Purpose: Ultrasound (US) is the standard imaging modality used to screen for developmental dysplasia of the hip in infants. Currently, radiologists or orthopaedic surgeons review scan images and judge them to be adequate or inadequate for interpretation. If considered adequate, diagnostic dysplasia metrics are determined; however, there is no standardized method for this process. There is significant inter-observer variability in this manual process which can affect misdiagnosis rates. To eliminate this subjectivity, we developed an automatic method to identify adequate US images and extract dysplasia metrics. The purpose of this study was to validate the efficacy of this automatic method by comparing results with observer-determined dysplasia metrics on a set of US images.

Methods: A total of 693 US images from scans of 35 infants were analyzed. Trained clinicians at a single institution labeled each image as adequate or inadequate, and subsequently measured $\alpha$ and $\beta$ angles on adequate images to diagnose dysplasia. We trained our image classifier on random sets of 415 images and used it to assess the remaining 278 images. $\alpha$ and $\beta$ angles were automatically estimated on all adequate images. We compared the manual and automatic methods for discrepancies in adequacy determination, metric variability and incidences of missed early diagnosis or over-treatment.

Results: There was excellent agreement between the automatic and manual methods in image adequacy classification (Kappa coefficient =0.912). On each adequate US image, $\alpha$ and $\beta$ angle measurements were compared, producing mixed levels of agreement between methods. Mean discrepancies of $1.78^\circ \pm 4.72^\circ$ and $8.91^\circ \pm 6.437^\circ$ were seen for $\alpha$ and $\beta$ angles, respectively. Standard deviations of the angle measures across multiple images from a single patient scan were significantly reduced by the automatic method for both $\alpha$ ($p<0.05$) and $\beta$ ($p<0.01$) angles. Additionally, the automatic method classified three hips (two patients) as Graf type II and two hips (two patients) as type III, while the manual method classified them as type I and II, respectively. Both cases flagged as type III patients by the automatic system subsequently failed Pavlik harness treatment and were booked for surgery.

Conclusion: The automatic method produced excellent agreement with radiologists in scan adequacy classification and significantly reduced measurement variability. Good agreement between methods was found in Graf classification. In instances of disagreement, subsequent clinical findings seemed to support the classification of the automatic method. This proposed method presents an alternative automatic, near-real-time analysis for US images that may potentially significantly improve dysplasia metric reliability and reduce missed early diagnoses without increasing over-treatment.