Pediatr Radiol (2010) 40:1634–1639
DOI 10.1007/s00247-010-1689-4

ORIGINAL ARTICLE

Evaluation of a training program for general ultrasound screening for developmental dysplasia of the hip in preventive child health care

S. Ramwadhdoebe · R. J. B. Sakkers · Cuno S. P. M. Uiterwaal · Magda M. Boere-Boonekamp · Frederik J. A. Beek

Received: 3 August 2009 / Revised: 12 March 2010 / Accepted: 21 March 2010 / Published online: 12 June 2010

Abstract
Background A research study in the Netherlands showed that general ultrasound (US) screening was cost-effective in the detection of developmental dysplasia of the hip (DDH). This study was followed by a pilot implementation study. Part of this pilot implementation study is to investigate whether professionals of the infant health care (IHC) system, with no previous US experience, would be able to perform US of the hip.
Objective This study looks at health care worker ability to classify US images into a modified Graf system.
Materials and methods After theoretical and practical training, seven nurses and physicians of the participating IHC centers reported their findings on sonographic images of 80 children. This was repeated five months later. From the two evaluation moments the intraobserver agreement and the interobserver agreement was determined.
Results The average estimated interobserver Cohen's kappa for both sessions was for nurses 0.6 and for physicians 0.5. The second evaluation showed a decrease from an average of 4.3% missed cases per screener to 2.3% and an increase of an average of 5% false positives per screener to 9.1%.
Conclusion The inter- and intra-observer agreement is comparable to similar studies in which the participants had a professional background in US examination. The level of agreement of the trainees in the perspective of the screening process was considered sufficient for the pilot implementation project.

Keywords Developmental dysplasia of the hip · Ultrasound · Child

S. Ramwadhdoebe (✉) · R. J. B. Sakkers
Department of Orthopedics, UMC Utrecht,
Heidelberglaan 100,
3584 CX Utrecht, The Netherlands
e-mail: s.ramwadhdoebe@umcutrecht.nl

R. J. B. Sakkers
e-mail: r.sakkers@umcutrecht.nl

C. S. P. M. Uiterwaal
Julius Center for Health Sciences and Primary Care,
UMC Utrecht,
Utrecht, The Netherlands

F. J. A. Beek
Department of Radiology, UMC Utrecht,
Utrecht, The Netherlands

M. M. Boere-Boonekamp
School of Governance, University Twente,
Enschede, The Netherlands

Introduction

In the Netherlands, all children regularly visit infant health care centres (IHC) in the first four years of life in the context of the national program for child health surveillance. Screening for developmental dysplasia of the hip (DDH) is part of this program. The current screening consists of physical examination by infant health care physicians. Physical examination for DDH is done at the ages; 1 month and 3 months. Babies with abnormal findings are referred. Those with risk factors are referred to a hospital at the age of three months because imaging at an earlier age leads to a large number of immature and marginally dysplastic hips which are likely to normalize spontaneously (Graf 2a/c) [1, 2].

As this screening for DDH by physical examination is strongly challenged in terms of its cost-effectiveness in the context of missed cases and a high percentage of false positives, alternatives are being explored based on the results of a former US study in the Netherlands [1]. One

alternative would be general US screening according to Graf's method for the detection of DDH [3–8]. A Dutch study in several IHCs, in which sonographers screened all children for DDH with US, indicated a sufficiently high agreement with a certified paediatric radiologist to allow for cost-effective screening by the technicians as compared to the current method of screening [9]. The next logical step would be to start a pilot implementation program to explore all possible organizational problems for national implementation of US screening for DDH at IHCs. Therefore, a pilot implementation study was sponsored by government funds (ZonMW nr 5501) in two IHC regions in the Netherlands. Since there are currently no sonographers employed at IHCs and the number of experienced sonographers is too small to execute a national implementation program of US screening it was decided to start a training program for nurses and physicians of the IHCs in the pilot regions and to investigate whether these IHC professionals would be able to perform US of the hip. This report describes the evaluation of the trainees within the first month and five months later, the intra- and interobserver variation of the trainees, and discusses this evaluation of training in the context of the pilot implementation study in which 4,600 infants are invited for US screening in one urban and one rural region in the Netherlands.

Materials and methods

The implementation program was started in 15 IHCs in a rural region and an urban region in the Netherlands. The exam protocol was reviewed by the Medical Ethics Board which judged that approval was not necessary because no tests were applied on humans that were subject to the law on medical scientific research.

The screening team consisted initially of three IHC physicians and four IHC nurses. A Terason T3000 (Terason Ultrasound, Burlington, MA, USA) equipped with a 5– to 12–mHz linear array probe was used.

