Radiation Protection Dosimetry (2010), Vol. 139, No. 1–3, pp. 26–36
Advance Access publication 10 February 2010

doi:10.1093/rpd/ncq007

EVALUATING IMAGING SYSTEMS: PRACTICAL APPLICATIONS

Magnus Båth[1,2,*]
[1]Department of Radiation Physics, University of Gothenburg, Gothenburg SE-413 45, Sweden
[2]Department of Medical Physics and Biomedical Engineering, Sahlgrenska University Hospital, Gothenburg SE-413 45, Sweden

*Corresponding author: magnus.bath@vgregion.se

There are many ways in which imaging systems can be evaluated. The aim of the present paper is to provide an overview of a number of selected approaches to evaluating imaging systems, often encountered by the medical physicist, and discuss their validity and reliability. Specifically, it will cover (i) characterisation of an imaging system in terms of its detective quantum efficiency using linear-systems analysis; (ii) attempts to calculate relevant measures directly in images using the Rose model and the pixel signal-to-noise ratio; (iii) task-based methods incorporating human observers such as receiver-operating characteristics and (iv) visual grading-based methods using experienced radiologists as observers.

INTRODUCTION

Operationalisation—or: what was actually measured?

Operationalisation is the process of defining measurable variables thought to describe the phenomenon which is the subject of study. The operationalisation process is of the utmost importance in any scientific task since the reliability and validity of the results from a study are strongly connected to the success of the operationalisation. The reliability describes the precision of the measurement; a high reliability demanding small stochastic errors. The validity describes how well the variables describe the phenomenon; a high validity demanding a small systematic error. Successful operationalisation therefore requires both high validity and high reliability.

Image quality in medical imaging is a phenomenon of enormous complexity. It is extremely task dependent—the demands on noise level, resolution and contrast differing from discipline-to-discipline. It also involves many processes that are not fully understood and described, such as the effects of image processing and the anatomical background in an image on the signal detection and interpretation by the human observer. It is therefore easy to understand the difficulty in defining a general image quality measure that has high validity. On the other hand, a measure with high validity for a specific task has the inherent property of being less generalisable. This difficulty in performing successful operationalisation for an image quality measure has led to the diverse methods of evaluating imaging systems in use today.

The purpose of the present paper is to provide an overview over some common methods for evaluating imaging systems from an operationalisation point of view, i.e. to discuss the validity and reliability associated with the methods and hence their suitability for different practical applications. The overview is not intended to cover the field, but will focus on a limited number of groups of methods encountered by the medical physicist in routine work or research. The groups contain (i) methods related to linear-systems analysis (LSA), (ii) quantitative measurements in images, (iii) receiver-operating characteristics (ROC) analysis and (4) visual grading.

LSA FOCUSED ON DETECTIVE QUANTUM EFFICIENCY

A desire to describe the imaging properties of an imaging system in an objective way, without taking the specific imaging task into account, has led to the application of LSA to medical imaging systems. LSA, based on linear-systems theory(1), can be used to give measures of the ability of the system to pass a signal, as well as of the noise characteristics of the system. The reasoning behind the use of LSA is to give general and detailed descriptions of the imaging system in terms of properties that are believed to influence the clinical performance of the system.

Since the 1940s, many attempts have been made to quantify the efficiency of radiation detectors(2). In the first attempts, the quantum efficiency was based on the ratio of the number of output events to the number of input events, and was termed the responsive quantum efficiency (RQE). However, the RQE has several drawbacks, the major one being that it 'links the input/output numbers in quantity but not in quality'(2). Amplification anywhere along the signal chain increases the RQE by a proportional amount, which leads to the fact that the RQE does not have an upper limit. It is therefore impossible to compare a real detector with an ideal one in order to obtain an absolute quality measure. However, by comparing the *fluctuations* at the output stage to those at the input stage, a measure with an upper limit of unity is obtained for a linear system. Based on the ratio of fluctuations, the detective quantum efficiency (DQE) can be defined as follows:

$$\mathrm{DQE} = \frac{\mathrm{SNR}^2_{\mathrm{out}}}{\mathrm{SNR}^2_{\mathrm{in}}}. \tag{1}$$

where $\mathrm{SNR}_{\mathrm{out}}$ is the signal-to-noise ratio at the output stage and $\mathrm{SNR}_{\mathrm{in}}$ that at the input stage. The DQE is as such closely connected to the quantum nature of radiation. A measurement of radiation is always associated with an uncertainty, but by comparing the SNR at the output and input stages, the inherent fluctuations of the radiation are excluded from the characterisation of the detector, and the detector is compared with an ideal detector—a detector that detects all incoming quanta without adding any noise to the signal.

