Mario Mosconi. “Interpreting Neonatal Hip Sonography: Intraobserver and Interobserver Variability.” Journal of Pediatric Orthopaedics B 29, no. 3 (May 2020): 214–218. https://doi.org/10.1097/BPB.0000000000000670.

**[18]** Karakus, Ozgun, Ozgur Karaman, Ahmet Sinan Sari, Mehmet Mufit Orak, and Hasan Hilmi Muratli. “Is It Difficult to Obtain Inter-Observer Agreement in the Measurement of the Beta Angle in Ultrasound Evaluation of the Paediatric Hip?” Journal of Orthopaedic Surgery and Research 14, no. 1 (July 17, 2019): 221. https://doi.org/10.1186/s13018-019-1263-1.

**[19]** Quader, Niamul, Emily K. Schaeffer, Antony J. Hodgson, Rafeef Abugharbieh, and Kishore Mulpuri. “A Systematic Review and Meta-Analysis on the Reproducibility of Ultrasound-Based Metrics for Assessing Developmental Dysplasia of the Hip.” Journal of Pediatric Orthopaedics 38, no. 6 (July 2018): e305. https://doi.org/10.1097/BPO.0000000000001179.

**[20]** Gwynne Jones DP, Vane AG, Coulter G, Herbison P, Dunbar JD. Ultrasound measurements in the management of unstable hips treated with the pavlik harness: reliability and correlation with outcome. J Pediatr Orthop. 2006 Nov-Dec;26(6):818-22. doi: 10.1097/01.bpo.0000234999.61595.ec. PMID: 17065955.

**[21]** Ghasseminia S, et al. Inter-Observer Variability of Hip Dysplasia Indices on Sweep Ultrasound for Novices, Experts, and Artificial Intelligence

**[22]** Rymaruk, S., R. Rashed, K. Nie, Q. Choudry, and R.W. Paton. “ANALYSIS OF THE POSITIVE PREDICTIVE VALUE IN CLINICAL NEONATAL HIP SCREENING FOR INSTABILITY IN DEVELOPMENTAL DYSPLASIA OF THE HIP.” Orthopaedic Proceedings 99-B, no. SUPP_11 (June 1, 2017): 8–8. https://doi.org/10.1302/1358-992X.99BSUPP_11.BSCOS2017-008.

**[23]** Carney, Patricia A., Edward A. Sickles, Barbara S. Monsees, Lawrence W. Bassett, R. James Brenner, Stephen A. Feig, Robert A. Smith, et al. 2010. “Identifying Minimally Acceptable Interpretive Performance Criteria for Screening Mammography.” Radiology 255 (2): 354–61. https://doi.org/10.1148/radiol.10091636.

**[24]** Hareendranathan AR, Chahal B, Ghasseminia S, Zonoobi D, Jaremko JL. Impact of scan quality on AI assessment of hip dysplasia ultrasound. J Ultrasound. 2021 Mar 5. doi: 10.1007/s40477-021-00560-4. Epub ahead of print. PMID: 33675031.

**[25]** Mulpuri K, Schaeffer EK, Kelley SP, Castañeda P, Clarke NM, Herrera-Soto JA, Upasani V, Narayanan UG, Price CT; IHDI Study Group. What Is the Impact of Center Variability in a Multicenter International Prospective Observational Study on Developmental Dysplasia of the Hip? Clin Orthop Relat Res. 2016 May;474(5):1138-45. doi: 10.1007/s11999-016-4746-y. PMID: 26891895; PMCID: PMC4814398.

# CHAPTER 4 DISCUSSION AND CONCLUSIONS

## Thesis Overview

This thesis studied and evaluated the benefits of acquisition of cine sweeps / 3D ultrasound images versus the ordinary 2D ultrasound images and the use of artificial intelligence in reading and interpreting those images (2D, 3D, cine sweeps) and detecting DDH. Due to known issues of Graf method for classification and diagnosis of hip 2DUS (high inter/intra-rater variability), this thesis discussed the advantages of saving sweep images and elaborated on how it can allow the reader as well as artificial intelligence to have a more complete and comprehensive view of the hip and thus provide a more accurate diagnosis with lower inter-rater variability (higher agreement) and higher reproducibility.

Chapter One presented a historical narrative on developmental dysplasia of the hip, discussed its various definitions, explored several associated risk factors, multiple methods to diagnosing it and numerous classifications that have been suggested for it. Chapter One also explored the use of different medical imaging modalities (i.e., CT and Ultrasound) as well as their applications in diagnosing and managing DDH. In the same Chapter, the opportunities for automation of the analysis of hip ultrasound were discussed and the potential for future improvements were reviewed.

Chapter 2 assessed reliability of sweep/3D image acquisition and AI analysis. As such it had two main goals; first to evaluate the effect of the readers' background and their level of expertise in reading hip ultrasound and second, to assess the agreement (hence the variability) between artificial intelligence and gold-standard in classification of hip ultrasound. In Chapter Two, the reliability of AI as an individual reader in processing 2D and sweep ultrasound was

compared against DDH-sub-specialist as well as non-DDH-sub-specialist medical imaging expert readers.

while Chapter Two focused on assessing reliability across a pseudo-random sample of hips, to establish validity and accuracy of the combined new method (3D acquisition and AI interpretation), it was necessary to perform a much larger study. Through a two-center, multi-year study, Chapter Three explored the hypothesis that artificial intelligence analysis of 3D ultrasound can produce high agreement with human experts' diagnosis for hip dysplasia (high sensitivity) and can measure the alpha angle in similar range to human experts.

