Ultrasonography of the hip in developmental hip dysplasia

E. Bar-On, S. Meyer, G. Harari, S. Porat
From Schneider Children’s and Rabin Medical Center, Alyn Hospital and Hadassah Medical Center

Ultrasonography of the hip was performed sequentially by two different examiners in 75 infants. The ultrasound strips were reviewed twice by three paediatric orthopaedic surgeons and classified by the Graf method. The intraobserver and interobserver agreement between the interpretations was analysed using simple and weighted kappa coefficients calculated for agreement on the Graf classification and for grouping as normal (types 1A to 2A), and abnormal requiring treatment (types 2B to 4).

When examining the same ultrasound strip, intraobserver agreement for the Graf classification was substantial (mean kappa 0.61), but interobserver agreement was only moderate (kappa 0.50). For the grouping into normal and abnormal, the mean kappa value for intraobserver agreement was 0.67 and for interobserver agreement 0.57. Because of the significant differences in agreement between normal and abnormal hips, we analysed a subgroup of those with at least one abnormal interpretation. Intraobserver agreement within this subgroup showed moderate reliability (kappa 0.41), but interobserver agreement was only fair (kappa 0.28). Interpretations of two different strips performed sequentially showed significantly lower agreement with an intraobserver kappa value of 0.29 and an interobserver value of 0.28. In the subgroup with at least one abnormal reading, the intraobserver kappa was 0.09 and the interobserver 0.1.

Our findings suggest that both the technique of performing ultrasonography and the interpretation of the image may influence the result.

Received 18 September 1997; Accepted after revision 11 November 1997

Since its introduction in 1980, ultrasonography of the hip has gained wide acceptance as the primary method for the screening, diagnosis, and monitoring of the treatment of developmental hip dysplasia (DDH) in infants. The two techniques most commonly used are the static method of Graf and the dynamic method of Clarke et al. In the first, a coronal plane image is obtained and the hip is classified on the qualitative assessment of the bony and cartilaginous acetabular components and on quantitative measurements between these components and the ilium. The alpha angle represents the bony roof and the beta angle the cartilaginous roof of the acetabulum.

Ultrasonography is performed either independently or as part of a general clinical review in which a history and physical examination are obtained.

We have used ultrasonography since 1986 and have performed over 10,000 examinations using the Graf method. Our aim in this study was to examine the reliability and reproducibility of our findings.

Materials and Methods

We randomly selected 75 infants from our ultrasound hip clinic. There were 52 girls and 23 boys with a mean age of 14 weeks (2 to 40). All had a standard assessment with a history and physical examination. Ultrasonography of both hips using a 5 MHz linear transducer of the Siemens Sonoline 1 unit (Siemens, Erlangen, Germany) was performed by a paediatric orthopaedic surgeon who then classified the hips according to the Graf method and made recommendations for treatment. After obtaining parental consent, the infants then had a second ultrasonography by a different examiner who did not obtain a history and physical examination and did not see the first recording. Thus, a total of 150 hips was examined each having two ultrasonographs (300 hard-copy strips).

Both hard-copy strips were then coded and assessed.
twice by each of three paediatric orthopaedic surgeons with at least one month between the two assessments. The examiners selected the best image from the strip, measured the alpha angle using a goniometer, and classified the hips according to the Graf classification. Each hip was therefore assessed six times which, in addition to the original measurement, gave a total of 1950 readings.

**Statistical analysis.** The data were analysed using SAS software (SAS Institute, Cary, NC).

**Measurement of the alpha angle.** Intraobserver and interobserver measurements were paired and Pearson correlation coefficients were calculated between the pairs. Measurements of the alpha angle were analysed using the paired $t$-test and the non-parametric sign-rank test. The limits of agreement, the range within which 95% of the differences between two readings of the same scan could be expected to lie, were calculated.\textsuperscript{9}

**Categorical parameters.** These were subjected to chi-squared testing, and McNemar’s test was applied. All the tests were two-tailed and a $p$ value of 5% or less was considered significant. We used the simple kappa coefficient between paired observers and the weighted kappa coefficient between multiple observers. This gave a measure of agreement between observations relative to random agreement. A value of 1.00 represents full agreement, that of 0.00 represents random agreement and a negative value indicates less than random agreement. The kappa values were classified as poor to excellent using the criteria laid down by Landis and Koch.\textsuperscript{10}

We analysed the data for ‘academic’ agreement on the exact Graf classification and also for ‘practical’ agreement, by grouping the hips into ‘normal’ (types 1A to 2A) requiring no treatment and ‘abnormal’ (types 2B to 4), requiring treatment according to Graf’s recommendations.