Although being a fundamental property of a detector, the DQE expressed as in Eq. (1) does not give enough information about an imaging system to be useful, since it does not take the resolution properties of the detector into account. This problem can be solved by expressing the DQE as a function of spatial frequency:

$$\mathrm{DQE}(u, v) = \frac{\mathrm{SNR}(u, v)^2_{\mathrm{out}}}{\mathrm{SNR}(u, v)^2_{\mathrm{in}}}, \tag{2}$$

where u and v denote orthogonal spatial frequencies for a two-dimensional imaging system. As such, the DQE describes the efficiency of the imaging detector completely since, for any given spatial frequency, it states the efficiency of the system in detecting that frequency compared with that of the ideal detector, which is ideal both in terms of detection and localisation of the incoming signal. Thus, the DQE describes the efficiency of an imaging system in the sense that it describes to what extent it utilises the information given as input to it. That the DQE takes both the sensitivity and resolution properties of an imaging system into account can be made even more explicit be expressing the DQE in the following way[3]:

$$\mathrm{DQE}(u, v) = \frac{\mathrm{MTF}(u, v)^2}{\mathrm{NNPS}(u, v)\mathrm{SNR}^2_{\mathrm{in}}}, \tag{3}$$

where the MTF is the modulation transfer function of the system, describing to what extent the amplitude of a given spatial frequency passing through the system is preserved, and the NNPS is the normalised noise power spectrum, which describes the variance of image intensity spread over the spatial frequencies in the image.

The use of the DQE as a fundamental measure of the imaging properties of a detector has shown a steady increase in recent years (Figure 1). The DQE concept is natural in any situation where the image quality is dependent on the number of detected photons. It has been widely used in projection radiography[4–7], where it is commonly accepted to be the most important measure of the imaging properties of a detector, but has also been proposed and used in CT[8] and nuclear medicine[9]. However, the DQE as a measure of how the imaging system maintains the SNR is general and should be applicable to any imaging system. Furthermore, it has been argued that the DQE is a better measure of the resolution properties of an imaging system than the MTF[10]. As the sharpness of a digital image can be altered using image processing, the MTF of the system does not describe the sharpness in the final image. As the amount of contrast enhancement that can be applied is limited by the noise, the DQE better describes the resolution that can be obtained.

The use of DQE for characterising the imaging properties of a detector is almost undisputed. However, it must be remembered that the DQE is only a descriptor of a single component in the imaging chain, namely the detector. Firstly, it gives little information about the final appearance of the resulting image, which is dependent on, for example, dose level, image processing and display characteristics. For example, regarding the importance of image processing, Sund *et al.*[11] performed a study in which the DQE of four different digital radiographic systems were compared with the clinical image quality, determined by letting experienced radiologists rate their opinion about the quality of chest radiographs collected with the four systems at similar dose levels and presented on one and the same monitor. A low correlation between the DQE of the system and the radiologists' impression of the

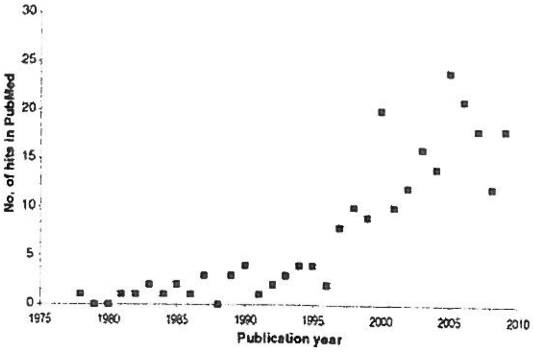

Figure 1. The number of published papers per year with 'DQE' as a keyword, according to a search in PubMed on 1 November 2009.

images was found. The major reason for such a finding is that the human is a limited observer, meaning that not only the information content in an image but also the appearance of an image is of importance. In the study by Sund *et al.*, the difference in image processing between the systems apparently was of larger importance for the clinical image quality than the difference in noise level resulting from the difference in DQE. Secondly, the detection of pathology may not be mainly hindered by stochastic noise such as quantum noise. For many detection tasks in radiography, for example, it has been shown that it is the anatomical background that limits the performance of the observer(11–20). Thus, the influence of the DQE on the image quality may be small for many examinations, as the DQE does not take the anatomical background into account. This also means that for a given system, the parameter setting resulting in the highest DQE may not be the optimal setting of the system. For systems where the stochastic noise level is higher, such as CT or gamma camera systems, the importance of the DQE on the resulting clinical image may be higher, meaning that the validity in the use of the DQE as a predictor of the clinical usefulness of a system is higher, but as long as the anatomy of the patient itself disturbs the observer, it cannot be taken for granted that a system with higher DQE results in a better performance in the clinical task.