## Statistical Analysis

Throughout this thesis, for quantitative measurement of agreements between various readers, intraclass correlation coefficient (ICC(2,1), single random rater) was used for continuously-valued data (e.g. alpha angle measurements) and Randolph's Kappa was used for categorical data (e.g. borderline hip versus normal hip).

Randolph's Kappa is a free-marginal implementation of Fleiss-Kappa. Although Fleiss-Kappa solves the issue of having more than two raters which is not possible with Cohen's Kappa, Brennan and Prediger (1981) [1-3] suggest using free-marginal Kappa when raters do not necessarily have to assign a certain number of cases to each category of reading. This was exactly the case for the studies of this thesis.

As mentioned before, single random raters intraclass correlation coefficient statistics (ICC(2,1)) were calculated by traditional techniques for continuously-valued data, together with 95% confidence intervals for each. Single random raters ICC (ICC(2,1)) is preferable when each matter has been rated or measured by each rater, and the reliability is based on a single measurement. This was exactly how this thesis's studies were designed and conducted. The other

types of ICCs are briefly described below, for further clarification on why they were not applicable to this thesis.

- **ICC(1,1)**

  Each matter is rated/measured by a different set of randomly selected raters and the reliability is based on a single measurement. This was not applicable to this thesis as every image was read by every reader.

- **ICC(1,k)**

  Each matter is rated/measured by a different set of randomly selected raters but the reliability is the average of k raters measurements. Similarly, since every image was read by every reader, this method was not applicable.

- **ICC(2,k)**

  Is similar to ICC(2,1) which is used in this study but the reliability is measured as the average of the k ratings, which was not done in this study.

- **ICC(3,1)**

  each matter is rated by a single rater who is the sole rater of interest and the reliability is based on a single rating. Again this was not how this study was designed.

- **ICC(3,k)**

  Is similar to ICC(3,1) but the reliability is measured as the average of the k ratings.

In summary, a special variant of kappa which was most appropriate, and traditional techniques for ICC, were used for this thesis.

## Inter-observer agreement in measuring 2D Ultrasound vs Sweep Ultrasound

Performing a hip ultrasound that is adequate to make a diagnosis on is prone to errors and sometimes very difficult even for experienced sonographers. Optimizing the probe orientation and

finding/approximating the perfect Graf plane can be very challenging. Therefore, the variabilities in measuring the indices on static 2D images of the hip are reported high (low inter/intra rater agreement). With recent technology advancements sweep images (records of short videos of the probe view through the entire hip) are more easily and more frequently acquired. These images show the bone far more comprehensively and allow the reader to choose the best frame from the video to do their readings.

Chapter Two illustrated that the inter-observer reliability for alpha angle and coverage was highest for 2D versus the sweeps. Sonographers had the highest inter-rater agreement for reading both the 2D and sweep images (within their group versus the radiologists, the clinicians, and the medical imaging researchers groups). While examining the hypothesis that AI can measure hip images just like a human reader, Chapter Two confirmed that Randolph kappa agreement score between AI and the human gold-standard is similar to an individual human reader.

Among the medical imaging expert readers, those who had sub-speciality in DDH had higher inter-observer agreement with the gold-standards compared to those who were not sub-specialized in DDH. The results in Chapter Two also showed that the DDH sub-specialist readers had higher ICC scores for both 2D and sweep images. Furthermore, the difference between the agreement scores of the two groups was wider for the sweeps for normal, borderline, and dysplastic hips.

In the same Chapter, the results showed that although AI algorithm had poorer agreement with the gold-standard for 2D images compared to human readers, it still performed like the non-sub-specialist medical imaging expert readers for sweeps.

Finally, per the scope of this thesis and its literature reviews, it is the first study (that we are aware of) on the assessment of inter-observer variability on sweep ultrasound images. It seems

that the extra step of choosing the appropriate image on sweeps, however, introduces challenges for the readers especially with less experience compared to those with more. This results in lower reliability on the sweep images compared to 2D images, but a more direct comparison would be to inter-scan error on 2D images, which is not well studied in the literature.

## Performance of the AI in measuring 3D Ultrasound

As elaborated in Chapter Three, percentage agreement between AI and clinical diagnosis was proved high across the entire dataset in both data centers (Edmonton, Canada and Melbourne, Australia). AI was also highly sensitive in correctly detecting abnormal hips (failed for only one case). Similarly, looking at the AI classification of the 3DUS and comparing it to the clinical diagnosis on the corresponding 2DUS, ICC scores were either high or near perfect for both centers.

We also tested AI on a purely image-based diagnosis (based on Graf categories) and defined three sub-classes as Class I & IIa (normal hips) and class IIb and above (requiring further management). This classification was based on the combination of measured alpha angle on 2DUS and the patients' age (if available). It was observed that AI similarly had a high kappa score agreement with the clinical results in classification of the hips.

Bland-Altman and scatter plots of the alpha angle measurements (AI on 3DUS versus the clinical measurement on the 2DUS) determined no systematic bias and demonstrated a high correlation between the two measurements as the r2 values were calculated 0.99 for each center and across the entire dataset. This was supplemented by an average difference of 6.3 degrees (SD = 5.7) on the measurement of alpha angle which showed AI slightly down-measured comparing to the clinical values.

Considering the outliers in the two plots (scatter plots and Bland-Altman) it was realised that they could be the result of (1) dissimilarity between the 2D and 3D images suggesting that

there were errors in labelling the images (left instead of right or vice-versa), (2) human errors (large differences between the original clinical measurement and the outcome of the re-measurement of the same images during the study team assessment), and (3) low quality 3D images which would not be adequate to make a diagnosis on.

