Comparing ultrasound measurements of neonatal hips using the methods of Graf and Terjesen

A. Falliner, D. Schwinzer, H.-J. Hahne, J. Hedderich, J. Hassenpflug
From University Hospital Schleswig-Holstein, Kiel, Germany

In a prospective study, 232 neonates were examined sonographically using the methods of Graf and Terjesen. In order to determine the reproducibility of the methods, 50 hips were evaluated by two skilled examiners. In an inter-observer study, five physicians and five medical students evaluated 24 images, which were evaluated on ten occasions at two-weekly intervals by one of the authors. Statistical evaluation used the Bland-Altman approach.

The neonates (110 females, 122 males) were less than four days old. The mean α angle was 62.4˚ and mean femoral head cover was 55.4%. According to Graf’s method, 1.3% of hips were pathological, compared with 4.1% according to Terjesen. Spearman’s correlation coefficient between femoral head cover and α angles was 0.552. The Bland-Altman approach shows greater variation for femoral head cover than for α, if measured by experienced examiners. The Bland-Altman approach shows almost equal reproducibility for α and femoral head cover in the inter-observer test, but better repeatability for α in the intra-observer test.

The Graf results relate better than Terjesen’s to the well-known frequency of 1% to 2% hip dysplasia in the European population. Kappa statistics indicate a fair agreement between the two methods. Inter-observer evaluation shows an equal reproducibility of both methods, whereas intra-observer tests reveal better repeatability with Graf’s method.

The foundations for the sonographic diagnosis of hip dysplasia in the newborn were described by Graf, and his method was rapidly established in the German-speaking countries.

In the English-speaking world and Scandinavia, Harcke et al2 and Terjesen, Rundén and Tangerud3 and Terjesen, Bredland and Berg4 developed their own method of investigation. Whereas Graf’s method measures two angles, α and β, Harcke et al2 and Terjesen et al3,4 determine the amount of the femoral head covered by the osseous part of the acetabulum.

Ultrasound measurement of anatomical specimens in a water bath gave comparable reproducibility for the two methods, but so far there have been few such clinical studies.

This study aimed to compare the Graf and Terjesen methods and to clarify whether, under clinical conditions, differences in the reproducibility of the methods can be assessed. Intra- and inter-observer variability is demonstrated by different examiners with variable experience in hip sonography and by repeated evaluations by one of the authors (DS).

Materials and Methods
In a prospective study between February and September 2003, 232 consecutive neonates were investigated sonographically by the methods of Graf and Terjesen. The examinations were performed with a 7.5 MHz linear transducer. The images were taken such that the landmarks of both methods could be identified. Documentation was by a video printer (Sony Video Graphic Printer UP 890 MD; Sony, Tokyo, Japan) with a magnification factor of 1.4 (Fig. 1) and the images were evaluated as follows.

Correlation between Graf and Terjesen. Direct comparisons of the α angle and the linear variables a and b for the determination of femoral head cover were made by one examiner from the same image. We did not evaluate the β angle because it is not relevant in normal hips (more than 98% of our hips) and we wished to limit our data. Correlation analysis used Spearman’s coefficient because the data were not parametric. In order to assess agreement between the two methods with respect to clinical outcome, the Graf type was considered
either normal (types I and IIa) or pathological (IIc, D, IIIa) and a kappa (K) statistic was calculated.

Reproducibility in experienced hands. Two of the authors (AF, DS) twice evaluated 50 hips from different images according to the methods. Reproducibility was analysed by the Bland-Altman method. Based on this approach, the limits of agreement were determined by the mean difference plus or minus the coefficient of reproducibility (CRa), where CRa was calculated as 1.96 x (standard deviation (SD) of the differences).

Inter-observer reliability. Ten examiners – five sonographically experienced physicians and five medical students with one hour’s explanation of the methods and analysis, each evaluated 24 images according to both methods. The means and SDs of the measurements for each hip were calculated separately for the physicians and the students. Each standard deviation was plotted against the relevant mean in a Bland-Altman plot. In order to compare results, an intra-class correlation coefficient (CRb) was calculated as:

\[ 2.83 \times \sqrt{\text{mean of the variances}} \]

Results
We investigated 232 neonates (age < four days, 110 female, 122 male). Of the 464 hips, six (1.3%) were pathological according to the Graf method and 19 (4.1%) according to the Terjesen method (Table I) and κ statistics indicate a fair agreement between the methods. The mean α angle was 62.4° (95% confidence interval (CI) 52.0 to 70.0) and mean femoral head cover value was 55.4% (95% CI 42.6 to 65.6).

