Visual Grading Analysis of digital neonatal chest phantom x-ray images: impact of detector type, dose and image processing on image quality

Smet¹ MH, Breyssem¹ L, Mussen¹ E, Bosmans^{1,2} H, Marshall^{1,2} NW, Cockmartin¹ L

¹Department of Radiology, University Hospital Leuven,
Herestraat, 49

3000 – Leuven, Louvain, Belgium

² Medical Imaging Research Center,
Medical Physics and Quality Assessment,
Katholieke Universiteit Leuven,
3000 - Leuven, Louvain, Belgium

marleen.smet@uzleuven.be

luc.breysem@uzleuven.be

elvier.mussen@uzleuven.be

hilde.bosmans@uzleuven.be

nicholas.marshall@uzleuven.be

lesley.cockmartin@uzleuven.be

Corresponding author:

Maria Helena Smet
Department of Radiology, University Hospital Leuven
Herestraat, 49
3000 – Leuven, Louvain, Belgium
marleen.smet@uzleuven.be

Tel: +32 16 34 37 81

Fax: +32 16 34 37 65

Abstract

Objectives To evaluate the impact of digital detector, dose level and post-processing on neonatal chest phantom X-ray image quality (IQ).

Methods A neonatal phantom was imaged using four different detectors: a CR powder phosphor (PIP), a CR needle phosphor (NIP) and two wireless CsI DR detectors (DXD and DRX). Five different dose levels were studied for each detector and two post-processing algorithms evaluated for each vendor. Three paediatric radiologists scored the images using European quality criteria plus additional questions on vascular lines, noise and disease simulation. Visual grading characteristics and ordinal regression statistics were used to evaluate the effect of detector type, post-processing and dose on VGA score (VGAS).

Results No significant differences were found between the NIP, DXD and CRX detectors ($p > 0.05$) whereas the PIP detector had significantly lower VGAS ($p < 0.0001$). Processing did not influence VGAS ($p = 0.819$). Increasing dose resulted in significantly higher VGAS ($p < 0.0001$). Visual grading analysis (VGA) identified a detector air kerma/image (DAK/image) of $\sim 2.4 \mu\text{Gy}$ as an ideal working point for NIP, DXD and DRX detectors.

Conclusions VGAS tracked IQ differences between detectors and dose levels but not image post-processing changes. VGA showed a DAK/image value above which perceived IQ did not improve - potentially useful for commissioning.

Key points

- *A VGA study detects IQ differences between detectors and dose levels*
- *The NIP detector matched the VGAS of the CsI DR detectors*
- *VGA data are useful in setting initial detector air kerma level*
- *Differences in NNPS were consistent with changes in VGAS*

Keywords Newborn infant; Thoracic radiography; Image enhancement; Visual perception; Imaging phantoms

Abbreviations

AUC _{VGC}	Area under the VGC curves
CI	Confidence interval
DAK	Detector Air Kerma
DAP	Dose Area Product
DQE	Detector quantum efficiency
DRL	Dose Reference Levels
DRX	Needle crystal based wireless CsI DR Carestream [®] DRX 2530C
DXD	Needle crystal based wireless CsI DR Agfa [®] DX-D 35C
ESAK	Entrance Surface Air Kerma
FPDs	Flat panel detectors
ICC	Intra-class correlation coefficient
IQ	Image quality
M2	Musica 2 Agfa [®] post-processing
M3	Musica 3 Agfa [®] post-processing
MTF	Modulation transfer function
NICU	Neonatal intensive care unit
NIP	CR crystal based phosphor Image Plate Agfa [®] HD5.0
NNPS	Normalized noise power spectrum
OR	Odds ratio
PIP	CR phosphor Image Plate Agfa [®] MD4.0
Proc1	Carestream [®] post-processing 1
Proc2	Carestream [®] post-processing 2
ROC	Receiver Operating Characteristics
VGA	Visual Grading Analysis
VGAS	Visual Grading Analysis score
VGC	Visual grading characteristics

Introduction

Optimization in neonatal digital chest x-ray imaging remains a challenging task-to provide sufficient diagnostic image quality (IQ) at the lowest possible dose [1,2]. As detector type, dose and post-processing may have significant impact on clinical IQ, optimization requires an approach that can cope with all these parameters. In medical imaging different methods are used to evaluate IQ. Physical IQ can be measured quantitatively by detector characteristics such as modulation transfer function (MTF) and detective quantum efficiency (DQE) [3]. However, these objective measurements only test the imaging hardware, and a large part of the complex clinical imaging chain, such as patient anatomy, processing, display and observer perception is excluded. Clinical IQ investigations cover the whole imaging chain, from acquisition to observer perception. A gold standard method for objective quantitation of diagnostic performance is Receiver Operating Characteristics (ROC) analysis [4]. However, ROC analysis needs a large number of cases with and without pathological findings and knowledge of a ground truth and such studies can be complex and time consuming to set up [5].