## Limitations

As clinical measurements and diagnosis are routinely made on the 2DUS, throughout the studies in this thesis we had to overcome the challenge of defining human gold standard on the sweeps and 3D images. In our variability study (Chapter Two) we took the median of the readings from a sub-set of the readers who were sub-specialised in DDH as our gold standard. This may have suggested some bias in our results especially while comparing the performance of the two medical imaging expert readers sub-groups (DDH sub-specialists versus non-DDH sub-specialists). Similarly, in our 3D study (Chapter Three), we compared AI measurements on 3D images with gold-standard values which had been extracted from the clinical results on corresponding 2DUS. This we believe was the main reason for some of the outliers in the dataset. We tried to minimize this issue by a holistic and extensive data validation/preparation followed by a comprehensive assessment of the questionable outliers, communicating our doubts with the two centers' representatives and under circumstances when we were completely certain that there was an error, removing the images from the dataset.

The other main limitation was the very subjective nature of hip dysplasia and its classifications methods. Patients are prescribed for treatment based on the clinicians' assessments, but these assessments are not necessarily reflective of Graf methodology. Because of the same issue, a hip image with an alpha angle that had been measured very similarly by both the clinician

and the AI, could have been categorized very differently. So quantitatively AI and the clinician were in perfect agreement, but qualitatively they were not.

Finally, all subjects in our studies were referred for hip ultrasound based on clinical suspicion of DDH. In fact, this was our initial inclusion criterion. Therefore, some of the patients could be beyond the age that ultrasound would be the modality of choice for DDH screening. So, although not ideal, but these were in fact part of clinical routine in the two centers that were participating in our study. This is unfortunately typical of a population where there is no universal screening, and diagnosis can happen late.

## Future Directions

In our two studies, human readers, and AI, made their measurements and diagnosis in absolute isolation. An interesting scenario for future work would be to utilize the AI algorithm as an assistive tool to make the initial measurement on the 2DUS or choose the best slice for making a diagnosis on from a sweep or 3DUS before the human reader can modify or confirm. The effect of this may help with increasing the low intra/inter-rater agreement between hip ultrasound readers.

One of our main observations throughout this study was the considerable extent to which the quality of the images can affect the diagnosis. A human reader is always entitled to reject an image due to inadequacy, but that is not necessarily the case for AI. We noted that most of the questionable results and outliers were the result of low-quality images. Hareendranathan AR et al [4] in their article have discussed the impact of scan quality on AI assessment of hip dysplasia ultrasound and reported AI accuracy of 57% in low quality images versus 89% in other cases. A beneficial avenue to explore would be to add an extra layer of quality check on the images which would reject non-adequate ones.

Furthermore, as we already mentioned in the introduction of this thesis that 2D sweeps and 3D images are very similar, one important bit of future work could be to confirm that these two types of images can be viewed as equivalent in medical imaging and further in the use of AI in interpreting them.

## Conclusion

In this thesis we studied the reliability and validity of a novel approach to infant hip ultrasound combining a new acquisition technique (multi-frame images, either by 2D sweep or 3D capture) with a new image interpretation tool (AI). We assessed interobserver reliability for assessment of hip dysplasia from ultrasound sweep images, for human readers and artificial intelligence. This was the first study of the sort that we are aware of. We found that interobserver reliability is slightly lower on these images than on single 2D images, as expected due to the additional reader task of selecting the optimal image from the sweep for review. Greater reader experience, subspecialist status, and background as a sonographer increased reliability. An AI network functioned with intermediate reliability between experienced and inexperienced human readers, encouraging further study of AI to facilitate DDH diagnosis and treatment. We then observed, in a large multi-center study, that artificial intelligence can classify infant hips 3DUS into normal and dysplastic categories with high accuracy compared to gold-standard clinical diagnosis using 2DUS.

This thesis essentially tried to evaluate the hypothesis that AI can eventually be relied on to measure images of hip that have been acquired by non-expert users, when multi-frame imaging (2D sweep or 3D) is applied. The images used in this thesis were acquired by people with diverse levels of experience. It was intended to represent the spectrum in a typical clinical practice, from

medical imaging researchers who had been trained to take hip ultrasound to expert sonographers who had spent the entirety of their careers in screening infants' hips.

Despite the minor errors mostly pertinent to the strict results of rule-based diagnosis (versus humans' experience/gut-based diagnosis), AI was very sensitive and specific in picking abnormal hips. It does not seem too far away that these minor errors will eventually be minimized to very negligible numbers to enable the health industry to develop and maintain universal screening schemes more easily and more robustly. This will likely be possible only by incorporating similar technologies in every step of the screening process and not solely the image processing step (e.g., image quality check, patients' demographics management etc.).

## References.

**[1]** Randolph, J. J. (2005). Free-marginal multirater kappa: An alternative to Fleiss´ fixed-marginal multirater kappa. Paper presented at the Joensuu University Learning and Instruction Symposium 2005, Joensuu, Finland, October 14-15th, 2005. (ERIC Document Reproduction Service No. ED490661)

**[2]** Warrens, M. J. (2010). Inequalities between multi-rater kappas. Advances in Data Analysis and Classification, 4(4), 271-286. doi:10.1007/s11634-010-0073-4

**[3]** Brennan, R. L., & Prediger, D. J. (1981). Coefficient Kappa: Some uses, misuses, and alternatives. Educational and Psychological Measurement (41)3, 687-699.

**[4]** Hareendranathan AR, Chahal B, Ghasseminia S, Zonoobi D, Jaremko JL. Impact of scan quality on AI assessment of hip dysplasia ultrasound. J Ultrasound. 2021 Mar 5. doi: 10.1007/s40477-021-00560-4. Epub ahead of print. PMID: 33675031.

