We studied the reproducibility of ultrasonographic screening examination of the hip when read by diagnostic radiographers. In order to determine interobserver variability, 200 ultrasonograms were classified according to Graf’s method by five observers (four radiographers and one radiologist). The kappa values for interobserver variability indicated moderate agreement (kappa 0.47) for the exact Graf classification and substantial agreement (kappa 0.65) for the classification of normal (type I) versus abnormal (type IIa-IV). Agreement was significantly different for normal, immature and abnormal hips. Comparison of the findings in our interobserver study with existing information based on other examinations and treatment revealed that only a small number of infants with mildly dysplastic hips would have been typed as normal by some observers as a result of observer variability. In conclusion, the interobserver agreement on the ultrasound assessment of the hip was good enough for screening purposes. Observer variability did not result in any severe cases being missed.

Received 1 October 2002; Accepted after revision 11 March 2003
Table I. Graf’s classification of 189 ultrasonograms by five observers

<table>
<thead>
<tr>
<th>Hip type</th>
<th>Radiographer 1 (%)</th>
<th>Radiographer 2 (%)</th>
<th>Radiographer 3 (%)</th>
<th>Radiographer 4 (%)</th>
<th>Radiologist (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I (normal)</td>
<td>70</td>
<td>73</td>
<td>74</td>
<td>74</td>
<td>62</td>
</tr>
<tr>
<td>IIA/IIA+IIA- (immature)</td>
<td>15</td>
<td>14</td>
<td>15</td>
<td>14</td>
<td>22</td>
</tr>
<tr>
<td>IIb/IIc/D (minor dysplasia)</td>
<td>10</td>
<td>7</td>
<td>6</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>IIIa/IIIb/IV (major dysplasia)</td>
<td>4</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

Table II. Interobserver agreement between radiographer 1 and radiographer 2

<table>
<thead>
<tr>
<th></th>
<th>Radiographer 1</th>
<th>Radiographer 2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>124</td>
<td>9</td>
<td>133</td>
</tr>
<tr>
<td>Abnormal</td>
<td>21</td>
<td>46</td>
<td>67</td>
</tr>
<tr>
<td>Total</td>
<td>145</td>
<td>55</td>
<td>200</td>
</tr>
</tbody>
</table>

Patients and Methods

Between October 1998 and July 2000, more than 5000 infants were screened at different ages within the framework of the Soundchee Study, comprising over 28,000 hips. The examinations were performed by diagnostic radiographers under the supervision of the project team (radiologist, orthopaedic resident, orthopaedic surgeon and child health-care physician). A portable sonograph (Hitachi EUB-405; Hitachi Medical Systems BV, Reewijk, The Netherlands) with a linear array transducer operating on an ultrasound frequency of 7.5 or 5 MHz was used.

Inter- and intraobserver variability. For the study of interobserver variability, 119 normal (Graf type I), 31 immature (Graf type Ila, Ila+ or IIa-) and 50 dysplastic (Graf type IIb or worse) images were selected on the basis of the original assessment of the ultrasonogram. The observers did not know the distribution of the types of hip. The median age of the infants was 11 weeks (3 to 21). The 200 images were read by four diagnostic radiographers and one radiologist. Two of the diagnostic radiographers and the radiologist had attended a hip ultrasonography course given by Professor Graf, whereas the other two diagnostic radiographers had been trained by the project team according to Graf’s principles. All the images were projected as right-sided and presented on a standard classification form with Graf’s principles. All the images were projected as right-sided and presented on a standard classification form with Graf’s principles.

Interobserver agreement was estimated by calculating the proportion of agreement between the observers. The kappa value for each pair of observers ranged from 0.68 to 0.79, with kappa values ranging from 0.43 to 0.54. The multirater kappa value was 0.47 (95% confidence interval (CI) 0.40 to 0.55). According to Landis and Koch, these kappa statistics indicate moderate agreement. Since the exact Graf classification is more detailed than necessary for screening purposes, the ultrasonograms were grouped into ‘normal’ (type I) and ‘abnormal’ (types IIa to IV and not readable) in the subsequent analyses. As an example, the agreement between radiographers one and two is presented in Table II. The proportion of interobserver agreement of this pair was 0.85, with a kappa value of 0.65. For all ten observer pairs, the proportion of agreement ranged from 0.82 to 0.91, with kappa values of 0.60 to 0.76 (Table III).

For the different observer pairs, the proportion of agreement on the exact Graf classification (10 categories) ranged from 0.68 to 0.79, with kappa values ranging from 0.43 to 0.54. The multirater kappa value was 0.47 (95% confidence interval (CI) 0.40 to 0.55). According to Landis and Koch, these kappa statistics indicate moderate agreement. Since the exact Graf classification is more detailed than necessary for screening purposes, the ultrasonograms were grouped into ‘normal’ (type I) and ‘abnormal’ (types IIa to IV and not readable) in the subsequent analyses. As an example, the agreement between radiographers one and two is presented in Table II. The proportion of interobserver agreement of this pair was 0.85, with a kappa value of 0.65. For all ten observer pairs, the proportion of agreement ranged from 0.82 to 0.91, with kappa values of 0.60 to 0.76 (Table III).

