Evaluation of lumbar spine images with added pathology

Anders Tingberg*[1], Clemens Herrmann[2], Jack Besjakov[3], Karsten Rodenacker[2], Anja Almén[1], Patrik Sund[4], Sören Mattsson[1], Lars Gunnar Månsson[4]

[1]Department of Radiation Physics at Malmö, Lund University, Malmö University Hospital, SE-205 02 Malmö, Sweden
[2]GSF - National Research Center for Environment and Health, D-85764 Neuherberg, Germany
[3]Department of Diagnostic Radiology at Malmö, Lund University, Malmö University Hospital, SE-205 02 Malmö, Sweden
[4]Department of Radiation Physics, Göteborg University, Sahlgrenska University Hospital, SE-413 45 Göteborg, Sweden

ABSTRACT

Optimisation of radiographic procedures require solid tools for evaluation of the image quality in order to ensure that it is sufficient to answer the clinical question at the lowest possible absorbed dose to the patient. Lumbar spine radiography is an examination giving a relatively high dose and good methods for evaluation of image quality as well as dose are needed. We have developed and used a method for the addition of artificial pathological structures into clinical images. The new images were evaluated in a study of detectability (free-response forced error experiment). The results from the study showed that the methodology can be used to detect differences in the screen-film systems used to produce the images, indicating that the method can be used in a study of image quality. The results of the study of detectability were compared with the outcome of a visual grading analysis based on the structures mentioned in the European Quality Criteria. The comparison indicated that a linear correlation exists between the two methods. This means that the simpler VGA can be used in the evaluation of clinical image quality.

Keywords: Diagnostic radiology, lumbar spine hybrid images, added pathology, free-response forced error (FFE) experiment, visual grading analysis (VGA), European Quality Criteria, evaluation methods for clinical image quality

1. INTRODUCTION

Determination of image quality using various types of physical parameters, like signal-to-noise ratio or detective quantum efficiency, often give valuable information about the performance of an imaging system. However, to investigate the complete chain from the patient to the observer, observer performance studies must be employed. Receiver operating characteristics (ROC)[1] in its various forms offer the possibility to investigate an imaging system in a way which considers both the sensitivity and the specificity of the system. However, the truth must be known and if patient images with (or without) real pathological structures are used, the "true" state (positive or negative) of the image has to be settled by a panel of experienced radiologists or by autopsy.

To simplify the set-up of an ROC study, clinical radiographs with added known structures (artificial or real) that resemble tumours can be used. This has been done especially in chest radiography. Samei et al.[2] manufactured Teflon objects with gaussian shaped thickness profiles to be placed on the surface of the patient. Sherrier et al.[3] used digitised chest images to extract tumours that they then used in a detection study. For a study to be successful the artificial structure must look like a tumour would have done if it was present in the patient at the time of the exposure, including effects of attenuation. Lumbar spine radiography is an examination with a relatively high absorbed dose to the patient. Objective methods for evaluation of the quality are needed in optimising this examination with respect to sufficient image quality for correct diagnosis and as low absorbed dose to the patient as possible.

* Correspondence: E-mail: anders.tingberg@rfa.mas.lu.se; Telephone: +46 40 33 11 55; Fax: +46 40 96 31 85

We have earlier used a variant of ROC, free response forced error (FFE) experiment described by Chakraborty and Winter[4]. The task for the observer is to correctly localise multiple lesions in an image and to rank them in order of confidence. The set-up of an ROC study is however difficult and time consuming. It is therefore desirable to find other methods that are both precise and objective to evaluate clinical image quality. In visual grading analysis (VGA) the task for the observer is to compare the quality of a whole image or parts of an image with a particular reference image. Any set of clinical images can be used in a VGA study. We have previously performed VGA studies with good results[5], using the structures mentioned in the European Quality Criteria[6]. It would therefore be desirable to find a correlation between the results of the detection study and the much simpler visual grading analysis.

The aims of this study were to create a digital model for producing lumbar spine images with added pathology ("hybrid images"), to use the images in a study of detectability and to compare the results of this study with the results of a visual grading analysis (VGA) based on the structures mentioned in European Quality Criteria.