# COMPLETE BIBLIOGRAPHY

**[1]** Hip Dysplasia, Understanding and Treating Instability of the Native Hip, Paul E. Beaulé Editor

**[2]** Dupuytren G. Original or congenital displacement of the heads OF THIGH-bones. Clin Orthop Relat Res. 1964;33:3–8.

**[3]** Klisic PJ. Congenital dislocation of the hip--a misleading term: brief report. J Bone Joint Surg Br. 1989 Jan;71(1):136. doi: 10.1302/0301-620X.71B1.2914985. PMID: 2914985.

**[4]** Seringe R, Bonnet J-C, Katti E. Pathogeny and natu- ral history of congenital dislocation of the hip. Orthop Traumatol Surg Res. 2014;100(1):59–67. https://doi. org/10.1016/j.otsr.2013.12.006.

**[5]** Bialik V, Bialik GM, Blazer S, Sujov P, Wiener F, Berant M. Developmental dysplasia of the hip: a new approach to incidence. Pediatrics. 1999 Jan;103(1):93-9. doi: 10.1542/peds.103.1.93. PMID: 9917445.

**[6]** BARLOW TG. EARLY DIAGNOSIS AND TREATMENT OF CONGENITAL DISLOCATION OF THE HIP. Proc R Soc Med. 1963 Sep;56(9):804-6. PMID: 14080075; PMCID: PMC1897214.

**[7]** Barlow TG. Congenital dislocation of the hip. Early diagnosis and treatment. Lond Clin Med J. 1964;5:47–58.

**[8]** Ortolani M. Congenital hip dysplasia in the light of early and very early diagnosis. Clin Orthop Relat Res. 1976;119:6–10.

**[9]** Weinstein SL, Mubarak SJ, Wenger DR. Fundamental concepts of developmental dysplasia of the hip. Instr Course Lect. 2014;63:299–305.

**[10]** Tönnis D. Normal values of the hip joint for the evaluation of X-rays in children and adults. Clin Orthop Relat Res. 1976 Sep;(119):39-47. PMID: 954321.

**[11]** Noordin S, Umer M, Hafeez K, Nawaz H. Developmental dysplasia of the hip. Orthop Rev (Pavia). 2010 Sep 23;2(2):e19. doi: 10.4081/or.2010.e19. PMID: 21808709; PMCID: PMC3143976.

**[12]** Narayanan, Unni MBBS, MSc, FRCS(S)*; Mulpuri, Kishore MBBS, MS (Ortho), MHSc(Epi)†; Sankar, Wudbhav N. MD‡; Clarke, Nicholas M.P. ChM, FRCS, FRCS.Ed§; Hosalkar, Harish MBBS, MD‖; Price, Charles T. MD, FAAP¶ International Hip Dysplasia Institute Reliability of a New Radiographic Classification for Developmental Dysplasia of the Hip, Journal of Pediatric Orthopaedics: July/August 2015 - Volume 35 - Issue 5 - p 478-484 doi: 10.1097/BPO.0000000000000318

**[13]** Akiyama M, Nakashima Y, Fujii M, Sato T, Yamamoto T, Mawatari T, Motomura G, Matsuda S, Iwamoto Y. Femoral anteversion is correlated with acetabular version and coverage in

Asian women with anterior and global deficient subgroups of hip dysplasia: a CT study. Skeletal Radiol. 2012 Nov;41(11):1411-8. doi: 10.1007/s00256-012-1368-7. Epub 2012 Feb 13. PMID: 22327395.

**[14]** Nepple, Jeffrey J. MD1,a; Wells, Joel MD, MPH1; Ross, James R. MD2; Bedi, Asheesh MD3; Schoenecker, Perry L. MD1; Clohisy, John C. MD1 Three Patterns of Acetabular Deficiency Are Common in Young Adult Patients With Acetabular Dysplasia, Clinical Orthopaedics and Related Research: April 2017 - Volume 475 - Issue 4 - p 1037-1044 doi: 10.1007/s11999-016-5150-3

**[15]** Fujii M, Nakashima Y, Sato T, Akiyama M, Iwamoto Y. Acetabular tilt correlates with acetabular version and coverage in hip dysplasia. Clin Orthop Relat Res. 2012 Oct;470(10):2827-35. doi: 10.1007/s11999-012-2370-z. Epub 2012 Apr 28. PMID: 22544668; PMCID: PMC3441999.

**[16]** Hartofilakidis G, Stamos K, Ioannidis TT. Low friction arthroplasty for old untreated congenital dislocation of the hip. J Bone Joint Surg Br. 1988;70(2):182–6.

**[17]** Hartofilakidis G, Yiannakopoulos CK, Babis GC. The morphologic variations of low and high hip disloca- tion. Clin Orthop Relat Res. 2008;466(4):820–4. Published online 2008 Feb 21. https://doi.org/10.1007/ s11999-008-0131-9.

**[18]** Hartofilakidis G, Stamos K, Karachalios T, Ioannidis TT, Zacharakis N. Congenital hip disease in adults. Classification of acetabular deficiencies and opera- tive treatment with acetabuloplasty combined with total hip arthroplasty. J Bone Joint Surg Am. 1996;78(5):683–92.

**[19]** Wilkin GP, Ibrahim MM, Smit KM, Beaulé PE. A Contemporary Definition of Hip Dysplasia and Structural Instability: Toward a Comprehensive Classification for Acetabular Dysplasia. J Arthroplasty. 2017 Sep;32(9S):S20-S27. doi: 10.1016/j.arth.2017.02.067. Epub 2017 Mar 3. PMID: 28389135.

**[20]** Wells J, Nepple JJ, Crook K, Ross JR, Bedi A, Schoenecker P, Clohisy JC. Femoral Morphology in the Dysplastic Hip: Three-dimensional Characterizations With CT. Clin Orthop Relat Res. 2017 Apr;475(4):1045-1054. doi: 10.1007/s11999-016-5119-2. PMID: 27752989; PMCID: PMC5339134.