The training program was provided by a paediatric radiologist and a paediatric orthopaedic surgeon. First a theoretical training of two days was given. This theoretical training was centered on pathology of the hip, the theory of Graf, the recognition of Graf type 1, 2a, 2b, 2c, D, 3 and 4, and on practical aspects of US screening. Hereafter, the trainees started with the general US screening of children at the IHCs. In the first four months of the screening period the trainees worked side by side with an experienced sonographer. During this period each trainee performed approximately 350 US hip examinations under supervision. An additional theoretical training of half a day was given three months after the beginning of training, focusing on missed and equivocal cases which were selected from the first three months of actual screening. In the first month of the training and four months later (1 month after the end of screening under supervision), the skills of the trainees were tested with an exam consisting of images of 80 subjects printed in an exam book.

All images were retrieved from a hospital picture and archiving communications system. For the images, 80 children with an age between three and four months were selected. The total selection comprised: Graf type 1 (n=35), Graf 2b (n=14), Graf 2c (n=6), Graf type D (n=6), Graf type 3 (n=7), Graf type 4 (n=6), and six cases were "technically insufficient". Images were defined as "technically insufficient" when the images were too dark or had insufficient penetration. The sonograms also needed to be sufficient according to Graf criteria for a correct scan plane, dictating that three points need to be visible: the lower limb of the bony ilium, the midportion of the acetabular roof, and the acetabular labrum [10]. In the second session the exam book consisted of the same images in a different order. The screeners were instructed to evaluate and classify all 80 cases.

Analysis

The aim of the screening program is to differentiate between normal and abnormal hips and hips were classified as such. To allow further distinction between minor and major types of dysplasia the screeners were instructed to subdivide the abnormal hips into minor and major dysplasia. Graf 2b and 2c hips were classified as minor and hips Graf D, 3 and 4 as major dysplasia. Furthermore we wanted to see whether the participants were able to distinguish a technically sufficient from an insufficient image which led to the addition of the category technically insufficient. This results in a classification with four categories: (1) normal, (2) DDH without (sub)luxation, (3) DDH with (sub)luxation, and (4) "technically insufficient image". The participants had to classify the hip on morphological grounds added by their measurement of the alpha and beta angles. The classification was standardized with the alpha and beta angles. It was decided that in the screening process in the implementation study every child with a hip with an alpha angle <60° was to be referred. For the evaluation of the two sessions, no additional dynamic test was done [11].

When the participants completed the first exam book, their experience was limited to the theoretical training which included approximately 50 cases and approximately 70 sonographic hip examinations done at the IHC in the first month of supervised screening. At the time of completion of the second book the participants had performed an additional 350 screening examinations.

Based on the results of the exam books the inter- and intra-observer agreement of the screeners was determined and evaluated with the use of Kappa statistics. Cohen's kappa was estimated and interpreted based the Landis and

Kochs [12, 13] classification.The results of the screeners were also categorized in missed cases and false positives in the context of the screening process.

Results

An overview of the outcomes of the individual trainees is given in Table 1.

On average the nurses had a slightly higher and more consistent score as compared to the physicians (Fig. 1).

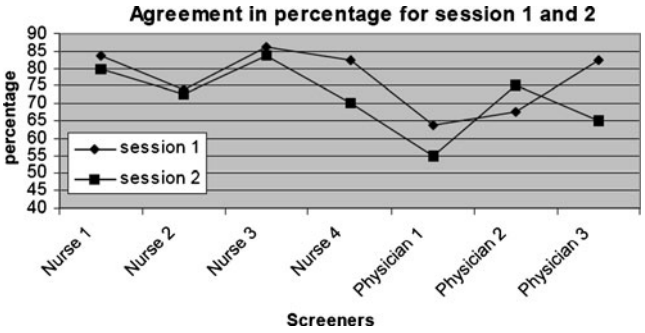

Fig. 1 The graph shows agreement in percentage for sessions 1 and 2

Outcomes of the first evaluation centered on the diagnosis

In session one, 23 cases of the 140 in the category DDH without (sub)luxation were classified as normal. One of the 133 cases in the category DDH with (sub)luxation was classified as normal. On the other hand, 26 of the 245 normal cases were classified as DDH without (sub)luxation, one of the 245 as DDH with (sub)luxation, and 17 of the 245 cases were classified as "technically insufficient".

Outcomes of the second evaluation centered on the diagnosis

In session two, six of 140 cases in the category DDH without (sub)luxation were classified as normal as well as seven of the 42 cases that were "technically insufficient". On the other hand, 44 of the 245 normal hips were classified as DDH without (sub)luxation, two of the 245 as DDH with (sub) luxation) and 20 of the 245 normal cases were judged "technically insufficient".