2. MATERIAL AND METHODS

1. Digital data acquisition and display

Fifteen conventional clinical frontal lumbar spine radiographs (70kV tube voltage, Kodak Lanex Regular screen, Kodak T-Mat L Film, grid) were used together with detailed information about the exposure parameters, the screen-film system and the patient configuration[5]. The H/D curve as well as the MTF and Wiener spectrum of the screen-film system was measured. The radiographs were digitised with a high resolution film scanner (ULTRASCAN 5000, Vexcel Imaging, Graz, Austria), which was calibrated in terms of optical density. The spatial resolution was 40 μm, and the dynamic resolution was 16 bit linear to transmission. For image display, a high resolution laser imager (Scopix LR5200, Agfa-Gevaert, Munich, Germany) was available, with a nominal spatial resolution of 40 μm and a dynamic resolution of eight bit. The knowledge of both scanner and printer calibration was an essential prerequisite, and the quality of the digital data was sufficient for the simulations and manipulations described in the following paragraphs. Details will be reported elsewhere[7].

2. Simulation of different image quality levels

The images reported here were processed with a computer workstation (O2, Silicon Graphics Inc.) with one GB main memory and about 20 GB hard disk capacity. For the image manipulations, functions and procedures were developed using the software package IDL™. With respect to the large amount of data, the image files were treated in parts - if necessary overlapping.

The spatial resolution - measured in terms of modulation transfer function (MTF) -, and the noise of screen-film systems (Wiener or noise power spectrum) are important image quality parameters which are closely linked to the dose: in general, screen-film systems which require a low dose, show a decreased MTF and increased noise compared to systems requiring a high dose. Information about these image quality parameters of the original screen-film systems was supplied by the Institute of Applied Radiation Physics of the University Hospital of Lausanne, following as close as possible the exposure conditions employed for the original radiographs.

The aim of the image manipulations was to produce, for each patient, three images which were different with respect to the physical image quality level: Level A corresponded to the quality of the original film radiographs, and for levels B and C the noise was increased and the spatial resolution was decreased in two steps. Level B corresponded to the properties of a screen-film system of nominal speed class 600 to 800, and level C was even worse and expected to be hardly acceptable for radiologists.

For the manipulation of the spatial resolution, the scanned digital image data was filtered by means of an appropriate digital filter in the spatial frequency domain. The image noise was increased by digitally adding a noise distribution with zero mean and a certain spectral characteristic: by means of a random number generator, arrays with spectrally white density variations and zero mean were calculated, digitally filtered with the corresponding MTF of the screen-film system, and finally added to the digitised radiographs (compare Buhr et al., 1993[8]). The density dependence of the noise was taken into account by using corresponding noise distributions according to different average density levels into which the radiographs had been segmented before. This procedure was performed without contouring effects. Details will be reported elsewhere[9]. Totally three image sets, according to the three different combinations of noise and spatial resolution, with 15 images each, were produced and evaluated in a visual grading analysis study.

3. Simulation of pathological structures

In the lumbar spine both "positive" (denser bone - sclerotic lesions) and "negative" lesions (destruction of bone - osteolytic lesions) can be present. Fixing objects on the patient's body can simulate the positive lesions during the exposure, but negative lesions are very hard to mimic. With a digital lesion, however, this can be accomplished. To simulate a certain type of pathology in a computer, an approach was adopted which was originally proposed by Samei et al., 1997[2] for chest nodules. Samei had demonstrated that subtle lung nodules can be represented by circular Teflon phantoms with a Gaussian thickness profile. Our computer model simulates the effect of such an object, and the contrast and the lateral size of the object can be adjusted independently. To simulate the limited spatial resolution of a radiographic image, the model includes digital filtering with the modulation transfer function (MTF) of a conventional screen-film system. Such a nodule is finally represented by a two dimensional array of logarithmic exposure values, which enables the nodule contrast in terms of optical density to be automatically adjusted according to the H/D curve: e.g., added to an anatomical background with an optical density of about 0.5, the nodule contrast is about 0.06 OD, and at a background of D=1.8 the same object produces a contrast of more than 0.14 OD, caused by the increase of the film gradient.