Correlation between Graf and Terjesen. The plot of α angles against femoral head cover values shows the correlation between these values, which are differently scaled (Fig. 2). Spearman’s correlation coefficient is 0.55, with a 95% CI from 0.48 to 0.62.

Reproducibility in experienced hands. The Bland-Altman approach of double measurement of the α angles and femoral head cover values by the two experienced investigators shows a larger variability for femoral head cover (CRa 10.25) than for α angle (CRa 6.15).

Inter-observer reliability. The Bland-Altman approach indicates no obvious difference between the physicians and the medical students. The intra-class correlation coefficient (α) was 0.72 for the physicians and 0.74 for the students and the intra-class correlation coefficient (femoral head cover) was 0.61 for the physicians and 0.77 for the students.
Intra-observer reliability. An author’s (DS) measurements of 24 hips show better repeatability for Graf (CR₉, 9.56) than Terjesen (CR₉, 8.22; Table II).

Discussion
Although the Graf¹ and Terjesen et al³⁴ methods are widely used in Europe and have similar sonographic standards, there are few studies that compare them clinically. Czubak et al¹¹ investigated 657 neonates and found a correlation of 0.57 (p < 0.01) between the α angle and femoral head cover. Our measurements are nearly the same. Comparison of α angles with femoral head cover values can only be approximate because Graf measures an angle whereas Terjesen calculates femoral head cover. This depends not only on the acetabulum but also on the position of the femoral head. Czubak et al¹¹ found 29% of hips to be physiologically immature according to Graf, whereas only 14% were ‘suspected’ dysplastic hips according to Terjesen. He also infers from this that the latter method has better specificity. Conversely, our results show a higher portion of dysplastic hips according to Terjesen (4.1%) than with Graf (1.2%). In our opinion, the latter value better matches the well-known frequency of hip dysplasia in 1% to 2% in the European population.¹²

Irha et al¹³ reported a better sensitivity using the Graf method than that of Harcke, which uses the same variables as Terjesen. They conclude that measurement of angles is better for the diagnosis and classification of pathological hips.

In our comparison of reproducibility, we found a greater variability between the two experienced investigators for femoral head cover values than for α angles.

Inter-observer tests for Graf’s method were also performed by Wiese⁴¹ and Roovers et al.¹⁵ Wiese⁴¹ found a variability of 6.5° for the α angle and of 13.3° for the β angle, whereas Roovers et al¹⁵ found the inter-observer SD to be 3.2° for the α and 6.0° for the β angle. Because we compared not only different examiners using one method, but also two differently scaled methods, we had to use other statistical techniques. In order to compare the inter-observer reliability of Graf’s and Terjesen’s methods we had to compute the intra-class correlation coefficient.⁹ Surprisingly, we found no outstanding difference between the experienced physicians and the medical students in the reliability of α angles. For femoral head cover values the students showed better reliability than the physicians. Our results show a funnel effect, meaning there is more variability when the magnitude of the measurements is greater. As in our test for reproducibility the intra-observer test by one of the authors shows a better reliability of α angles than for femoral head cover.

Supplementary Material
A further opinion by Mr Mark Paterson is available with the electronic version of this article on our website at www.jbjs.org.uk

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

Table II. Inter-class correlation coefficient (ICC) with 95% confidence interval (CI) for inter- (physicians and students) and intra-observer tests (DS). Inter-observer tests show nearly equal reliability for α angles but better reliability for femoral head cover in the student group. Intra-observer tests reveal a slight superiority of Graf’s method

<table>
<thead>
<tr>
<th>Method</th>
<th>Inter-observer Physicians</th>
<th>Intra-observer Students (DS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>α</td>
<td>0.72</td>
<td>0.74</td>
</tr>
<tr>
<td>95% CI</td>
<td>0.65 to 0.85</td>
<td>0.77 to 0.93</td>
</tr>
<tr>
<td>Femoral head cover</td>
<td>0.608</td>
<td>0.773</td>
</tr>
<tr>
<td>95% CI</td>
<td>0.43 to 0.78</td>
<td>0.64 to 0.88</td>
</tr>
</tbody>
</table>

α: Angle; CI: Confidence interval; DS: Doctor’s score.