Statistical analysis

On average the screening nurses had the highest intra- and inter-observer kappa; (Table 2). The average agreement among the screeners was highest for the category "technically insufficient image" and lowest for the category DDH with (sub)luxation for both sessions.

Results in the context of the screening process

The first evaluation resulted in an average of 4.3% missed cases per screener (1.25–7.5%). Of these missed cases 96% were in the category DDH without (sub)luxation (Fig. 2). The screeners considered an average of 5% of the images as abnormal (1.25–10%) in disagreement with the exam book diagnosis and these cases would have been false positives in the actual screening process. Also here, 96% of these false positives were in the category DDH without (sub) luxation. An average of 3% per screener (0–8.75%) was classified as "technically insufficient" instead of normal. In the screening setting a child would have been invited for a second US screening.

The second evaluation showed a decrease from an average of 4.3% missed cases per screener to 2.3% (0–5%) and an increase of an average of 5% false positives per screener to 9.1% (0–21.25%). The average percentage of "technically insufficient" classifications of normal images increased from 3 to 3.6% (0%–7.5%).

Examples of US images obtained are shown in Fig. 3.

Table 1 Hip classification of 160 images by seven screeners

Hip type (number according to gold standard in parentheses)	Nurse 1	Nurse 2	Nurse 3	Nurse 4	Physician 1	Physician 2	Physician 3
Session 1							
Normal (35)	30	38	34	31	33	25	34
DDH without (sub)luxation (type 2b/c) (20)	27	18	23	26	17	17	18
DDH with (sub)luxation (type D and 3/4) (19)	15	5	15	18	2	13	12
Technically insufficient (6)	8	19	8	5	28	25	16
Session 2							
Normal (35)	33	26	29	18	34	29	24
DDH without (sub)luxation (type 2b/c) (20)	22	17	30	41	20	14	18
DDH with (sub)luxation (type D and 3/4) (19)	19	11	17	18	13	17	11
Technically insufficient (6)	6	26	4	3	13	20	27

Table 2 Analysis of inter- and intraobserver agreement of the first and second session for four classes normal DDH without (sub) luxation, DDH with (sub)luxation and technically insufficient

	Session 1 Interobserver kappa mean, (SD)	Session 2 Interobserver kappa mean, (SD)	Intraobserver kappa mean, (SD)
Nurse 1	0.66, (0.0692)	0.53, (0.0708)	0.48, (0.0696)
Nurse 2	0.54, (0.0626)	0.69, (0.0706)	0.7, (0.0694)
Nurse 3	0.69, (0.07)	0.48, (0.0652)	0.49, (0.0663)
Nurse 4	0.62, (0.071)	0.55, (0.06)	0.78, (0.0705)
Physician 1	0.4, (0.0581)	0.51, (0.052)	0.24, (0.0608)
Physician 2	0.48, (0.0605)	0.54, (0.0635)	0.54, (0.0646)
Physician 3	0.66, (0.0659)	0.42, (0.0592)	0.59, (0.0652)

Discussion

The hiring and education of IHC personnel for US screening of the hip for DDH in the Dutch national IHC program is under investigation in a government subsidized pilot implementation program. One of the first steps was an evaluation of the training program of the IHC personnel. Both IHC nurses and IHC physicians were hired from the existing staff of the IHCs to evaluate the abilities of both categories of IHC personnel.

In the evaluation of the training program no major difference was seen between the performance of physicians and nurses. The level of general medical education does not seem to influence the capability of these health care professionals to recognize images of the different types of hip dysplasia. The drawing of conclusions from this evaluation is of course limited by the use of an exam book which only consists of static images. For example, it might be difficult for the trainees to differentiate between DDH with hip luxation and a "technically insufficient" image since it was not possible to check the correct plane of the image or the position of the femoral head by moving the probe. However, as a first step in the evaluation it seems logical to assess whether the trainees can distinguish a static image of a normal hip from an abnormal hip.

The great value attached to the alpha angle can be questioned. Morphological features and stability of the hip are of great importance. In the outline of the implementation study it was decided that referral of a child should be based on a quantitative parameter: the alpha angle. This will undoubtedly lead to referral of stable, morphologically normal hips with an alpha angle of 59° but it provides the investigators with an "objective" parameter.