This procedure results in a patient image with a nodule which appears as a blurred circular spot with decreased optical density due to the simulated increased X-ray absorption (sclerotic type of nodule). The computer model offers the chance not only to add but also to subtract the nodule. The subtraction corresponds to a type of disease which destroys the bone material in the spine and decreases the X-ray absorption (osteolytic or destructive type of nodule). Consequently, the nodule appears as a blurred circular spot with increased optical density on the film. It must be emphasised that such a lesion cannot be simulated by fixing objects externally to a patient. A destructive lesion involving cortex in a vertebral body destructs this cortical line and to simulate this behaviour image processing was performed to erase the cortex line. Details will be published elsewhere[10].

4. Model for hybrid image production

From the 15 original lumbar spine radiographs, the image of the patient that was closest to the "average patient" regarding age, weight and height, was selected and used as a basis for the production of the images for the FFE experiment. The image was divided into two regions. The lumbar spine and the sacrum defined region 1 while the winged portion of the iliac bone defined region 2. A digital grid structure was created for positioning the lesions in the two regions. The lesions were only positioned in the bony parts of the image. The diameter of the lesions was 10 mm in region 1 and 6 mm in region 2. One hundred and forty-nine positions were available in region 1 and 101 positions in region 2. The positions of the lesions were randomly generated and between 1 and 10 lesions were positioned in each region of each image. The type of lesion was also randomly chosen with 70% probability that the lesion would be a destruction and 30% probability that the lesion would be a sclerosis. Fifty different lesion distributions were produced with a random number generator. The same three combinations of noise and spatial resolution as above were applied to the hybrid images and thus 150 different images were created for the FFE-study.

5. Image evaluation

The FFE-images were randomly batched 10 by 10, and the VGA-images originating from the same patient were batched, to facilitate the viewing process and were evaluated individually by seven European expert radiologists on conventional viewing boxes. All identification tags on the films were removed and the films were assigned a randomly generated code. The illumination in the room was dim and kept constant. There was no limitation with regard to viewing time or viewing distance.

The hybrid images were evaluated according to the FFE procedure: The most apparent lesion in an image is marked with "1", the second most apparent with "2" and so on, until an error was made and thus a false positive image has been created. For each combination of noise and spatial resolution, the FFE-score, A_1, was calculated with equation 1[4]:

$$A_1 = \frac{\sum_{i=1}^{I}\sum_{o=1}^{O}(TPF_{o,i})}{I \times O} \qquad \text{Equation 1}$$

where

$(TPF_{o,i})$ = m_i/n_i, the quotient between the number of correct findings (m) before observer o makes an error and the total number of lesions (n) in image i.

I = Number of images (fifty for each combination of noise and spatial resolution in this study).

O = Number of observers (seven in this study).
The standard errors given represent the variation of the fraction of correct findings for the images for one combination of noise and spatial resolution.

The two types of lesions in the hybrid images could either be treated together (as described above) or separately. When the "normal" FFE scores have been calculated based on the correctly located lesions of both types the findings could be separated in correctly located destructive lesions and correctly located sclerotic lesions before the first error was made. With this division, an investigation can be made whether there is a difference in detectability between destructive and sclerotic lesions for the same attenuation differences compared to the surrounding tissue. In other words: is a positive lesion more visible than a negative? Each image will thus provide six measurement points (two regions; all lesions plus destructive and sclerotic lesions).

The radiologists evaluated the 45 images without added pathological structures with VGA. The image quality of seven structures mentioned in the European Quality Criteria (Table 1)[6] was visually compared to the same structures in a reference image, and graded on a five-level scale (Table 2).

Table 1. The structures used for the visual grading analysis. These structures are mentioned in the European Quality Criteria[6].

1. Upper and lower-plate surfaces
2. Pedicles
3. Lateral cortex
4. Intervertebral joints
5. Spinous processes
6. Transverse processes
7. Adjacent soft tissue

Table 2. The scoring system for the visual grading analysis. The image to be graded and the reference image always originated from the same patient.