A more practical means of studying clinical IQ was introduced by Bth and Mnsson: Visual Grading Analysis (VGA) with visual grading characteristics (VGC) statistics [6]. This method shares some similarities with ROC analysis, where observers state their confidence concerning the fulfillment of a criterion [7]. VGA assumes that the visibility of normal anatomy strongly correlates with the detectability of pathology [3,8,9]. VGC analysis with non-parametric statistics then enables quantification of visual qualitative verbal ratings, making them suitable for comparative analysis of different imaging techniques [10,11]. In the literature, VGA has gained acceptance as an established and valuable tool for studying observer performance, producing results similar to clinical diagnosis decision making by radiologists [3,7]. Clinical interpretation of physical measurements for neonatal applications remains difficult [12,13]. This is probably due to difficulties in making a clear definition of the IQ required to perform some clinical task, either quantitatively or qualitatively [12]. The clinical image optimization process is even more complex in children than in adults because of the variability in sizes and higher sensitivity to the stochastic effects of irradiation [1,14,15]. There is currently no objective means of predicting an optimal balance between dose and IQ for a specific diagnostic task and this prompted the current study: whether a subjective VGA study might be a helpful tool in finding trade-offs between IQ and dose. This work reports on a VGA method to study the impact of different digital detectors, dose levels and post-processing algorithms on neonatal chest phantom image quality.

Materials and methods

A prospective VGA study was performed using X-ray images generated with an anthropomorphic neonatal chest phantom (Gammex^R610 Neonatal Chest Phantom, Gammex^R, Middleton, USA) (Fig. 1a). Lung inserts mimicking hyaline membrane disease and pneumothorax and two additional vascular lines (Premicath^R28G) were included within the phantom (Fig. 1b).

Four different detectors for neonatal bed-side imaging were assessed: a CR powder phosphor Agfa^R MD4.0 (PIP), a CR needle crystal (CsBr) based phosphor Agfa^R HD5.0 (NIP), a needle crystal (CsI) based wireless DR Agfa^R DX-D 35C (DXD) and a needle crystal based wireless CsI DR Carestream^R DRX 2530C (DRX) (Table 1). Agfa^R mobile DX-D100 and Carestream^R DRX-Revolution mobile x-ray systems were used to generate the Agfa^R and Carestream^R images respectively. The images were acquired using a small focus (0,6mm), 3,1mm Al total filtration, 63kV tube voltage, 1m focus-detector distance and 300cm² field of view at the phantom exit plane. Although the European Guidelines [16] recommend additional filtration, no (spectral) pre-filter was used in this study. Guidelines in ICRP Publication 121 [17] do not recommend the use of pre-filters for neonates and very small infants because of the low tube potentials used. Furthermore, selectable pre-filters were not available on the mobile system routinely used in our neonatal department. Exposure settings were identical for all detectors, covering 5 dose levels: detector air kerma (DAK) was 0.8μGy, 1.4μGy, 2.4μGy, 3.7μGy and 8.3μGy, estimated using the calibrated exposure index [18]. Four different multiscale post-processing algorithms optimized for paediatric examinations were applied: Agfa^R Musica2 (M2), Agfa^R Musica3 (M3), Carestream^R1 (Proc1) and Carestream^R2 (Proc2). All X-ray images were blinded for the image acquisition parameters and randomly presented on a calibrated 2 mega pixel diagnostic medical display system (Barco^R-Nio2MP E2-621, Kortrijk, Belgium).

Clinical image quality evaluation was performed by three experienced paediatric radiologists (M.S., L.B. and E.M.) with respectively 32, 29 and 16 years of experience. The reproduction of important normal anatomical structures was scored following European quality criteria [6,16], with additional questions on the visibility of vascular lines, noise and disease simulation (Table 2). Criteria on peripheral airways, diaphragm, costophrenic angles and pneumothorax (numbers 3,4,5,10) were used to evaluate VGA scores (VGAS) in high contrast regions (lung fields) while criteria on trachea, bronchi, spine and catheters (numbers 1,2,6,7,8,9) were used to evaluate VGAS in low contrast areas (mediastinum). A

five-point Likert scale (Table 3) was used for scoring via the Viewdex^R software program [19,20]. The readers were allowed to adjust image brightness, contrast and magnification. No time and viewing distance limits were imposed. Before starting the study, a short training session was used to get the readers accustomed to the scoring methodology. Additionally, 50 images were read twice to evaluate scoring consistency for each reader.