In an attempt to take the anatomical background into account, the concept of a generalised DQE has been proposed(21). For the generalised DQE, not only the stochastic noise sources contribute to the noise power but also the power spectrum of the anatomical background is included. In the same way, as the conventional DQE is related to an observer for which the anatomical background has no effect on detection at all, the generalised DQE is related to an observer for which the entire anatomical background acts as random noise. However, as it has been shown that the extent to which the anatomical background acts as random noise for a human observer substantially differs between different combinations of type of pathology and type of background(13,18), ranging from having almost no effect at all to almost completely acting as random noise, it is not obvious that the generalised DQE in general actually is more related to the human observer than is the conventional DQE.

Linear-systems analysis, with emphasis on the determination of DQE, has mainly been used to compare the imaging properties of different detectors. However, it has also been proposed as part of quality assurance programmes to ensure that a certain level of image quality is maintained(22,23). A substantial amount of effort has been made over the years to ensure that DQE determinations can be performed in a reliable way. Until recently, evident deviations between different methods of determining DQE have been reported, mainly due to differences in the MTF determination(24,25). However, there now exist standards for DQE determinations in general radiography(26), mammography(27) as well as dynamic imaging(28), for which the reliability has been shown to be high(29). The validity of the above-mentioned standards in providing accurate data on the information transfer abilities of the imaging detector is high, whereas, as discussed above, the validity of using the DQE as a measure of the entire imaging system is lower.

THE ROSE MODEL AND PIXEL SNR

A common simplistic approach to describe the visibility of an object in an image is to determine the ratio of the mean signal of the object (the difference between the average pixel value in the object and the average pixel value in the background) and the pixel standard deviation in the background. This SNR, referred to as pixel SNR, or $\mathrm{SNR_p}$, is sometimes used as a measure of image quality. As will be described below, the validity of this measure of image quality is very limited.

Although frequently used for objects of different size, the $\mathrm{SNR_p}$ is the special case when the Rose model(30) is applied to an object with a size given by one pixel. The Rose model is an attempt to describe how the human observer detects a flat-topped sharp-edged signal of area A in a uniform background containing uncorrelated Poisson noise. The count level in the background is $\langle n_\mathrm{b}\rangle$ expected number of photons per unit area, whereas the signal contains $\langle \Delta n_\mathrm{s}\rangle$ extra photons per unit area, resulting in a contrast $C = \langle \Delta n_\mathrm{s}\rangle / \langle n_\mathrm{b}\rangle$ for the object. The Rose model SNR for such an object is defined as follows(31):

$$\mathrm{SNR_{Rose}} = C\sqrt{A\langle n_\mathrm{b}\rangle}. \qquad (4)$$

The Rose model has been shown to agree well with the human observer, given that the requirements of the model are fulfilled. Rose aimed to find a threshold value for the $\mathrm{SNR_{Rose}}$ for an object to be visible for a human observer. Commonly, it is stated that a threshold value of 5 is needed for the object to be detected. Although the detection of an object is dependent on the confidence threshold of the observer, the given threshold value corresponds well with the typical threshold used by the human observer. In Figure 2, a number of disc-like objects of different size and contrast are presented in a uniform background containing uncorrelated Poisson noise. As can be seen, the objects for which the $\mathrm{SNR_{Rose}}$ is above five match closely those that are detected by the human eye.

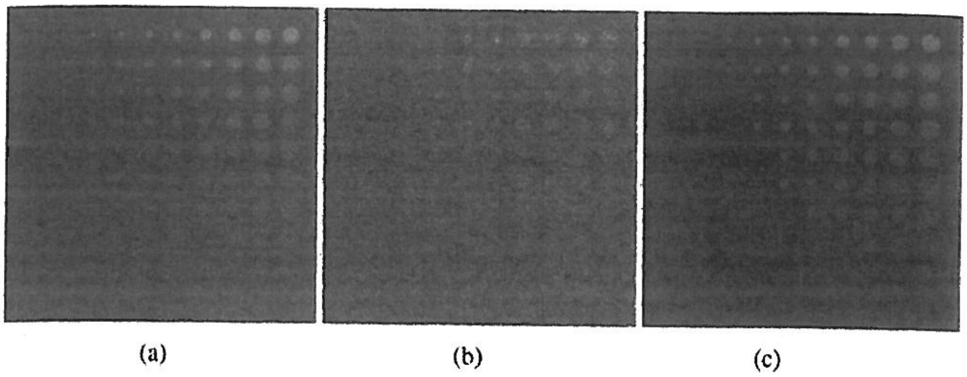

Figure 2. (a) An image containing 10 × 10 disc-like objects, grouped into columns by increasing size (radius = 2, 4, 6, ..., 20 pixels) and rows by increasing contrast (C = 0.5, 1, 1.5, ..., 5 %), embedded in a constant background pixel value of 100. (b) The image in (a) after an exchange to a uniform background containing uncorrelated noise sampled from a Poisson distribution with a mean of 100. (c) An image containing the discs in (a) for which the SNR_{Rose} in (b) is 5 or higher. Note that the discs in (c), predicted by the Rose model to be detectable, correspond well with the discs that actually are visible in (b).