**[21]** Jaremko JL, Wang CC, Dulai S. Reliability of indices measured on infant hip MRI at time of spica cast application for dysplasia. Hip Int. 2014 Jul-Aug;24(4):405-16. doi: 10.5301/hipint.5000143. Epub 2014 May 30. PMID: 24970320.

**[22]** Hesham K, Carry PM, Freese K, Kestel L, Stewart JR, Delavan JA, Novais EN. Measurement of Femoral Version by MRI is as Reliable and Reproducible as CT in Children and Adolescents With Hip Disorders. J Pediatr Orthop. 2017 Dec;37(8):557-562. doi: 10.1097/BPO.0000000000000712. PMID: 28323254; PMCID: PMC5368029.

**[23]** Jia, H., Wang, L., Chang, Y. et al. Assessment of irreducible aspects in developmental hip dysplasia by magnetic resonance imaging. BMC Pediatr 20, 550 (2020). https://doi.org/10.1186/s12887-020-02420-2

**[24]** Rosenbaum, Daniel G., et al. “MR Imaging in Postreduction Assessment of Developmental Dysplasia of the Hip: Goals and Obstacles.” RadioGraphics, no. 3, Radiological Society of North America (RSNA), May 2016, pp. 840–54. Crossref, doi:10.1148/rg.2016150159.

**[25]** Onaç O, Alpay Y, Yapıcı F, Bayhan Aİ. Correlation of postoperative magnetic resonance image measurements with persisting acetabular dysplasia in open reduction of developmental hip dysplasia. Jt Dis Relat Surg. 2021;32(2):461-467. doi: 10.52312/jdrs.2021.48. Epub 2021 Jun 11. PMID: 34145825; PMCID: PMC8343841.

**[26]** Shi XT, Li CF, Cheng CM, Feng CY, Li SX, Liu JG. Preoperative Planning for Total Hip Arthroplasty for Neglected Developmental Dysplasia of the Hip. Orthop Surg. 2019 Jun;11(3):348-355. doi: 10.1111/os.12472. Epub 2019 Jun 13. PMID: 31197911; PMCID: PMC6595139.

**[27]** Albers CE, Rogers P, Wambeek N, Ahmad SS, Yates PJ, Prosser GH. Preoperative planning for redirective, periacetabular osteotomies. J Hip Preserv Surg. 2017 Sep 14;4(4):276-288. doi: 10.1093/jhps/hnx030. PMID: 29250336; PMCID: PMC5721378.

**[28]** Tallroth, Kaj, and Jyri Lepistö. “Computed Tomography Measurement of Acetabular Dimensions: Normal Values for Correction of Dysplasia.” Acta Orthopaedica, no. 4, Informa UK Limited, Jan. 2006, pp. 598–602. Crossref, doi:10.1080/17453670610012665.

**[29]** Shalaby, Mennatallah Hatem, et al. “CT Measurement of Femoral Anteversion Angle in Patients with Unilateral Developmental Hip Dysplasia: A Comparative Study between 2D and 3D Techniques.” The Egyptian Journal of Radiology and Nuclear Medicine, no. 3, Springer Science and Business Media LLC, Sept. 2017, pp. 639–43. Crossref, doi:10.1016/j.ejrnm.2017.02.007.

**[30]** A. Ahmed, Amin, and Mohie El Din Fadel. “Role of Intraoperative Arthrogram in Decision Making of Closed versus Medial Open Reduction of Developmental Hip Dysplasia.” International Journal of Research in Orthopaedics, no. 6, Medip Academy, Oct. 2019, p. 1037. Crossref, doi:10.18203/issn.2455-4510.intjresorthop20194200.

**[31]** Grissom L, Harcke HT, Thacker M. Imaging in the surgical management of developmental dislocation of the hip. Clin Orthop Relat Res. 2008 Apr;466(4):791-801. doi: 10.1007/s11999-008-0161-3. Epub 2008 Feb 21. PMID: 18288547; PMCID: PMC2504666.

**[32]** Graf R. Classification of hip joint dysplasia by means of sonography. Arch Orthop Trauma Surg. 1984;102(4):248–55.

**[33]** Graf R. Fundamentals of sonographic diagno- sis of infant hip dysplasia. J Pediatr Orthop. 1984;4(6):735–40.

**[34]** Graf R. Ultrasonography-guided therapy. Orthopade. 1997;26(1):33–42.

**[35]** Rosendahl K, Aslaksen A, Lie RT, Markestad T. Reliability of ultrasound in the early diagnosis of developmental dysplasia of the hip. Pediatr Radiol. 1995;25(3):219-24. doi: 10.1007/BF02021541. PMID: 7644309.

**[36]** 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 Med Wkly. 2004 Nov 13;134(45-46):671-7. PMID: 15611889.

**[37]** Roovers EA, Boere-Boonekamp MM, Geertsma TS, Zielhuis GA, Kerkhoff AH. Ultrasonographic screening for developmental dysplasia of the hip in infants. Reproducibility of assessments made by radiographers. J Bone Joint Surg Br. 2003 Jul;85(5):726-30. PMID: 12892198.

**[38]** 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 J Orthop. 2015 Nov-Dec;49(6):610-4. doi: 10.4103/0019-5413.168753. PMID: 26806967; PMCID: PMC4705726.

**[39]** Dias JJ, Thomas IH, Lamont AC, Mody BS, Thompson JR. The reliability of ultrasonographic assessment of neonatal hips. J Bone Joint Surg Br. 1993 May;75(3):479-82. doi: 10.1302/0301-620X.75B3.8496227. PMID: 8496227.