VGC analysis was then applied following Bth and Mnsson [6]. First, the VGA score (VGAS) i.e. mean value of all ratings, was calculated for each dose level, each detector and all criteria together or for the criteria defining image quality in lung fields, mediastinum, catheters and noise separately. Secondly, for each detector type and processing algorithm the relative number of each confidence rating was tabulated. Next, two frequency distributions, one for each detector type, of the five-point Likert scales were calculated by accumulatively increasing the decision threshold that indicates if the quality criterion was fulfilled for each comparison of two detectors/processing. This results in the corresponding points of the VGC curve. Finally, the area under the VGC curve (AUC_{VGC}) was calculated using the trapezoidal rule; if the confidence interval of AUC_{VGC} includes 0.5, then the difference between the two detectors can be considered not statistically significantly different [6]. In order to analyze the effect of dose in our study, ordinal regression was performed to investigate the effect of dose and processing on VGA score while taking into account potentially confounding effects of independent variables such as detector type and reader [21]. The intraclass correlation coefficient (ICC) was applied to quantify the agreement between the image quality scores of the three observers as well as the agreement between repeated reader ratings. Statistical analysis was performed using the IBM Statistical Package for Social Sciences (SPSS version 13); statistical significance level was 0.05.

Results

Table 4 shows the area under the VGC curves (AUC_{VGC}) together with the 95% confidence interval (CI) for the detector comparisons and for the comparison between Agfa^R (M2, M3) and Carestream^R (Proc1, Proc2) image processing. No significant differences in image quality scores were found between the NIP, DXD and CRX detectors whereas scores for the PIP were significantly worse. Comparison between the two processing algorithms showed no significant difference for any of the detectors.

Results of ordinal regression analysis are reported in Table 5. The logit model was used, treating the median score of 11 image quality criteria and 3 readers as dependent variables, the categorical variables of 4 different detectors and 2 different processing

techniques for Agfa^R and Carestream^R images as independent factors, while the dose, a continuous variable, served as an independent covariate. Increasing dose levels resulted in significantly higher VGAS with an odds ratio (OR) of 22.92 (95% CI, 14.01 to 37.49), $p < .0001$).

Additionally, Table 5 shows that images processed with M2 resulted in similar image quality scores compared to images processed with M3, with an OR of 0.98 (95% CI, 0.60 to 1.61) and no significant effect ($p=0.93$). The same was found for Proc1 versus Proc2 (OR=1.06 (95% CI, 0.46 to 2.49, $p=0.89$).

Furthermore it was confirmed that only the PIP resulted in significantly different image quality scores, with an OR=0.03 (95% CI, 0.01 to 0.08, $p<0.0001$) compared to the DRX presented as a reference here.

Next, the VGAS, i.e. the mean ratings of all readers and all criteria for a given dose level, detector type and processing algorithm, were calculated. These scores increased with increasing dose (Fig. 2). The third dose tested (DAK~2.4 μ Gy), is the current clinically used dose level; below this dose, VGAS decreased while higher dose did not improve VGAS, except for the PIP detector.

In Figure 3, the VGAS is plotted against dose for image quality criteria 3,4,5,10 in the high contrast lung region (Table 2). The DRX gave better scores compared to the other detectors, even at lower dose levels. The PIP always had the lowest score. Data points for the different post-processing algorithms were overlapping, indicating no difference in image quality between M2 and M3, and between Proc1 and Proc2.

Fig. 4 shows the VGAS versus dose for image quality criteria 1,2,6,7,8,9 (Table 2) in the low contrast environment of the mediastinum. The DRX had the lowest scores at lower dose levels compared to the DXD and NIP detectors. Overall, the NIP had a good score, even at lower dose levels, while the PIP always gave the lowest scores. Again, no difference was found between the different post-processing algorithms. Similar trends in VGAS were found for the 2 simulated vascular catheters (Fig. 5).

Curves of VGAS versus dose for perceived noise are illustrated in Fig. 6a. The DRX had low scores at lower dose levels, while the PIP always had the lowest scores. The difference in VGAS for the post-processing between M2 and M3 was larger than for the other criteria, especially for the NIP and PIP detectors.