As described above, the SNR_p is related to the Rose model, being equal to SNR_{Rose} when the object has the size of one pixel and the requirements of the Rose model are fulfilled. Thus, the SNR_p is a decent description of the possibility for a human observer in detecting an object of size one pixel in a background of uncorrelated Poisson noise. However, as the SNR_p does not take the size of the object into account, its correlation with the human observer is in general low. The SNR_p for the objects in each row in Figure 2b is constant, clearly demonstrating that SNR_p is not a relevant measure for objects of different size. Furthermore, the SNR_p is often used as a measure of image quality when several of the requirements for the Rose model are not fulfilled. For example, the SNR_p is often determined in images containing noise other than uncorrelated Poisson noise. As the human observer is sensitive for the texture of the noise, the noise description used in SNR_p (the standard deviation of the pixel fluctuations) is overly simplistic. This can be seen in Figure 3, where the same objects are embedded in two different backgrounds with different noise texture, but the same pixel standard deviation of the noise. Clearly, the objects in Figure 3b are better visualised than the objects in Figure 3a, although the SNR_p is the same in the two images. A third frequent misconception relates to the pixel size. Given the same imaging conditions, the SNR_p is lower in an image with smaller pixels. Based on this, it is often stated that it is necessary to use a larger number of photons for the image if a smaller pixel size is used in the image. However, the human observer, usually not interested in single pixel values, integrates information over an area in an image and is rarely affected by the pixel-to-pixel fluctuations. Thus, the validity of the statement that smaller pixels require a higher dose is in general low. A fourth limitation of the SNR_p is that if the ROI used to determine the standard deviation is positioned in a region of the image where non-homogeneous anatomy is present, the variation in pixel value related to the anatomy may heavily influence the measure. The anatomical background rarely affects the human observer as if it had been pure noise[13,18].

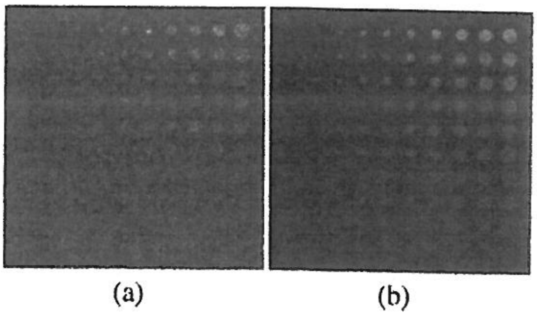

Figure 3. (a) The image in Figure 2b after smoothing with a 5 × 5 boxcar filter. (b) The image in Figure 2a after an exchange to a uniform background containing uncorrelated noise sampled from a Poisson distribution with a mean of 2500, resulting in the same pixel standard deviation of the noise as in (a). The SNR_p is equal for all objects in the same rows in (a) and (b), clearly showing that the SNR_p is not a relevant measure for objects of different size or noise with different texture.

Thus, the SNR_p is related to an observer that decides whether a signal is present or not by looking at the deviation of a single pixel value in the object from the pixel-to-pixel fluctuations in the background, without taking into account the size of the object, the size of the pixels or the texture of the noise. Such an observer has very little in common with the human observer. Consequently, the validity of using SNR_p as a meaningful measure of image quality is in general very low and its use should be avoided in the comparison of different imaging systems or different image processing techniques etc. However, as the reliability of the measure in general is high, it may be suitable for, for example, constancy control, as the purpose of such a programme is to detect changes over time.

ROC ANALYSIS

The fundamental task for an observer in medical imaging is to state whether an image belongs to a healthy patient or whether the patient has a disease. This has led to the need for characterising the performance of the observer. An intuitive measure of the quality of the observer might be the number of correct responses. However, such a measure has a serious drawback in that it is strongly dependent on the prevalence of signal (or disease). As an example, imagine an image data set corresponding to a selection of patients of which only 1 % suffers from a specific disease. If the observer in this case would state that the patient is healthy in all cases, he would, despite the failure of not detecting a single pathological case, end up with the impressive number of 99 % correct responses. Thus, it is easily understood that a relevant measure needs to be independent of the prevalence of signal. Sensitivity (the probability that a patient with an actual disease is determined as having a disease by the observer) and specificity (the probability that a healthy patient is determined as being healthy by the observer) are two common measures that fulfil the requirement of independence of the prevalence of signal. However, for a given observer the sensitivity and the specificity are closely correlated in that an increase in sensitivity, stemming from a change in the decision threshold in most cases, inevitably results in a decrease in specificity. This dependency on decision threshold leads to difficulties in comparing different observers.