**Results**

**Alpha angle.** The mean Pearson correlation coefficients for intraobserver agreement was 0.71 (0.67 to 0.79) and for interobserver agreement 0.61 (0.49 to 0.74) with a $p$ value of 0.0001. The mean limit of agreement of the alpha angles was $10^\circ$ (7 to 12) for intraobserver and $13^\circ$ (8 to 16) for interobserver measurements.

**Graf classification**

**Interpretation of the same strip.** Academic intraobserver agreement on the Graf classification was substantial with a mean kappa coefficient of 0.61 (0.49 to 0.69). Interobserver agreement was moderate with a mean kappa value of 0.50 (0.45 to 0.55). The practical agreement of indications for...
treatment showed even higher reliability with kappa values of 0.67 for intra- and 0.57 for interobserver agreement. There was, however, a significant difference between normal and abnormal hips. A hip read as normal had a 98% chance of being read again as normal both by the same or by a different observer (Fig. 1), but a hip interpreted as abnormal by one observer had only a 69% chance of being reinterpreted as abnormal by the same observer and a 56% chance of being read as normal by a different observer (Fig. 2).

We therefore isolated a subgroup of abnormal hips which included 47 strips in which at least one reading had been as abnormal. Agreement in this group was found to be significantly lower. Intraobserver reliability was only moderate with a mean kappa value of 0.41 (0.18 to 0.71) and interobserver reliability was fair with a mean kappa value of 0.28 (0.14 to 0.41) (Fig. 3).

**Interpretation of different strips.** When we examined agreement between the interpretation of two ultrasound strips performed consecutively by two different examiners, we found a substantial decrease in reliability (Fig. 4). When all strips were included, the mean intraobserver reliability coefficient was found to be 0.29 (–0.03 to +0.49) and the mean interobserver coefficient was 0.28 (0.12 to 0.55). When only those strips containing abnormal readings were considered, kappa values were found to be even lower with a mean intraobserver value of 0.09 (–0.16 to +0.29) and an interobserver value of 0.10 (–0.103 to +0.41) (Fig. 5).

**Effect of the history and physical examination.** We examined the kappa coefficient for two readings of the same strip by the same examiner; one determined blindly and one performed during the original clinic examination in which a history and physical examination had been done. This value was compared with that between two blind examinations of the same strip performed by the same examiners. We found a decrease in kappa values from 0.67 to 0.56 for all strips, and from 0.41 to 0.37 when only those strips containing abnormal readings were considered.

**Discussion**

Although widely accepted as the primary diagnostic tool in the management of DDH, ultrasonography has several potential sources of error both in performance and interpretation.

The examination is generally performed either by radiologists or by orthopaedic surgeons and may not always include a history and physical examination. The technique of positioning the infant and placing the transducer, although delineated by Graf, can vary and may influence the image obtained. Other potential sources of discrepancy include the placement of the lines which depict the alpha angle, the positioning of the goniometer for its measure-
ment, and variability between different goniometers. The correlation coefficients and the limits of agreement for the alpha angles measured in our study closely correlate with those found by Dias et al.\textsuperscript{11} The interpretation of the image, the Graf classification and recommendations for treatment are based on the quantitative measurement of the alpha angle and also on qualitative parameters which are more prone to subjective interpretation.

Our results indicate that the technique of performing the examination and the interpretation of the ultrasonographic image obtained may influence the result. The availability of a history and physical examination may also have some influence, although to a less degree.

The examinations carried out in our study were undertaken in identical settings on the same positioning device and using the same ultrasound equipment. All three examiners were paediatric orthopaedic surgeons working in the same location and communicating on a regular basis; two of them had attended a formal training course in ultrasonography. When examinations are performed by examiners of differing backgrounds working in variable settings additional factors affecting discrepancy will be introduced. We found no statistically significant differences in agreement between the examiners who were formally trained or between them and the examiner who was not.

Ultrasonography is used in the management of DDH both to identify normal and abnormal hips, to determine the severity of the dysplasia, and to monitor treatment. The high level of agreement in the identification of normal hips is encouraging, since it reinforces the use of ultrasonography as a method of screening. The reliability in classifying abnormal hips was significantly lower and methods to improve this should be explored. Our study did not attempt to assess the accuracy of the ultrasound examination. To do this would need the establishment of formal standards which are beyond the scope of our study.

No benefits in any form have been received or will be received from a commercial party related directly or indirectly to the subject of this article.

References