Grading	Visibility of structure
- 2	Clearly inferior to
- 1	Slightly inferior to
0	Equal to
+ 1	Slightly better than
+ 2	Clearly better than the reference image

The VGA-images were individually masked with black cardboard so that a rectangular window only showed the relevant parts for this study[11]. The masking forced the observers to view exactly the same area of the images. The top of the rectangular window was located at the centre of L2 and the bottom was located at the centre of L4. In the lateral direction the window extended 2 cm outside the edges of the transverse processes of L3 (Figure 1).

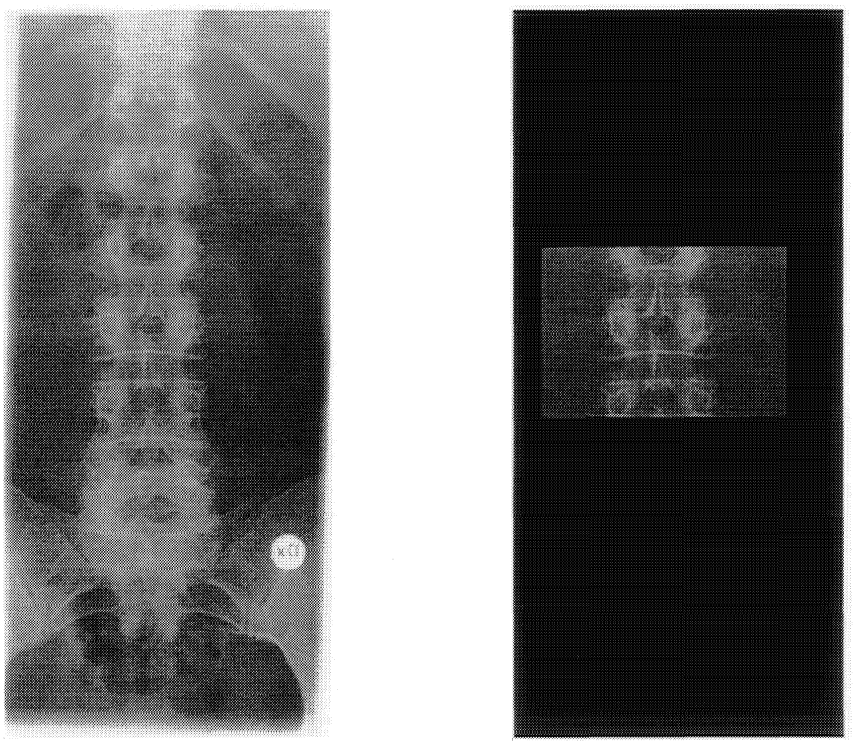

Figure 1. The images for the VGA study were individually masked and only an area around the third vertebral body was visible to the observers.

The original images (no additional noise or reduced spatial resolution) were used as reference images and were always from the same patient as the image being evaluated, and the three images of one patient were evaluated one after the other, but in random order. For each combination of noise and spatial resolution, a visual grading analysis score (VGAS) was calculated with the following equation:

$$VGAS = \frac{\sum_{i=1}^{I}\sum_{s=1}^{S}\sum_{o=1}^{O} G_{i,s,o}}{I \times S \times O} \qquad \text{Equation 2}$$

where

$G_{i,s,o}$ = Grading (-2, -1, 0, +1 or +2) for image i, structure s and observer o.
I = Number of images (fifteen for each combination of noise and spatial resolution in this study)
S = Number of structures (seven in this study)
O = Number of observers (seven in this study)

The standard errors given represent the variation of the scores for the images for one combination of noise and spatial resolution.

A comparison was made between the results of the VGA study and the FFE experiment. The structures used for the VGA are only present in region 1 (the lumbar spine and sacrum) (see Figure 1 and Table 1). Therefore the comparison between FFE and VGA is only performed for the results from region 1. It should be noted that for each combination of noise and spatial resolution there were 50 images for the FFE experiment and only 15 images for the VGA study.