The intra-class correlation coefficients (ICC) showed excellent inter-reader variability (ICC=0.79 (95%CI 0.75-0.82), indicating consistent image quality perception among the three radiologists. Additionally, the intra-reader agreement was excellent within each reader for 50 repeated image ratings with ICC values of 0.93 (95%CI 0.92-0.94), 0.77 (95%CI 0.72-0.80) and 0.88 (95%CI 0.86-0.90) for the three readers respectively.

Discussion

Two significant findings in current VGA study were the poor VGAS of PIP and the influence of dose on VGAS.

The inferior VGAS of the PIP is consistent with other studies comparing PIP with the newer NIP and DR detectors. Smans et al [22] confirmed for neonatal chest imaging the superiority of the NIP over the PIP in a computer model image simulation technique at 65kV, 3.8mm Al total filtration and phantom entrance air kerma of respectively 22, 80 and 110 μ Gy. Using a VGC image validation study on feline chest radiographs, Ludewig et al [23] found the image quality of the NIP clearly superior to that for the PIP, allowing a 50% possible dose reduction without relevant loss of IQ. Cohen et al reported a possible dose reduction of ~20% with the NIP compared to the PIP, using an image preference study for neonatal chest radiographs [24]. Higher phosphor packing and less light spread in the needle phosphor detectors improve X-ray absorption and image sharpness [25]. We have quantified the necessary dose increase for the PIP to achieve the IQ of NIP or DR detectors. However, increasing dose is the least acceptable solution for improving IQ in neonatal imaging.

VGA scores between the NIP, DXD and DRX detectors were not significantly different. The equal performance of the NIP versus FPDs (flat panel detectors) has also been described in the literature [26,27], however, to the best of our knowledge, no study found IQ of NIP to be equal to FPD for neonatal chest. They are substantially cheaper than FPD devices but disadvantages are the cassette handling and associated reduction in workflow. FPDs offer instant images, sometimes wireless, increased workflow, equally high IQ - but at a much higher price than the NIP.

Another significant factor in our VGA study was the influence of dose.

In digital X-ray systems noise strongly impacts IQ with the signal to noise ratio (SNR) being the most significant indicator of IQ [28]. The efficiency of SNR transmission of a detector is expressed in its DQE (detector quantum efficiency) [28]. However, at the

lowest doses in our study, VGAS was lower for the DRX than for the NIP in the mediastinum (Fig. 4 and 5). This can be explained by the increased contribution from electronic noise at these low doses. The importance of electronic noise was described by Menser et al in a Monte Carlo pediatric chest simulation [29]. Physical IQ measurements by Marshall et al [30] also showed that electronic noise can reduce DQE of FPDs compared to NIP CR detectors at lower doses.

The influence of DAK on image noise can be investigated quantitatively using the normalized noise power spectrum (NNPS) (Fig. 6b). The NNPS curve at $\sim 2.5\mu\text{Gy}$ for the PIP is clearly higher, corresponding to a noisier image appearance. NNPS for the NIP, DXD and DRX at $\sim 2.5\mu\text{Gy}$ all lie at a similar level, with the exception of the DRX at spatial frequencies greater than 3mm^{-1} . This is consistent with the similar VGAS achieved at a dose of $2.4\mu\text{Gy}$ for these detectors, including the DRX. Increasing the DAK to $8\mu\text{Gy}$ brings the NNPS curve for the PIP to the same level as for the other detectors. The VGA data show that in order to match the VGAS of the NIP, DXD and DRX detectors, the PIP has to be operated at higher dose i.e. $\sim 8\mu\text{Gy}$, the same dose level as found for the NNPS.

This VGA study showed (Fig. 2) that a dose of $2.4\mu\text{Gy}$ DAK is a good trade-off where a further increase of dose did not improve VGAS, except for the PIP, while a reduction in dose reduced VGAS, giving “poor” and “very poor” IQ scores (VGAS 2 and 1). A DAK of $2.4\mu\text{Gy}$ was the set working point for these detectors at the beginning of the study for the detectors used in house (NIP, DXD and DRX). PIP detectors are no longer in use, but had been used previously and at higher dose. It is remarkable that radiologists, blinded for the dose, found exactly this clinically used $2.4\mu\text{Gy}$ DAK to be acceptable for NIP, DXD and DRX. As a consequence of this study, we have not changed our clinically working point; it is possible that the DAK of $2.4\mu\text{Gy}/\text{image}$ can serve as a guideline for other clinical sites. This is consistent with German guidance proposing detector classes compatible with $\leq 2.5\mu\text{Gy}$ in neonates [31]. ESAK (entrance surface air kerma) levels for neonatal chest X-rays proposed by the EC [16] and NRPB [32] are respectively $80\mu\text{Gy}$ and $50\mu\text{Gy}$. The trade-off point of $2.4\mu\text{Gy}$ DAK in our study resulted in a measured ESAK of $20\mu\text{Gy}$. This is lower than the recommended ESAK of EC and NRCB and also lower than proposed ESAK levels in Belgium ($45\mu\text{Gy}$) [33], Finland ($34\text{--}66\mu\text{Gy}$) [34] and Austria ($50\mu\text{Gy}$) [35]. Martin et al [11], using VGAS to validate IQ and progressive dose reduction in NICU chest X-rays on CR, found acceptable ESAK of 20 to $30\mu\text{Gy}$, which is similar to our findings. We are reporting ESAK values, rather than DAP values as this is a more relevant comparison, given the wide range in patient sizes and variable collimations.