However, the varying choices of the decision threshold are the essence of ROC analysis. The method provides a natural distinction between the inherent detectability of the signal (or disease) and the judgement of the observer, reflected in the positioning of the decision threshold. By deliberately varying the decision thresholds, the trade-offs between the true positive fraction (TPF = sensitivity) and the false positive fraction (FPF = 1 − specificity) can be established. In Figure 4a, four such choices of decision thresholds are shown. (A confidence scale with five steps leads to four decision thresholds.) In principle, an infinite number of decision thresholds can be considered, thus generating a continuous ROC curve with the TPF given as a function of the FPF (Figure 4b). Such curves allow one to directly compare the inherent diagnostic capabilities of different diagnostic procedures. A more accurate procedure will generate a curve closer to the top left corner than a less accurate one.

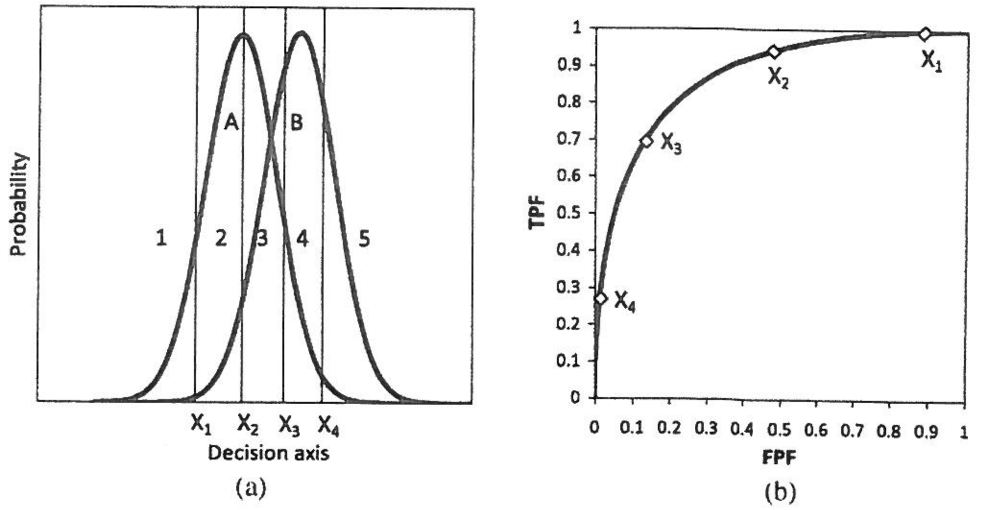

Figure 4. (a) Probability distributions (A: no signal present, B: signal present) for a detection task showing four levels of decision thresholds X_1–X_4. Values of $X \leq X_1$ correspond to the first rating category (1), $X_1 < X \leq X_2$ to the second (2), etc. and $X > X_4$ to the last (5). (b) The resulting ROC curve, giving the TPF as a function of the FPF. The four operating points corresponding to the four decision thresholds in (a) are given by the diamonds in (b).

Curves situated on or near the diagonal represent totally non-informative procedures, the results of which are no better than pure guesswork. Thus, an ROC curve describes all possible compromises between true positive and false positive decisions inherent in a diagnostic procedure. Like sensitivity and specificity, it is independent of the prevalence of signal (or disease). Furthermore, ROC analysis is independent of any effect that different decision thresholds might have on the diagnostic process.

The accuracy index most often used is the area under the ROC curve, A_z, for which the range of values is [0.5, 1.0], where 0.5 represents detection governed by chance only and 1.0 represents perfect detection. As such, it is a useful quantitative measure of the performance of the observer. In fact, the accuracy index A_z has a direct relation to statistical decision theory. It can be shown(32) to be equal to the proportion of correct choices in two alternative forced-choice experiments, in which the observer is presented with two images—one image containing a signal, the other not—at the same time, with the task of deciding which of the two images contains the signal.

The most complete way of performing an ROC study comparing different modalities is to let a number of readers interpret the same cases in all modalities. Such a study design, commonly referred to as the multiple-reader multiple-case (MRMC) study design, is needed to be able to generalise the conclusions of a study to the population of readers and cases, but has the inherent property that the observer data are correlated, which makes the statistical analysis cumbersome. However, a method of correctly analysing such correlated ROC data has been described by Dorfman, Berbaum and Metz (DBM) using a jack-knifing approach(33), and the DBM MRMC ROC methodology is generally considered the method-of-choice for ROC analysis.