3. RESULTS

The detectability (FFE-scores) for combinations of destructive and sclerotic lesions in region 1 and 2 is plotted in Figure 2, and the detectability of the lesions separated by type is plotted in Figure 3. The scores for the lowest noise and highest spatial resolution (the original images) were always the highest and the scores for the highest noise and lowest spatial resolution were always the lowest. There was no difference in detectability for the destructive and the sclerotic lesions in region 1, but in region 2 there was a significant difference for the worst combination of noise and spatial resolution. The detectability of the lesions in region 2 (the winged portion of the iliac bone) was higher than for the lesions in region 1 (the lumbar spine and the sacrum) for both types of lesion.

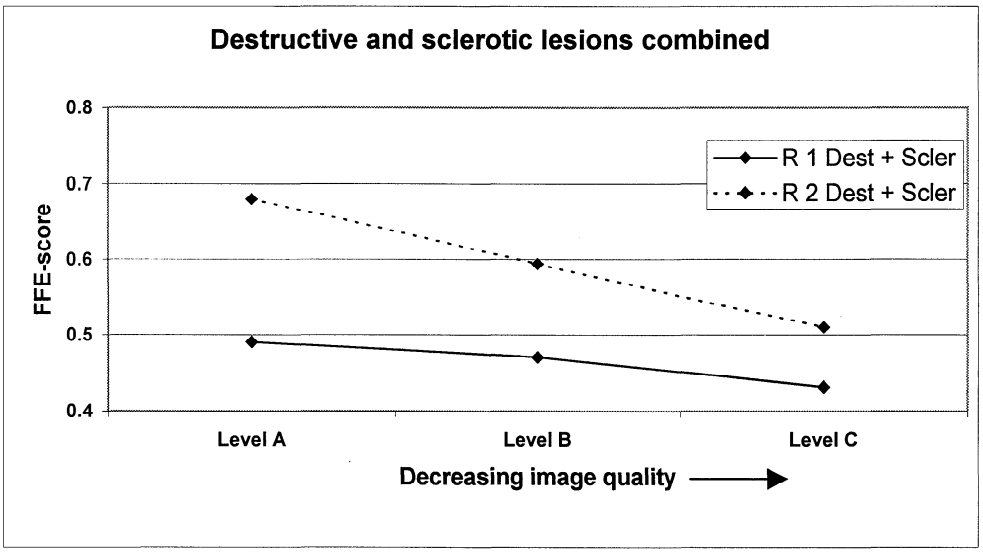

Figure 2. FFE-scores for combinations of destructive and sclerotic lesions for the three combinations of noise and spatial resolution (R1 = region 1; R2 = region 2).