The important role of post-processing in optimizing IQ in digital imaging is described in the literature. For chest radiography, especially, image post-processing differs for paediatric and adult patients [36]. Although differences in VGAS of noise perception were not significant, there were some trends: differences in VGAS between the post-processing algorithms tested as a function of dose were larger between M2 and the newer M3 for the PIP than the NIP (Fig. 6a). M2 scored higher than the newer M3 post-processing. A possible explanation is that the observers were accustomed to the look of M2 than M3 for a PIP detector. It is possible that the new algorithm was tuned more specifically for the NIP, giving poorer performance for the PIP. While M3 showed the noise more prominently than M2, it maintained low and high contrast visualization (Fig. 3, Fig. 4), and sharp representation of the catheters (Fig. 5). The result that no significant differences were found between the different post-processing algorithms might be explained by the use of an *absolute* VGA study in which the scoring of subtle image quality differences might be more difficult or less sensitive than the side-by-side image comparisons in a *relative* VGA study. Alves et al [37] remark that sometimes VGA is not sufficient to make distinctions between different techniques. Using a different validation method of rating and matching, Sensakovic et al [38] were able to demonstrate significant differences in perceived IQ for different image post-processing sets using the same Gammex^R neonatal chest phantom.

There are some shortcomings in phantom usage such as less variability of anatomy and pathology, no motion artifacts and absence of variable positioning and collimation seen in clinical practice. Therefore, conclusions from phantom studies certainly need further correlation with studies involving patients.

Although a VGA study with VGC statistics is an established method [3, 8-11] a potential limitation is its subjective character being prone to bias. To reduce this risk, well-defined clinically relevant anatomical standards were used for scoring [3]. Another disadvantage of visual scoring is its dependence on the experience of the observers [3,8]. In our study, all observers were experienced and showed an excellent inter-reader variability.

In conclusion, the VGA method found clear IQ differences between the detectors and dose levels but not between the different post-processing algorithms. VGA identified a trade-off detector air kerma level of 2.4 μ Gy, above which IQ score did not improve for the NIP and FPDs, while lower dose levels reduced IQ score. This 2.4 μ Gy DAK trade-off point may be helpful as a clinical working point for optimization protocols in neonatal

imaging techniques. While this VGA study has given clear insight into parameters that influence clinical image quality for a neonate anthropomorphic phantom, further studies on real patient images are needed before results can be extrapolated and considered finally validated.

Acknowledgements:

Dirk Vandenbroucke for commenting on the methods, making the phantom available and his technical interpretation of the results, Thomas De Clercq for assisting in this research project.