One can question whether general US screening for DDH is effective. One would expect that general screening lowers the number of children with late detected DDH [1]. Indeed, the percentage of first operative procedures for DDH is decreased after introduction of general US screening [14]. Another advantage would be that the number of children referred because of findings at physical examination, such as limited abduction and leg shortening, or because of risk factors like breech delivery or a positive family history, will decrease. These referrals are costly and general screening should limit this number [15, 16]. With the current screening in the Netherlands approximately 22% of all children are referred for imaging in a radiological hospital department [1]. On the other hand general US screening will lead to referral of children with marginal DDH that is likely to improve spontaneously. There will be overtreatment approaching 5% and of these children approximately 1% could suffer complications such as avascular necrosis of the femoral head [14]. In all large studies on US screening a small number of false-negative cases is present [17]. Overall the choice between general or selective screening is still a matter of debate [4, 8, 11, 18]. Regarding the Dutch situation: in the Netherlands a study was conducted with general US screening at the ages of one, two and three months. A US examination at eight months was done to detect possibly missed cases. The study showed that a general US screening at the age of three months would be most effective with the best equilibrium between delayed treatment and overtreatment. In this strategy physical examination at one month of age is still performed.

A limitation of our study is that it addressed only the intra- and inter-observer variability in reading a sonogram, while in the actual screening process variability can be present both in recording and in reading a sonogram. It is reported that the variability in recording is higher than in

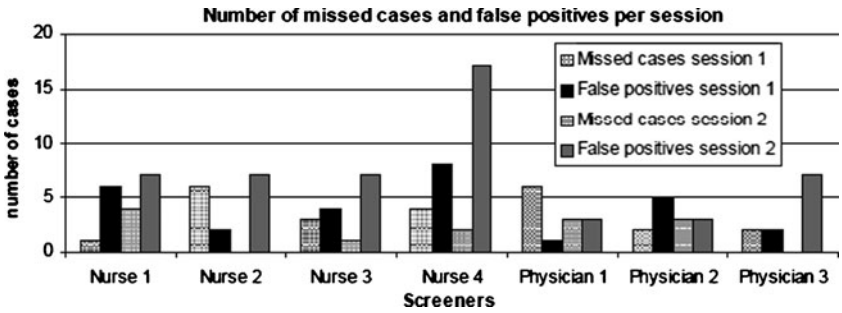

Fig. 2 The graph shows the number of missed cases and false positives per session

Fig. 3 Examples of hip US images. **a** Dysplasia with subluxation. Two screeners classified the image as dysplasia without subluxation. **b** Morphologically normal hip. The alpha angle is approximately 60°. Classified as normal by four screeners and as dysplasia without subluxation by three screeners. **c** The lower limb of the iliac bone is not well seen. The image was classified as technically insufficient by all screeners

reading [9, 19]. Another limitation is the rather small number of hips that can be realistically examined and the use of an exam book and the small number of screeners. On average the level of agreement with the paediatric radiology of the individual children was relatively high and stable, especially when compared to the scarce amount of other studies on this subject [9, 13, 20, 21]. The amount of inter- and intra-observer agreement in this study is comparable to those studies in which the participants had a professional background in US examination.

In other studies regarding the inter- and intra-observer variation in US diagnosis of DDH, the kappa estimates of agreement are generally between 0.5 and 0.7 [9, 13, 20, 21]. When using an exam book, as in this evaluation of a training program, these outcomes immediately put the gold standard in perspective. All inter- and intraobserver variation studies show a certain disagreement between normal and abnormal in borderline cases. Most disagreement in the context of missed cases in this evaluation was between the categories normal and DDH without (sub) luxation, which is in agreement with the normal effects of inter- and intraobserver variation.

The accuracy in terms of false positives and false negatives varied between participants but also between sessions. The total number of errors with respect to missed cases and false positives was 52 at the first session and 64 at the second session (Fig. 2), with a shift from an excess of false negatives (24 in session 1 and 13 in session 2) to false positives (28 in session 1 and 51 in session 2). It appears that in the second session trainees improved their skills in detecting cases of DDH with (sub)luxation. The implications for actual screening practice would therefore imply that more children would be incorrectly referred to hospital based on the screening diagnosis. On the other hand, there are fewer cases of late detected children with DDH who could have been treated earlier. We expected that the results of the screeners would show a learning curve with fewer missed cases and fewer false positives. However the overall result was disappointing with no real improvement evident. The increased number of false-positive cases in the second session signifies that close monitoring of the screening process with continual education and training are necessary. A possibility is to facilitate training on the job once every three months where screeners receive additional individual supervision. Regular meetings with fellow screeners and with orthopaedic surgeons during which cases and images are discussed and feedback is given could also enhance their skills and increase their self-confidence in borderline cases. If general US screening is implemented, one could consider a yearly test that should be passed.