Although ROC analysis is often referred to as the gold standard of evaluating performance in medical imaging, there are weaknesses when it is applied to localisation tasks. For example, if an observer misses the single true lesion in an image but erroneously identifies another location as containing a lesion, the observer makes two mistakes: a false negative (misses the true lesion) and a false positive (reports a non-lesion). However, the two mistakes effectively cancel out each other, and on the case level the observer is scored with a true positive. In the free-response paradigm, first being recognised as important for medical imaging by Bunch *et al.*(34), the analysis is conducted on the lesion level instead of on the case level, the so-called free-response ROC (FROC) analysis. The observer marks suspicious regions in each case, which may be either lesion localisations (if the mark coincides with a true lesion, which may be referred to as a true positive mark) or non-lesion localisations (if the mark does not coincide with a true lesion, which may be referred to as a false positive mark) The FROC curve is a plot of a lesion localisation fraction (the ratio of the number of lesion localisations and the number of lesions) versus non-lesion localisation fraction (the ratio of the number non-lesion localisations and the number of cases) as the threshold confidence level is varied, and is the FROC counterpart to the ROC curve in ROC analysis. In the same way, as ROC analysis had difficulties in correctly analysing data from an MRMC ROC study prior to the DBM approach, using the FROC curve for analysing MRMC FROC studies initially proved to be difficult. However, this was recently solved with the jackknife alternative FROC (JAFROC) analysis(35), which applies to MRMC FROC studies in the same way as the DBM applies to MRMC ROC studies(36). Solving the problem of correctly analysing the correlated data from an MRMC FROC study, JAFROC analysis proved to be a valuable analysis method that has earned great interest(37–43). A positive side effect of performing the analysis on the lesion level instead of on the case level is also the increased statistical power(44).

Due to the scientific soundness, well-established statistical analysis and close connection to the clinical task, ROC and ROC-related methods such as JAFROC analysis are suitable for large-scale image quality trials, in order to compare, for example, different modalities in terms of detectability of specific pathology. However, although conducting these types of experiments are the gold standard for image quality evaluation, concerns have been raised regarding their clinical relevance. For example, the observers—although being, for example, experienced radiologists—may behave differently in the laboratory situation resulting from the experiment compared with in the clinical environment(45). Also, conducting these types of studies may be cumbersome, since they are based on the establishment of truth for all cases and normally require a large number of cases in order to produce statistically significant results (meaning that the reliability is relatively low). For these reasons, ROC-related methods may not be the method of choice for the local optimisation task at a radiology department where the intention is to find, for example, the optimal image processing setting or dose level for a given examination.

VISUAL GRADING

A different approach to assessing image quality, involving human observers, is to let the observers rate the visibility of details in an image. Performing such a study in a controlled scientific manner is usually termed visual grading. Using visual grading of the reproduction of important anatomical structures in clinical images for evaluating image quality has become an established method for several reasons. First of all, the validity of such studies can be assumed to be high if the anatomical structures are selected based on their clinical relevance and if the observers are experienced radiologists. Second, visual grading methods have in special cases been shown to agree both with detection studies using human observers(46,47) and with advanced calculations of the physical image quality(48–50). This is important, and validates in some way the assumption that the possibility to detect pathology correlates to the reproduction of anatomy—the basic idea of visual grading. Discrepancies between the methods have been reported(51), but have been explained with the different tasks for the methods rather than low validity for visual grading. Third, visual grading studies are relatively easy to conduct, especially in comparison with ROC studies, which is important when optimising equipment at the local level. How to perform visual grading studies has been extensively described(52–56), and the learning threshold for conducting such studies is low. Fourth, the time consumption is moderate, at least for the observers, which means that it is realistic to believe that these methods can be implemented at almost any hospital. The workload on each participating radiologist is typically in the order of a few hours, which means that a study is easy to justify from an economical perspective for the hospital.

Arguments against the use of visual grading are often presented. Some of these relate to the subjective nature of the task and state that studies of this type amount to a 'beauty-contest'(57), meaning that they are prone to bias. For example, in a clinical trial of screen–film combinations in portable chest radiography where the observers were asked to indicate their preference without any criteria, they usually chose the modality in use in their department at the time(58)). However, according to Kundel(59) the images of the highest diagnostic quality are those that enable the observer 'to most accurately report diagnostically relevant structures and features'. To reduce the risk of bias, an international group of well-established radiologists and physicists developed the European quality criteria(60–62), which for specific examinations in general radiography, paediatric radiography and CT state important anatomical landmarks and their needed level of reproduction to aid accurate diagnosis (see Table 1 for an example of image criteria for an examination). Letting experienced radiologists rate the visibility of these landmarks is a study design closely related to the description of diagnostic image quality by Kundel.