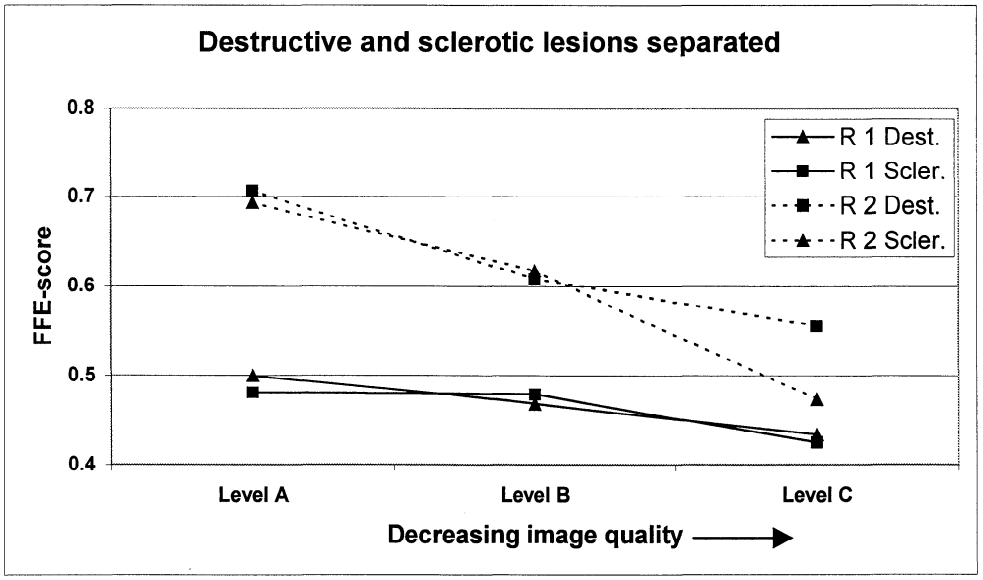

Figure 3. FFE-scores for the destructive and sclerotic lesions separated for the three combinations of noise and spatial resolution (R1 = region 1; R2 = region 2).

In Figure 4 the results from the VGA study is plotted against the results from the FFE experiment of both types of lesions in region 1. The results indicate a linear correlation between the objective evaluation method (FFE) and the subjective method (VGA). The standard errors for the VGAS were much lower than for the FFE-scores.

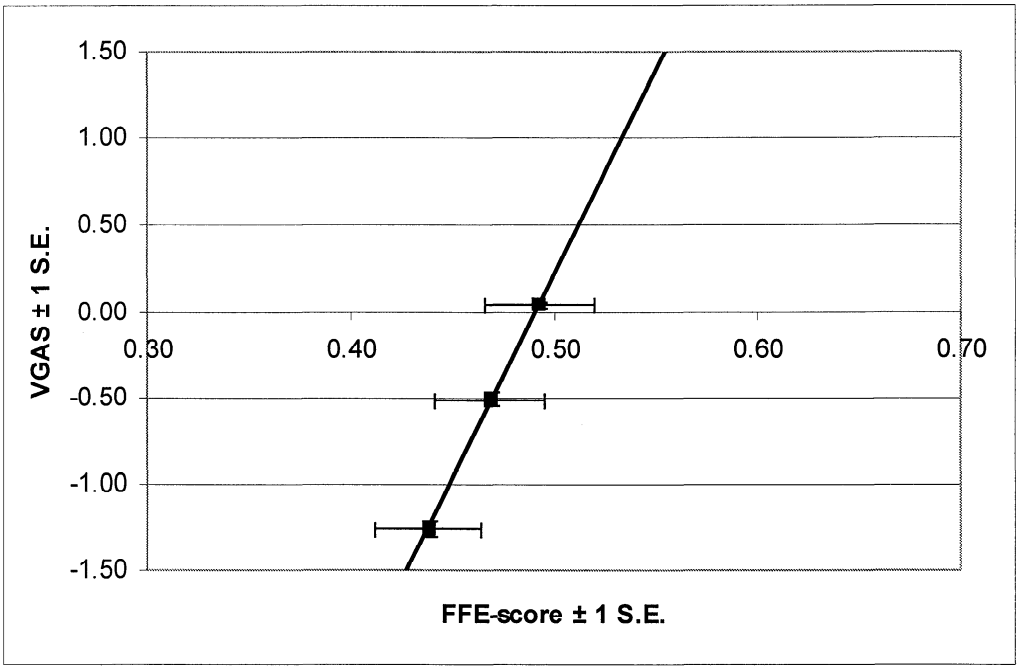

Figure 4. The results from the visual grading analysis (VGA) plotted against the results from the free response forced error (FFE) experiment for both types of lesions in region 1. The results indicate a linear relation between the outcome of the two methods. Note the small error bars for VGAS.

4. DISCUSSION

In this study we have developed a model for producing lumbar spine images with added pathology (hybrid images). The method uses a digitally simulated lesion that was inserted randomly in a digitised clinical image. Two types of lesions, destructive and sclerotic, which can be found in lumbar spine images, were simulated. This corresponds to a decreased and increased attenuation respectively. Three different combinations of noise and spatial resolution were applied to fifty different lesion distributions and evaluated in a study of detectability. The results from the study gave expected results: the images with the lowest noise and the highest spatial resolution had the highest scores and the images with the highest noise and the lowest spatial resolution had the lowest scores. This indicates that the method is capable of detecting differences in the speed-film system used.