The level of agreement of the trainees in this perspective may be considered sufficiently adequate for the implementation project. The outcomes of this evaluation of the training program cannot be predictive for the outcomes of the pilot implementation study as mentioned in the introduction. In this study, the trainees will perform US screening in more than 9,000 hips in children in a period of 2.5 years with a different distribution of the types of DDH among a vast majority of normal hips. The outcomes will represent not only the ability of the screeners to recognize hip pathology on an US image, but also the ability to obtain a level of quality in the depiction of the hip joint.

Conclusion

The amount of inter- and intra-observer agreement in this study is comparable to similar studies in which the participants had a professional background in US examination. The level

of agreement of the trainees in the perspective of the screening process was considered sufficiently adequate for the pilot implementation project. However, the increased number of false-positive cases in the second session signifies that close monitoring of the screening process with continual education and training are necessary.

References

1. Roovers EA, Boere-Boonekamp MM, Castelein RM et al (2005) Effectiveness of ultrasound screening for developmental dysplasia of the hip. Arch Dis Child Fetal Neonatal Ed 90:F25–30
2. Marks DS, Clegg J, al-Chalabi AN (1994) Routine ultrasound screening for neonatal hip instability. Can it abolish late-presenting congenital dislocation of the hip? J Bone Joint Surg Br 76:534–538
3. Wirth T, Stratmann L, Hinrichs F (2004) Evolution of late presenting developmental dysplasia of the hip and associated surgical procedures after 14 years of neonatal ultrasound screening. J Bone Joint Surg Br 86:585–589
4. Toma P, Valle M, Rossi U et al (2001) Paediatric hip–ultrasound screening for developmental dysplasia of the hip: a review. Eur J Ultrasound 14:45–55
5. Shipman SA, Helfand M, Moyer VA et al (2006) Screening for developmental dysplasia of the hip: a systematic literature review for the US preventive services task force. Pediatrics 117:e557–576
6. Rosendahl K, Markestad T, Lie RT (1996) Developmental dysplasia of the hip. A population-based comparison of ultrasound and clinical findings. Acta Paediatr 85:64–69
7. Luhmann SJ, Bassett GS, Gordon JE et al (2003) Reduction of a dislocation of the hip due to developmental dysplasia. Implications for the need for future surgery. J Bone Joint Surg Am 85-A:239–243
8. Dezateux C, Rosendahl K (2007) Developmental dysplasia of the hip. Lancet 369:1541–1552
9. Roovers EA, Boere-Boonekamp MM, Geertsma TS et al (2003) Ultrasonographic screening for developmental dysplasia of the hip in infants. Reproducibility of assessments made by radiographers. J Bone Joint Surg Br 85:726–730
10. Graf R (2006) Hip sonography: diagnosis and management of hip dysplasia, 2nd edn. Springer, Berlin, pp 28–29
11. Keller MS, Nijs EL (2009) The role of radiographs and US in developmental dysplasia of the hip: how good are they? Pediatr Radiol 39:S211–S215
12. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174
13. Simon EA, Saur F, Buerge M et al (2004) 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 134:671–677
14. Von Kries R, Ihme N, Oberle D et al (2003) Effect of ultrasound screening on the rate of first operative procedures for developmental hip dysplasia in Germany. Lancet 362:1883–1887
15. Rosendahl K, Markestad T, Lie RT (1994) Ultrasound screening for developmental dysplasia of the hip in the neonate: the effect on treatment rate and prevalence of late cases. Pediatrics 94:47–52
16. Clegg J, Bache CE, Raut VV (1999) Financial justification for routine ultrasound screening of the neonatal hip. J Bone Joint Surg Br 81:852–857
17. Holen KH, Tegnander A, Bredland T et al (2002) Universal or selective screening of the neonatal hip using ultrasound? A prospective, randomised trial of 15,529 newborn infants. J Bone Joint Surg Br 84-B:886–890
18. Sewell MD, Rosendahl K, Eastwood DM (2009) Developmental dysplasia of the hip. BMJ 339:b4454
19. Rosendahl K, Aslaksen A, Lie RT et al (1995) Reliability of ultrasound in the early diagnosis of developmental dysplasia of the hip. Pediatr Radiol 25:219–224
20. Bar-On E, Meyer S, Harari G et al (1998) Ultrasonography of the hip in developmental hip dysplasia. J Bone Joint Surg Br 80:321–324
21. Omeroglu H, Bicimoglu A, Koparal S et al (2001) Assessment of variations in the measurement of hip ultrasonography by the Graf method in developmental dysplasia of the hip. J Pediatr Orthop B 10:89–95