Table 1. Image criteria for a posteroanterior (PA) chest radiograph from European guidelines(60).

Image criterion	Description
1.1.1	Performed at full inspiration (as assessed by the position of the ribs above the diaphragm—either 6 anteriorly or 10 posteriorly) and with suspended respiration
1.1.2	Symmetrical reproduction of the thorax as shown by the central position of the spinous process between the medial ends of the clavicles
1.1.3	Medial border of the scapulae to be outside the lung fields
1.1.4	Reproduction of the whole rib cage above the diaphragm
1.1.5	Visually sharp reproduction of the vascular pattern in the whole lung, particularly the peripheral vessels
1.1.6	Visually sharp reproduction of: (a) the trachea and proximal bronchi, (b) the borders of the heart and aorta, (c) the diaphragm and lateral costo-phrenic angles
1.1.7	Visualisation of the retrocardiac lung and the mediastinum
1.1.8	Visualisation of the spine through the heart shadow

Some criteria depend on correct positioning and cooperation of the patient, whereas others reflect technical performance of the imaging system. The latter are suitable for visual grading studies.

Visualisation: characteristic features are detectable but details are not fully reproduced; features just visible.

Reproduction: Details of anatomical structures are visible but not necessarily clearly defined; details emerging.

Visually sharp reproduction: anatomical details are clearly defined; details clear.

Visual grading based on the European quality criteria or similar criteria has been used extensively, mainly in studies evaluating different settings—such as dose level, beam quality or image processing—of a given equipment or comparing different equipments with each other. Visual grading can be performed either by using the image criteria themselves, and letting the observers state whether they are fulfilled or not, or by identifying the anatomical structures that are most relevant to the criteria and letting the observers rate the visibility of these structures on a multi-step rating scale. The former is usually referred to as image criteria scoring and the resulting proportion of fulfilled criteria (usually averaged over all observers, cases and criteria) is referred to as an image criteria score (ICS). The latter is usually referred to as visual grading analysis (VGA) and can either be performed in a relative manner, where each image is compared with a reference image and the observer states whether the details in the image are reproduced better or worse than in the reference image, or in an absolute manner, where the observer gives a statement about the visibility of each detail on an absolute scale. For the analysis of the data collected in a VGA study, numerical values of the given ratings are often used to calculate a relative visual grading analysis score (VGAS) or an absolute VGAS, depending on the type of study, by averaging the given ratings over all cases, observers and structures.

The above-mentioned ways of analysing visual grading data can be questioned in at least two ways. The first refers to whether a complex phenomenon such as image quality really can be reduced to a single number, even if it is based on the visibility of relevant structures in a clinical image. However, more details can be obtained from a visual grading study if the summation over structures/criteria is omitted and a score is obtained for each structure/criterion. Since the visibility of different structures is sensitive to variations in contrast, resolution and noise in different ways, a separate analysis of each criterion or structure may reveal information that is hidden in the total score. The second objection refers to the way the originally qualitative (ordinal) scales are attributed quantitative properties(56,63,64). The central limit theorem states that the mean value of a variable that can take the value of either zero or unity, such as the ICS, is normally distributed for large samples(65). However, although the scale steps used in a VGA study may have been labelled with numerical values, they still belong to an ordinal scale. Calculating the mean value of data belonging to an ordinal scale, as is done for the VGAS, is a statistically forbidden (or meaningless) operation as such statistics imply knowledge of more than the relative rank order of data(66). This has earned some interest in recent years and attempts to analyse visual grading data without violating the statistical limitations of ordinal data have resulted in the development of visual grading characteristics (VGC) analysis(56) and visual grading regression (VGR)(64). Both methods treat the scale steps as ordinal and no assumptions about the distribution of the data need to be made. The basic VGC study is an expanded image criteria scoring study in which the observer uses a multi-step rating scale to state his opinion about the fulfilment of an image quality criterion(56). In this way VGC can be interpreted as a repeated image criteria scoring, where the observer changes his threshold for the fulfilment of each criterion in a similar way as when the scale steps in a ROC study are used by the observer to state the confidence of each decision. By plotting the cumulative distributions of the rating data for two compared systems against each other, a VGC curve is obtained which gives the ICS for one system (the evaluated system) as a function of the ICS for another system (the reference system). The area under the VGC curve (AUC_{VGC}) can be used as a measure of the difference in image quality between the two systems, where an AUC_{VGC} of 0.5 indicates an overall equal image quality for the two systems, an $AUC_{VGC} <$ 0.5 indicates that the image quality is higher for the reference system and an $AUC_{VGC} > 0.5$ indicates that the image quality is higher for the evaluated system. The ordinal data from a conventional absolute VGA study can also be analysed using the VGC approach. In VGR, ordinal logistic regression is applied to data from single-image and image-pair experiments with visual grading scores selected on an ordinal scale. The approach is applicable for situations where, for example, the effects of the choice of imaging equipment and post-processing method are to be studied simultaneously, while controlling for potentially confounding variables such as patient and observer identity. Hopefully, VGC analysis or VGR will in the coming years be expanded to enable MRMC visual grading studies to be accurately analysed in the same way as the DBM approach solves MRMC ROC studies and the JAFROC approach solves MRMC FROC studies.