The detectability of the lesions was higher in region 2 (the winged portion of the iliac bone) than in region 1 (the lumbar spine and the sacrum), even though the diameter of the lesions in region 2 was smaller than in region 1. The reason for this is probably the homogeneous appearance of the winged portion of the iliac bone compared to the lumbar spine and the sacrum. The circular shape of the lesions is thus more obscured in region 1 than in region 2. The results from the two regions show that the difficulty in finding the lesions was suitable for use in a study of detectability, 46% in region 1 and 59% in region 2. This means that the contrast of the lesions was subtle enough so that it was neither too easy (detectability→100%) nor too hard (detectability→0%). Figure 3 shows an interesting finding: In region 2 for the worst combination of noise and spatial resolution, the sclerotic lesions are significantly harder to detect than the destructive. This behaviour is not present in any of the other cases. The optical density in the winged portion of the iliac is rather low and the contrast of the film will therefore be low (the toe of the H/D curve). The sclerotic lesion lowers the optical density further, but this is hidden by the high noise and low spatial resolution of this image quality level. The corresponding destructive lesions is apparently not

obscured as much by the high noise and low spatial resolution. The small addition of optical density increases the contrast of the lesion and compensate for the high noise and low spatial resolution. This effect is not present in region 1 where the density is higher (linear part of the H/D curve) and the lesions are larger and therefore not affected as much by the MTF.

The results from the FFE study were compared with the results from a VGA study based on the structures mentioned in the European Quality Criteria[6]. The comparison was only done for region 1 since the structures used for the VGA study is positioned in the lumbar spine. FFE is based on detection of pathological structures in the images, while VGA is based on comparison of the visibility of anatomical structures. The results indicate a linear correlation between the outcomes of the two methods. The indication of a linear correlation is very promising, but more data points would of course be desirable to make a general conclusion (see also Tingberg et al. 1999[12]). Also, it would be interesting to test the correlation for images produced with other radiographic techniques (e.g. images taken at different tube voltages). The study has also been performed for chest radiography indicating a similar relationship[13].

The uncertainties in the VGAS values were lower than the uncertainties in the FFE values even though the number of images for each combination of noise and spatial resolution was more than three times higher in the FFE study. This shows that VGA is robust method and that statistically significant results can be achieved with few images. This makes the VGA suitable in a clinical environment as a quick test to evaluate whether a new method (e.g. a new screen-film combination or image processing algorithm) is as good as the old method. Care must however be taken when selecting the image that should be used as a reference image. The best solution is to use an image that originates from the same patient as the image that should be evaluated, which easily can be accomplished in digital radiography. This fact makes VGA an excellent method for evaluating the effect of image processing, where there is no need for exposing the same patient twice.

The time needed for setting up and running a VGA study is significantly shorter than for a FFE experiment. Since normal patient images can be used there is no need for - as in the various ROC methods - finding out whether there really is a lesion or not in the image (for example by using "gold standard observers" or when possible by autopsy) or to add artificial lesions to the image. The time needed for the observers to evaluate each image was in this study about three times shorter for the VGA images than for the hybrid images (there was no observation time limit when viewing the images). This is another important advantage for the VGA, since this kind of studies often last for several hours or days. Furthermore, in a strict sense, the results from ROC studies only apply to a certain kind of lesion that had to be detected whereas the approach of using the structures of the European Guidelines is much broader because it is based on normal anatomy, thus the results are more general.

Visual grading analysis is a simple method for evaluation of image quality, and it is not dependent on the access to images where the truth is known (i.e. if there actually are lesions present in the image). This is a well-known problem of ROC analysis. In this study we have overcome this limitation by introducing artificial lesions in the images. The simplicity of VGA makes it very suitable for use in the clinic for selecting the optimal radiographic technique.

5. CONCLUSIONS

We have developed a model for producing lumbar spine images with added pathology (hybrid images). The images were evaluated in a detection study and the results showed that the lesions which were added were subtle enough for the purpose and that the evaluation of the images could detect differences in the noise and spatial resolution of the system.