**[40]** Jaremko JL, Mabee M, Swami VG, Jamieson L, Chow K, Thompson RB. 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. 2014 Dec;273(3):870-8. doi: 10.1148/radiol.14140451. Epub 2014 Jun 25. PMID: 24964047.

**[41]** Mostofi, E., Chahal, B., Zonoobi, D. et al. Reliability of 2D and 3D ultrasound for infant hip dysplasia in the hands of novice users. Eur Radiol 29, 1489–1495 (2019). https://doi.org/10.1007/s00330-018-5699-1

**[42]** Zonoobi D, Hareendranathan A, Mostofi E, Mabee M, Pasha S, Cobzas D, Rao P, Dulai SK, Kapur J, Jaremko JL. Developmental Hip Dysplasia Diagnosis at Three-dimensional US: A Multicenter Study. Radiology. 2018 Jun;287(3):1003-1015. doi: 10.1148/radiol.2018172592. Epub 2018 Apr 24. PMID: 29688160.

**[43]** Quader N., Hodgson A.J., Mulpuri K., Cooper A., Abugharbieh R. (2017) A 3D Femoral Head Coverage Metric for Enhanced Reliability in Diagnosing Hip Dysplasia. In: Descoteaux M., Maier-Hein L., Franz A., Jannin P., Collins D., Duchesne S. (eds) Medical Image Computing and Computer Assisted Intervention − MICCAI 2017. MICCAI 2017. Lecture Notes in Computer Science, vol 10433. Springer, Cham. https://doi.org/10.1007/978-3-319-66182-7_12

**[44]** Quader, Niamul. "Automatic Characterization of Developmental Dysplasia of the Hip in Infants Using Ultrasound Imaging." University of British Columbia, University of British Columbia, 2018, doi:10.14288/1.0364129.

**[45]** Stoica Z, Dumitrescu D, Popescu M, Gheonea I, Gabor M, Bogdan N. Imaging of avascular necrosis of femoral head: familiar methods and newer trends. Curr Health Sci J. 2009 Jan;35(1):23-8. Epub 2009 Mar 21. PMID: 24778812; PMCID: PMC3945237.

**[46]** Resnick D, Niwayama G. Osteonecrosis: diagnostic techniques, special situations and complications. In: Resnick D, editor. Diagnosis of Bone and Joint Disorders. 3. Philadelphia: WB Saunders Co; 1995. pp. 3495–3558.

**[47]** Ntoulia A, Barnewolt CE, Doria AS, Ho-Fung VM, Lorenz N, Mentzel HJ, Back SJ. Contrast-enhanced ultrasound for musculoskeletal indications in children. Pediatr Radiol. 2021 Mar 30. doi: 10.1007/s00247-021-04964-6. Epub ahead of print. PMID: 33783575.

**[48]** Back SJ, Chauvin NA, Ntoulia A, Ho-Fung VM, Calle Toro JS, Sridharan A, Morgan TA, Kozak B, Darge K, Sankar WN. Intraoperative Contrast-Enhanced Ultrasound Imaging of Femoral Head Perfusion in Developmental Dysplasia of the Hip: A Feasibility Study. J Ultrasound Med. 2020 Feb;39(2):247-257. doi: 10.1002/jum.15097. Epub 2019 Jul 23. PMID: 31334874.

**[49]** Gornitzky AL, Georgiadis AG, Seeley MA, Horn BD, Sankar WN. Does Perfusion MRI After Closed Reduction of Developmental Dysplasia of the Hip Reduce the Incidence of Avascular Necrosis? Clin Orthop Relat Res. 2016 May;474(5):1153-65. doi: 10.1007/s11999-015-4387-6. PMID: 26092677; PMCID: PMC4814438.

**[50]** Tiderius C, Jaramillo D, Connolly S, Griffey M, Rodriguez DP, Kasser JR, Millis MB, Zurakowski D, Kim YJ. Post-closed reduction perfusion magnetic resonance imaging as a predictor of avascular necrosis in developmental hip dysplasia: a preliminary report. J Pediatr Orthop. 2009 Jan-Feb;29(1):14-20. doi: 10.1097/BPO.0b013e3181926c40. PMID: 19098638.

**[51]** Hareendranathan AR, Mabee M, Punithakumar K, Noga M, Jaremko JL. A technique for semiautomatic segmentation of echogenic structures in 3D ultrasound, applied to infant hip dysplasia. Int J Comput Assist Radiol Surg. 2016 Jan;11(1):31-42. doi: 10.1007/s11548-015-1239-5. Epub 2015 Jun 20. PMID: 26092660.

**[52]** El-Hariri, Houssam. 2020. “Reliable and Robust Hip Dysplasia Measurement with Three-Dimensional Ultrasound and Convolutional Neural Networks.” Electronic Theses and Dissertations (ETDs) 2008+. T, University of British Columbia. doi:http://dx.doi.org/10.14288/1.0389533.

**[53]** Çiçek, Özgün et al. “3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation.” Lecture Notes in Computer Science (2016): 424–432. Crossref. Web.

**[54]** Golan D., Donner Y., Mansi C., Jaremko J., Ramachandran M., on behalf of CUDL (2016) Fully Automating Graf’s Method for DDH Diagnosis Using Deep Convolutional Neural Networks. In: Carneiro G. et al. (eds) Deep Learning and Data Labeling for Medical Applications. DLMIA 2016, LABELS 2016. Lecture Notes in Computer Science, vol 10008. Springer, Cham. https://doi.org/10.1007/978-3-319-46976-8_14

**[55]** Z. Zhang, M. Tang, D. Cobzas, D. Zonoobi, M. Jagersand and J. L. Jaremko, "End-to-end detection-segmentation network with ROI convolution," 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), 2018, pp. 1509-1512, doi: 10.1109/ISBI.2018.8363859.