Statistical analysis. Inter- and intraobserver and interexaminer agreement was estimated by calculating the proportion of agreement and Cohen’s kappa for each pair of observers. In accordance with the medical literature, the arbitrary classification of Landis and Koch for kappa was used for interpretation. For interobserver agreement, the multirater kappa was also calculated. Systematic differences in measuring the alpha and beta angles between the observers were determined by comparing the mean values and standard deviations, and calculating differences in mean values and standard deviations. Variability in the measurements of the alpha and beta angle was assessed by the within-subject standard deviation.
The multirater kappa value was 0.65 (95% CI 0.58 to 0.71). There were significant differences in agreement between the assessments of normal, immature and abnormal hips. In the hips typed as normal at the original assessment, all observers read the hips as the same in 84%; agreement in the immature and abnormal hips was 13% and 66%, respectively (Table IV).

Intraobserver agreement of the radiographers ranged from 0.82 to 0.94, with kappa values of 0.64 (95% CI 0.52 to 0.77) to 0.86 (95% CI 0.77 to 0.95). Inter- and intraobserver kappa statistics indicated substantial agreement, while radiographer four had almost perfect intraobserver agreement (Table III).

Inter- and intraobserver variability of alpha and beta angles. Alpha and beta angles were measured by all four radiographers on 187 and 174 ultrasonograms, respectively. Missing values occurred because it was not always necessary to measure alpha and beta angles to characterise the hip, or because it was considered that the image had not been taken in the correct sectional plane. There were no systematic differences between the mean alpha angles. The mean beta angle of radiographer four was slightly smaller than those of radiographers one and two, while that of radiographer three was slightly larger (Table V). Wide variability in the measurements of alpha and beta on the same ultrasonogram was found. Mean interobserver within-subject standard deviations for the alpha and beta angles were 3.2° and 6.0°, respectively. There were no differences between the mean values of the measurements taken by the same observer. Intraobserver within-subject standard deviations for the alpha and beta angles ranged from 2.4° to 3.5° and from 4.5° to 4.8°, respectively (Table VI).
Interexaminer variability. Two radiographers examined 24 infants consecutively for the study on interexaminer variability. Because of the low incidence of DDH and since it was not possible to make a selection beforehand, most of the hips were normal. The radiographers agreed on 42 normal hips and three immature or abnormal hips. Two hips were typed as normal by one radiographer, but as immature by the other, while one hip was typed as normal by one radiographer, but as abnormal (type IIb) by the other. The proportion of agreement was 94.8%, with a kappa value of 0.63, which corresponded to substantial agreement according to Landis and Koch.13 However, chance played a major role in this estimate (95% CI for kappa 0.25 to 1.00).

Discussion

In our study, interobserver, intraobserver and interexaminer variability of ultrasonographic examinations of the hip were determined. The kappa values of interobserver and intraobserver agreement indicated substantial agreement based on the dichotomy of 'normal' versus 'abnormal'. There was very little difference between the intra- and interobserver kappa values, although we expected higher intraobserver agreement. The relatively low intraobserver kappa values can be explained by the higher percentage of immature and dysplastic hips in the intraobserver study. When the interobserver kappa values were calculated on the same set of images, the kappa values decreased slightly (range 0.56 to 0.75). Interexaminer agreement was also substantial, but the CI of this kappa value was very wide. There was no systematic difference in the measurements of the alpha and beta angles between the different observers, or by the same observer. However, variability of the angles measured on the same ultrasonogram was high, both between observers and by the same observer. These results support the conclusion of others that variance in alpha and beta angles is large, even when the angles are measured by the same observer.7-11

The radiologist read more of the ultrasonograms as immature or abnormal than the radiographers, although he had attended the same ultrasonography course and the radiographers were working under his supervision during the screening project. An explanation could be that he had higher a priori expectations of DDH based on his clinical experience. The effect on the percentages of agreement and kappa values was negligible. Our findings cannot be related directly to those of others, because it is not always clear how the different types of hip were distributed and because the hips were classified into a different number of groups. Bar-On et al10 and Ömeroglu et al9 found interobserver kappa values of 0.57 and 0.46, which were slightly lower than our kappa values. Ömeroglu et al9 classified the hips into three groups, which could explain their lower kappa values. The intraobserver kappa values reported by Bar-On et al10 and Ömeroglu et al9 were comparable with those of three of our radiographers. Rosendahl et al8 had higher intraobserver agreement, with kappa values of 0.7 and 0.8, despite division into four groups; their interexaminer kappa value was 0.5. Our interexaminer kappa value was higher, but had a wide CI. Rosendahl et al8 did not mention the CI of their kappa value, but it was probably smaller, because of the larger number of infants examined.

The ultrasonograms used in the intra- and interobserver study were taken within the framework of the Soundcheck Study. Therefore we also had existing patient information on other examinations and subsequent treatment. Comparison of this information with the findings in the present interobserver study enabled assessment of the consequences of observer variability. None of the hips which were typed as normal by all the observers had been treated; 19 of the 35 treated hips were judged to be abnormal by all the observers. In four infants, the hip was typed as normal by some of the observers, but the ultrasonogram of the contralateral hip, which was also selected for the study, was typed as abnormal by all of them. As a result, if both hips had been taken into account, these infants would not have been missed by the screening. In the remaining 12 cases, the hips were typed as normal by some of the observers and as type Ia, Ia+, Ia- or IIb by the others, showing that the missed cases were hips with less severe abnormalities.

In conclusion, inter- and intraobserver agreement of ultrasound assessment of the hip was sufficient for screening purposes. Observer variability did not result in any severe cases being missed. Although interexaminer agreement may be slightly lower, we conclude that, if examiners are well trained and supervision is arranged to ensure the quality of examination, ultrasound assessment of the hip has good potential as a screening tool for DDH. It remains to be seen whether the actual application of this tool in a screening programme will effectively reduce morbidity due to DDH.

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