References

1. Kim TH, Ryu JH, CW Jeong et al (2017) Reduced radiation dose and improved image quality using a mini mobile digital imaging system in a neonatal intensive care unit. *Clin Imaging* 42:165–171
2. Conradie A, Herbst CP (2016) Evaluating the effect of reduced entrance surface dose on neonatal chest imaging using subjective image quality evaluation. *Phys Med* 32:1368–1374
3. Båth M (2010) [Evaluating imaging systems: practical applications](#). *Radiat Prot Dosimetry* 139:26-36
4. Lança L, Silva A (2009) Digital radiography detectors – A technical overview: Part 2. *Radiography* 15:134-138
5. Tapiovaara MJ (2008) Review of relationships between physical measurements and user evaluation of image quality. *Radiat Prot Dosimetry* 129:244-248
6. Båth M, Månsson LG (2007) Visual grading characteristics (VGC) analysis: a non-parametric rank-invariant statistical method for image quality evaluation. *Br J Radiol* 80:169-176
7. [Zarb F](#), [McEntee ME](#), [Rainford L](#) (2015) Visual grading characteristics and ordinal regression analysis during optimisation of CT head examinations. [Insights Imaging](#) 6:393-401
8. Sund P, Båth M, Kheddache S, Månsson LG (2004) Comparison of visual grading analysis and determination of detective quantum efficiency for evaluating system performance in digital chest radiography. *Eur Radiol* 14:48-58
9. [Tesselaar E](#), [Dahlström N](#), [Sandborg M](#) (2016) Clinical audit of image quality in radiology using visual grading characteristics analysis. *Radiat Prot Dosimetry* 169:340-346
10. [Sandborg M](#), [Tingberg A](#), [Ullman G](#), Dance DR, Alm Carlsson G (2006) Comparison of clinical and physical measures of image quality in chest and pelvis computed radiography at different tube voltages. *Med Phys* 33:4169-4175
11. Martin L, Ruddlesden R, Makepeace C, Robinson L, Mistry T, Starritt H (2013) Paediatric x-ray radiation dose reduction and image quality analysis. *J Radiol Prot* 33:621–633
12. Shet N, Chen J, Siegel EL (2011) Continuing challenges in defining image quality. *Pediatr Radiol* 41:582-587
13. Don S (2011) Pediatric digital radiography summit overview: state of confusion. *Pediatr Radiol* 41:567-572
14. Armpilia CI, Fife IA, Croasdale PL (2002) Radiation dose quantities and risk in neonates in a special care baby unit. *Br J Radiol* 75:590–595
15. Berrington de González A, Darby S (2004) Risk of cancer from diagnostic X-rays: estimates for the UK and 14 other countries. *Lancet* 363:345-351
16. European Commission (1996) European guidelines on quality criteria for diagnostic radiographic images in paediatrics, Report EUR 16261EN
17. ICRP (2013) Radiological protection in paediatric diagnostic and interventional radiology. ICRP Publication 121. Ann. ICRP 42(2)
18. Shepard SJ et al (2009) An exposure indicator for digital radiography: AAPM Task Group 116 (Executive Summary) *Med Phys* 36:2898-2914
19. [Börjesson S](#), [Håkansson M](#), [Båth M](#) et al (2005) A software tool for increased efficiency in observer performance studies in radiology. *Radiat Prot Dosimetry* 114:45-52
20. [Håkansson M](#), [Svensson S](#), [Zachrisson S](#), Svålvik A, Båth M, Månsson LG (2010) Viewdex: an efficient and easy-to-use software for observer performance studies. *Radiat Prot Dosimetry* 139:42-51
21. Smedby O, [Fredrikson M](#), [De Geer J](#), Borgen L, Sandborg M (2013) Quantifying the potential for dose reduction with visual grading regression. *Br J Radiol*. DOI:10.1259/bjr/31197714
22. [Smans K](#), [Vandenbroucke D](#), [Pauwels H](#), [Struelens L](#), [Vanhavere F](#), [Bosmans H](#) (2010)