**[56]** 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, pp. 1356-1359, doi: 10.1109/ISBI.2018.8363823.

**[57]** Houssam El-Hariri, Antony J. Hodgson, Kishore Mulpuri, Rafeef Garbi, Automatically Delineating Key Anatomy in 3-D Ultrasound Volumes for Hip Dysplasia Screening, Ultrasound in Medicine & Biology, 2021, ISSN 0301-5629, https://doi.org/10.1016/j.ultrasmedbio.2021.05.011.

**[58]** Expert Panel on Pediatric Imaging: Jie C. Nguyen, MD, MSa ; Scott R. Dorfman, MDb ; Cynthia K. Rigsby, MDc ; Ramesh S. Iyer, MDd ; Adina L. Alazraki, MDe ; Sudha A. Anupindi, MDf ; Dianna M. E. Bardo, MDg ; Brandon P. Brown, MDh ; Sherwin S. Chan, MD, PhDi ; Tushar Chandra, MDj ; Matthew D. Garber, MDk ; Michael M. Moore, MDl ; Nirav K. Pandya, MDm; Narendra S. Shet, MDn ; Alan Siegel, MD, MSo ; Boaz Karmazyn, MD.p, Developmental Dysplasia of the Hip (DDH)–Child. Available at https://acsearch.acr.org/docs/69437/Narrative/. American College of Radiology. Accessed Aug 19, 2021

**[59]** 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.

**[60]** Mulpuri K, Song KM, Goldberg MJ, Sevarino K. Detection and Nonoperative Management of Pediatric Developmental Dysplasia of the Hip in Infants up to Six Months of Age. J Am Acad Orthop Surg 2015;23:202-5.

**[61]** American Journal of Roentgenology. 2014;203: 1324-1335. 10.2214/AJR.13.12449

**[62]** 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 Dec 6;362(9399):1883-7. doi: 10.1016/S0140-6736(03)14957-4. PMID: 14667743.

**[63]** Graf, R. “The Diagnosis of Congenital Hip-Joint Dislocation by the Ultrasonic Combound Treatment.” Archives of Orthopaedic and Traumatic Surgery, no. 2, Springer Science and Business Media LLC, Sept. 1980, pp. 117–33. Crossref, doi:10.1007/bf00450934.

**[64]** Jacobino, Bruno de Castro Paixão, et al. “Using the Graf Method of Ultrasound Examination to Classify Hip Dysplasia in Neonates.” Autopsy and Case Reports, no. 2, Editora Cubo, 2012, pp. 5–10. Crossref, doi:10.4322/acr.2012.018.

**[65]** Falliner, A., et al. “Comparing Ultrasound Measurements of Neonatal Hips Using the Methods of Graf and Terjesen.” The Journal of Bone and Joint Surgery. British Volume, no. 1, British Editorial Society of Bone & Joint Surgery, Jan. 2006, pp. 104–06. Crossref, doi:10.1302/0301-620x.88b1.16419.

**[66]** Harke HT, Grissom LE: Performing dynamic sonography of the infant hip. AJR Am J Roentgenol 1990;155:837–844.

**[67]** Shin, YiRang, et al. “Artificial Intelligence in Musculoskeletal Ultrasound Imaging.” Ultrasonography, no. 1, Korean Society of Ultrasound in Medicine, Jan. 2021, pp. 30–44. Crossref, doi:10.14366/usg.20080.

**[68]** Hareendranathan, A. R., D. Zonoobi, M. Mabee, D. Cobzas, K. Punithakumar, M. Noga, and J. L. Jaremko. “Toward Automatic Diagnosis of Hip Dysplasia from 2D Ultrasound.” In 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), 982–85, 2017. https://doi.org/10.1109/ISBI.2017.7950680.

**[69]** Quader N, Hodgson AJ, Mulpuri K, Schaeffer E, Abugharbieh R. Automatic Evaluation of Scan Adequacy and Dysplasia Metrics in 2-D Ultrasound Images of the Neonatal Hip. Ultrasound Med Biol. 2017 Jun;43(6):1252-1262. doi: 10.1016/j.ultrasmedbio.2017.01.012. Epub 2017 Mar 22. PMID: 28341489.

**[70]** Paserin O., Mulpuri K., Cooper A., Hodgson A.J., Abugharbieh R. (2017) Automatic Near Real-Time Evaluation of 3D Ultrasound Scan Adequacy for Developmental Dysplasia of the Hip. In: Cardoso M. et al. (eds) Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures. CARE 2017, CLIP 2017. Lecture Notes in Computer Science, vol 10550. Springer, Cham. https://doi-org.login.ezproxy.library.ualberta.ca/10.1007/978-3-319-67543-5_12

**[71]** Randolph, J. J. (2005). Free-marginal multirater kappa: An alternative to Fleiss´ fixed-marginal multirater kappa. Paper presented at the Joensuu University Learning and Instruction Symposium 2005, Joensuu, Finland, October 14-15th, 2005. (ERIC Document Reproduction Service No. ED490661)

**[72]** Warrens, M. J. (2010). Inequalities between multi-rater kappas. Advances in Data Analysis and Classification, 4(4), 271-286. doi:10.1007/s11634-010-0073-4

**[73]** Brennan, R. L., & Prediger, D. J. (1981). Coefficient Kappa: Some uses, misuses, and alternatives. Educational and Psychological Measurement (41)3, 687-699.

**[74]** Landis, J. Richard, and Gary G. Koch. "The Measurement of Observer Agreement for Categorical Data." Biometrics 33, no. 1 (1977): 159-74. Accessed December 18, 2020. doi:10.2307/2529310.