A method based on visual comparison of image quality compared to a reference image (VGA) based on the structures of the European Quality Criteria, which basically is a subjective method, proved to give the same results as the objective detection study (FFE). This result shows that a VGA study, which is much easier to set up because virtually any clinical image can be used and the evaluation procedure is less time consuming, can be used to evaluate the image quality of clinical images.

6. ACKNOWLEDGEMENTS

We are very grateful to the radiologists who evaluated the images and in other ways contributed to the study: Dr. Paloma Chimeno, Dr. Claudius Gückel, Dr. Susanne Kheddache, Prof. Mario Maffessanti, Dr. Dieter Saure, Prof. Graham Whitehouse.
Thanks are also due to Dr. Francis Verdun and his co-workers at the Department of Applied Radiation Physics, Lausanne, Switzerland, for performing the measurements of the H/D curves, MTF and Wiener spectra that were vital for this study.
The study was financially supported by the Radiation Protection Research Programme of the European Union no. FI4P-CT95-0005 "Predicitivity and optimisation in diagnostic radiology".

7. REFERENCES

1. C.E. Metz, "ROC methodology in radiologic imaging," *Invest. Radiol.* **21**(9), pp. 720-33, 1986.
2. E. Samei, M.J. Flynn and W.R. Eyler, "Simulation of subtle lung nodules in projection chest radiography [published erratum appears in Radiology 1997 Jun;203(3):884]," *Radiology* **202**(1), pp. 117-24, 1997.
3. R.H. Sherrier, G.A. Johnson, S.A. Suddarth, C. Chiles, C. Hulka and C.E. Ravin, "Digital synthesis of lung nodules," *Invest. Radiol.* **20**(9), pp. 933-7, 1985.
4. D.P. Chakraborty and L.H. Winter, "Free-response methodology: alternate analysis and a new observer-performance experiment," *Radiology* **174**(3), pp. 873-81, 1990.
5. A. Almén, A. Tingberg, S. Mattsson, J. Besjakov, S. Kheddache, B. Lanhede et al., "The influence of different technique factors on image quality of lumbar spine radiographs as evaluated by established CEC image criteria," *Submitted to Br. J. Radiol.*
6. European Commission. European guidelines on quality criteria for diagnostic radiographic images. Brussels: EUR 16260; 1996.
7. C. Herrmann, B. Lanhede, A. Tingberg, W. Panzer, A. Almén, S. Mattsson et al., "A system for the digital reproduction of conventional film radiographs of chest and lumbar spine," *manuscript in preparation.*
8. E. Buhr, C. Herrmann and D. Hoeschen, "Correlation between physical image quality parameters and visually perceptible image quality in X-ray diagnosis," *J. Phot. Sc.* **41**(3), pp. 90-1, 1993.
9. C. Herrmann, B. Lanhede, A. Tingberg, W. Panzer, A. Almén, S. Mattsson et al., "Methods for the digital simulation of certain image characteristics of conventional clinical radiographs of chest and lumbar spine," *manuscript in preparation.*
10. C. Herrmann, A. Tingberg, J. Besjakov and K. Rodenacker, "Simulation of nodule-like pathology in radiographs of the lumbar spine," *Submitted to Radiat. Prot. Dosimetry.*1999.
11. H. Manninen, E.O. Terho, M. Wiljasalo, S. Wiljasalo and S. Soimakallio, "An evaluation of different imaging chains in clinical chest radiography," *Br. J. Radiol.* **57**(683), pp. 991-5, 1984.
12. A. Tingberg, C. Herrmann, A. Almén, J. Besjakov, S. Mattsson, P. Sund et al., "Comparison of two methods for evaluation of the image quality of lumbar spine radiographs," *Submitted to Radiat. Prot. Dosimetry.*1999.
13. P. Sund, C. Herrmann, A. Tingberg, S. Kheddache, L.G. Månsson, A. Almén et al., "Comparison of two methods for evaluating image quality of chest radiographs," *SPIE Medical Imaging.* **3981**.2000.