- Validation of an image simulation technique for two computed radiography systems: an application to neonatal imaging. *Med Phys* 37:2092-2100
23. [Ludewig E](#), [Hirsch W](#), [Bosch B et al](#) (2010) Assessment of clinical image quality in feline chest radiography with a needle-image plate (NIP) storage phosphor system-an approach to the evaluation of image quality in neonatal radiography. [Rofo](#) 182:122-132
 24. Cohen M, Corea D, Wanner M et al (2011) Evaluation of a new phosphor plate technology for neonatal portable chest radiographs. *Acad Radiol* 18:197-198
 25. [Schaefer-Prokop C](#), [Neitzel U](#), [Venema HW](#), Uffmann M, Prokop M (2008) Digital chest radiography: an update on modern technology, dose containment and control of image quality. [Eur Radiol](#) 18:1818-1830
 26. [Bremicker K](#), [Gosch D](#), [Kahn T](#), [Borthé G](#) (2015) The future of bedside chest radiography: Comparative study of mobile flat-panels and needle-image plate storage phosphor systems. [Med Klin Intensivmed Notfmed](#) 110:603-608
 27. Fernandez JM, Ordiales JM, Guibelalde E, Prieto C, Vano E (2008) Physical image quality comparison of four types of digital detector for chest radiology. [Radiat Prot Dosimetry](#) 129:140-143
 28. Spahn M (2005) Flat detectors and their clinical applications. *Eur Radiol* 15:1934-1947
 29. [Menser B](#), [Manke D](#), [Mentrup D](#), [Neitzel U](#) (2016) A Monte Carlo simulation framework for joint optimisation of image quality and patient dose in digital paediatric radiography. [Radiat Prot Dosimetry](#) 169:371-377
 30. Marshall NW, Smet MH, Hofmans M, Pauwels H, De Clerq T, Bosmans H (2017) Technical characterization of five X-ray detectors for paediatric radiography applications. *Phys Med Biol*. DOI:10.1088/1361-6560/aa9599
 31. Bundesärztekammer (2007) Leitlinie der Bundesärztekammer zur Qualitätssicherung in der Röntgendiagnostik.
 32. Hart D, Wall BF, Shrimpton PC, Bungay DR, Dance DR (2000) NRPB- R318-Reference doses and patient size in paediatric radiology. NRPB-R318, NRPB, Chilton
 33. Dabin J, Struelens L, Vanhavere F (2014) Radiation dose to premature new-borns in the Belgian neonatal intensive care units. *Radiat Prot Dosimetry* 158:28-35
 34. Kiljunen T, Järvinen H, Savolainen S (2007) Diagnostic reference levels for thorax X-ray examinations of paediatric patients. *Br J Radiol* 80:452-459
 35. Billinger J, Nowotny R, Homolka P (2010) Diagnostic reference levels in pediatric radiology in Austria. *Eur Radiol* 20:1572-1579
 36. [Don S](#), [MacDougall R](#), [Strauss K](#) et al (2013) Image gently campaign back to basics initiative: ten steps to help manage radiation dose in pediatric digital radiography. [AJR Am J Roentgenol](#) 200:W431-W436
 37. [Alves AF](#), [Alvarez M](#), [Ribeiro SM](#), Duarte SB, Miranda JR, Pina DR (2016) Association between subjective evaluation and physical parameters for radiographic images optimization. [Phys Med](#) 32:123-132
 38. Sensakovic WF, O'Dell MC, Letter H et al (2016) Image quality and dose differences caused by vendor-specific image processing of neonatal radiographs. [Pediatr Radiol](#) 46:1606-1613

Table and Figure legends:

Table 1. Technical specifications of the detectors

Detector name	Technology	X-ray conversion layer	Pixel spacing (mm)	Pixel matrix
Agfa MD4.0	powder CR	BaSrFBr:Eu ²⁺	0.100	1728 X 2328
Agfa HD5.0	needle CR	CsBr:Eu ²⁺	0.100	1656 X 2280
Agfa DX-D 35C	needle DR	CsI	0.125	2192 X 2800
Carestream DRX 2530C	needle DR	CsI	0.139	2176 X 1792

Table 2. Image quality criteria

Criterion
1.Reproduction of the trachea
2.Reproduction of the proximal bronchi
3.Reproduction of the small peripheral airways
4.Visually sharp reproduction of the diaphragm
5.Visually sharp reproduction of the costophrenic angles
6.Reproduction of the spine
7.Visibility of the upper catheter in the mediastinum
8.Visibility of the tip of the upper catheter
9.Visibility of the tip of the lower catheter
10.Reproduction of the pneumothorax
11.Diagnostic acceptable image noise

Table 3. Absolute rating scale

Grade	Decision
A	Criterion definitely not fulfilled
B	Criterion probably not fulfilled
C	Indecisive whether criterion fulfilled or not
D	Criterion probably fulfilled
E	Criterion definitely fulfilled

Table 4. Area under the visual grading characteristic curve (AUC_{VGC}) with 95% confidence interval for all paired-wise comparisons between the 4 different detectors and two processing techniques for each detector. This AUC_{VGC} gives a measure of the difference in image quality between two compared detectors or processing techniques; if the confidence interval of AUC_{VGC}

includes 0.5, then the difference between the two detectors can be considered not statistically significantly different.