**[75]** Pedrotti, Luisella, Ilaria Crivellari, Alessandro Degrate, Federica De Rosa, Francesca Ruggiero, and Mario Mosconi. “Interpreting Neonatal Hip Sonography: Intraobserver and Interobserver Variability.” Journal of Pediatric Orthopaedics B 29, no. 3 (May 2020): 214–218. https://doi.org/10.1097/BPB.0000000000000670.

**[76]** Yildiz, Kadri, Hayrunnisa BEKİS BOZKURT Bekis, Türkhun Çetin, and Vahit Yildiz. “Interobserver Reliability in the Ultrasonic Evaluation with Graf Method of Developmental Dysplasia of the Hip: The Importance of Education for Ultrasonography Classification.” Journal of Health Sciences and Medicine 3, no. 2 (March 19, 2020): 11–124. https://doi.org/10.32322/jhsm.676820.

**[77]** Karakus, Ozgun, Ozgur Karaman, Ahmet Sinan Sari, Mehmet Mufit Orak, and Hasan Hilmi Muratli. “Is It Difficult to Obtain Inter-Observer Agreement in the Measurement of the Beta Angle

in Ultrasound Evaluation of the Paediatric Hip?" Journal of Orthopaedic Surgery and Research 14, no. 1 (July 17, 2019): 221. https://doi.org/10.1186/s13018-019-1263-1.

**[78]** Ömeroglu, Hakan, Ali Biçmoglu, Süha Koparal, and Sinan Seber. "Assessment of Variations in the Measurement of Hip Ultrasonography by the Graf Method in Developmental Dysplasia of the Hip§." Journal of Pediatric Orthopaedics B 10, no. 2 (April 2001): 89–95.

**[79]** Quader, Niamul, Emily K. Schaeffer, Antony J. Hodgson, Rafeef Abugharbieh, and Kishore Mulpuri. "A Systematic Review and Meta-Analysis on the Reproducibility of Ultrasound-Based Metrics for Assessing Developmental Dysplasia of the Hip." Journal of Pediatric Orthopaedics 38, no. 6 (July 2018): e305. https://doi.org/10.1097/BPO.0000000000001179.

**[80]** Ismiarto, YD, P Agradi, and ZN Helmi. "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 13, no. 3 (November 2019): 60–65. https://doi.org/10.5704/MOJ.1911.010.

**[81]** Hareendranathan AR, Chahal B, Ghasseminia S, Zonoobi D, Jaremko JL. Impact of scan quality on AI assessment of hip dysplasia ultrasound. J Ultrasound. 2021 Mar 5. doi: 10.1007/s40477-021-00560-4. Epub ahead of print. PMID: 33675031.

**[82]** Moraleda, L., J. Albiñana, M. Salcedo, and G. Gonzaléz-Morán. "Dysplasia in the Development of the Hip." Revista Española de Cirugía Ortopédica y Traumatología (English Edition) 57, no. 1 (January 1, 2013): 67–77. https://doi.org/10.1016/j.recote.2013.01.009.

**[83]** Shorter D, Hong T, Osborn DA. Cochrane Re- view: screening programmes for developmental dysplasia of the hip in newborn infants. Evid Based Child Health 2013;8(1):11–54.

**[84]** Dezateux C, Rosendahl K. Developmental dysplasia of the hip. Lancet. 2007 May 5;369(9572):1541-1552. doi: 10.1016/S0140-6736(07)60710-7. PMID: 17482986.

**[85]** Graf, R. & (1984). Fundamentals of Sonographic Diagnosis of Infant Hip Dysplasia. Journal of Pediatric Orthopaedics, 4 (6), 735-740.

**[86]** Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

**[87]** Gwynne Jones DP, Vane AG, Coulter G, Herbison P, Dunbar JD. Ultrasound measurements in the management of unstable hips treated with the pavlik harness: reliability and correlation with outcome. J Pediatr Orthop. 2006 Nov-Dec;26(6):818-22. doi: 10.1097/01.bpo.0000234999.61595.ec. PMID: 17065955.

**[88]** Ghasseminia S, et al. Inter-observer variability of hip dysplasia indices on sweep ultrasound for novices, experts, and artificial intelligence

**[89]** Rymaruk, S., R. Rashed, K. Nie, Q. Choudry, and R.W. Paton. "ANALYSIS OF THE POSITIVE PREDICTIVE VALUE IN CLINICAL NEONATAL HIP SCREENING FOR

INSTABILITY IN DEVELOPMENTAL DYSPLASIA OF THE HIP." Orthopaedic Proceedings 99-B, no. SUPP_11 (June 1, 2017): 8–8. https://doi.org/10.1302/1358-992X.99BSUPP_11.BSCOS2017-008.

**[90]** Carney, Patricia A., Edward A. Sickles, Barbara S. Monsees, Lawrence W. Bassett, R. James Brenner, Stephen A. Feig, Robert A. Smith, et al. 2010. "Identifying Minimally Acceptable Interpretive Performance Criteria for Screening Mammography." Radiology 255 (2): 354–61. https://doi.org/10.1148/radiol.10091636.

**[91]** Mulpuri K, Schaeffer EK, Kelley SP, Castañeda P, Clarke NM, Herrera-Soto JA, Upasani V, Narayanan UG, Price CT; IHDI Study Group. What Is the Impact of Center Variability in a Multicenter International Prospective Observational Study on Developmental Dysplasia of the Hip? Clin Orthop Relat Res. 2016 May;474(5):1138-45. doi: 10.1007/s11999-016-4746-y. PMID: 26891895; PMCID: PMC4814398.