AUC _{VGC} [95%CI]	DXD M2	NIP M2	PIP M2	DXD M3	NIP M3	PIP M3	DRX proc 1	DRX proc 2
DXD M2	NA	0.51 [0.48-0.55]	0.71 [0.67-0.74]	0.50 [0.46-0.53]	0.51 [0.48-0.54]	0.71 [0.68-0.74]	0.48 [0.44-0.51]	0.49 [0.45-0.52]
NIP M2	0.49 [0.45-0.52]	NA	0.70 [0.67-0.73]	0.50 [0.46-0.53]	0.50 [0.46-0.53]	0.71 [0.68-0.74]	0.46 [0.42-0.49]	0.47 [0.44-0.50]
PIP M2	0.29 [0.26-0.33]	0.30 [0.27-0.33]	NA	0.29 [0.26-0.32]	0.30 [0.27-0.33]	0.50 [0.47-0.54]	0.27 [0.24-0.30]	0.29 [0.26-0.32]
DXD M3	0.50 [0.47-0.54]	0.52 [0.48-0.55]	0.71 [0.68-0.74]	NA	0.51 [0.48-0.55]	0.71 [0.68-0.74]	0.48 [0.45-0.52]	0.49 [0.46-0.53]
NIP M3	0.49 [0.46-0.52]	0.50 [0.47-0.54]	0.70 [0.67-0.74]	0.49 [0.45-0.52]	NA	0.71 [0.68-0.74]	0.46 [0.43-0.50]	0.47 [0.44-0.51]
PIP M3	0.29 [0.26-0.32]	0.29 [0.26-0.33]	0.50 [0.46-0.53]	0.29 [0.25-0.32]	0.29 [0.26-0.32]	NA	0.28 [0.25-0.31]	0.28 [0.25-0.31]
DRX proc 1	0.52 [0.49-0.56]	0.54 [0.51-0.58]	0.73 [0.70-0.76]	0.52 [0.48-0.55]	0.54 [0.50-0.57]	0.74 [0.71-0.77]	NA	0.51 [0.47-0.54]
DRX proc 2	0.51 [0.48-0.55]	0.53 [0.50-0.56]	0.71 [0.68-0.75]	0.51 [0.47-0.54]	0.53 [0.49-0.56]	0.72 [0.68-0.75]	0.49 [0.46-0.53]	NA

NA = not applicable **red = 95%CI of AUC_{VGC} contains 0.5 = no significant difference**

Table 5. Results of ordinal logistic regression analysis

Model function: Logit	-2 Log Likelihood	Chi-square	df	p-value				
Intercept only	923.016							
Final	484.694	438.322	6	< 0.0001				
Parameter Estimates	Estimate	Std. Error	Wald	df	p-value	OR	95% Confidence Interval	
							Lower Bound	Upper Bound
[ratings = 1]	-1.31	0.40	10.98	1	0.001	0.27	0.12	0.59
[ratings = 2]	0.45	0.36	1.54	1	0.215	1.56	0.77	3.17
[ratings = 3]	1.23	0.36	11.75	1	0.001	3.42	1.69	6.90
[ratings = 4]	4.17	0.43	93.89	1	< 0.0001	64.44	27.75	149.67
[ratings = 5]	0	.	.	0	.			
dose	3.13	0.25	155.56	1	< 0.0001	22.92	14.01	37.49
[detector=DXD]	-0.02	0.40	0.00	1	0.963	0.98	0.45	2.14
[detector=NIP]	-0.08	0.40	0.04	1	0.848	0.93	0.43	2.02
[detector=PIP]	-3.39	0.44	60.48	1	< 0.0001	0.03	0.01	0.08
[detector=DRX]	0.00	.	.	0	.			
[processing=M2]	-0.02	0.25	0.01	1	0.930	0.98	0.60	1.61
[processing=M3]	0.00	.	.	0	.			
[processing=Proc1]	0.06	0.43	0.02	1	0.885	1.06	0.46	2.49
[processing=Proc 2]	0	.	.	0	.			

Fig. 1a The Gammex[®] neonatal chest phantom (Image courtesy of Gammex[®] Inc, USA). **b** X-ray image of the Gammex[®] neonatal chest phantom simulating hyaline membrane disease and pneumothorax: left subclavian vein line simulation (black arrowhead), tip of the simulated subclavian catheter (white arrow), tip of the simulated lower peripheral catheter (black arrow), pneumothorax (white arrowhead).

Fig. 2 VGAS significantly ($p < 0.0001$) increased with higher dose levels. Dose level of 2,4 μGy detector air kerma was our currently clinically used dose level. Dose levels below this level decreased VGAS scores while higher dose levels did not improve VGAS scoring, except for the PIP detector.

Fig. 3 VGAS versus dose for structures in a high contrast environment (lung fields).

Fig. 4 VGAS versus dose for structures in a low contrast environment (mediastinum).

Fig. 5 VGAS versus dose for the 2 simulated vascular catheters.

Fig. 6a VGAS versus dose level for the criterion perceived diagnostic acceptable noise. **b** Normalized Noise Power Spectrum (NNPS) measured at RQA3 beam quality for the four detectors, at a DAK level of $\sim 2.5 \mu\text{Gy}/\text{image}$. Also shown are NNPS curves for the PIP and NIP detectors at a DAK of $\sim 8 \mu\text{Gy}/\text{image}$.