As described above, the validity of a visual grading study is in general accepted to be high, as is it closely related to the general clinical task in medical imaging of assessing whether pathology is present or not in an image; for this task to be fulfilled it is essential that the anatomy be adequately reproduced. However, the reliability is in general relatively low and usually a large number of cases are needed, as in ROC analysis. However, as visual grading is commonly based on images of normal patients, collecting as large a number of cases as needed is relatively straightforward.

GENERAL DISCUSSION

In the present paper, an attempt has been made to provide an overview over some common methods for evaluating imaging systems from an operationalisation point of view, i.e. to discuss the validity and reliability associated with the methods and hence their suitability for different practical applications. There is a danger in mixing the concepts of validity and reliability in such a way that one is led to believe that a high reliability implies also a high validity. The reliability of a measurement can be very high, but the validity still non-existent if conclusions are drawn about a phenomenon the measure does not describe. This is especially poignant in medical imaging. An example is the use of LSA to evaluate system performance by describing the imaging properties of an imaging system through the use of the quantity DQE. This quantity can be determined with high reliability and the validity in terms of systematic errors is acceptable, as is discussed above. However, if the results are used to draw conclusions about phenomena other than that used as the basis for the original operationalisation, namely the transfer of SNR, the validity is naturally immediately reduced. Using the DQE as a measure of the clinical performance of a system without establishing the relationship between this quantity and measures taking into account the complete imaging chain, involving image processing, display and the response of the observer, is therefore an approach with low validity. The DQE should not be underestimated since the imaging properties of the detector constitute an important link in the imaging chain. However, in situations where the clinical image quality is more affected by disturbing anatomical structure than by quantum noise and system noise, the DQE is a less important parameter. There is therefore a risk of overestimating the importance of the DQE in such a way that it is assumed that a system with a higher DQE always results in higher clinical image quality. The validity of the DQE must be assessed in all clinical situations if it is to be used as a measure of clinical system performance.

Regarding quantitative measurements in images, it is the opinion of the author that these should be restricted to constancy control. In contrast to LSA, which is mostly used with the intention to describe the imaging system itself excluding the observer, measurements in an image are often used as a surrogate for human observers. Till date, there does not exist a model observer that correlates well with the performance of the human observer over the range of complex image backgrounds that do exist in medical images. Especially the frequently used SNR_p has been shown to have very little in common with the human observer. Thus, these types of calculations should be avoided in situations where the purpose is to compare different types of images. However, quantitative measurements are useful for constancy control. The inherent fluctuations in a human observer usually make the number of observations needed to achieve reliable results high. Nevertheless, the combination of a single observer and a single image is frequently used in quality assurance programmes when, for example, a medical physicist tries to determine whether the low-contrast visibility for given equipment has deteriorated over time. It has been shown that simple calculations directly in images have a much better chance of accurately detecting such a change than human observers[67,68].

To conclude, the inherent fluctuations in a human observer imply that all methods that involve human observers have limited reliability, meaning that a large number of observations are usually needed in order to obtain reliable results. On the other hand, as clinical images are evaluated by humans, human observers are mostly needed to obtain results that are valid for the entire imaging system. Although a lot of research over the years has aimed at bridging the gap between human observers, model observers and physical measurements, there is still a lot of work to be done until the goal is reached. Until then, when trying to evaluate an imaging system, it may be of value to bear in mind the words on a sign that Einstein kept on his wall: 'Not everything that counts can be counted; not everything that can be counted counts'[69].

ACKNOWLEDGEMENTS

The author would like to thank Angelica Svalkvist and Lars Gunnar Månsson for their contribution to the present work.

REFERENCES

1. Cunningham, I. A. *Applied linear-systems theory.* In: Handbook of Medical Imaging. **Vol. 1** Physics and Psychophysics, Beutel, J., Kundel, H. L. and Van Metter, R. L. Eds. (Bellingham: SPIE Press) pp. 79–159 (2000).
2. Dainty, J. C. and Shaw, R. *Image Science; Principles, Analysis and Evaluation of Photographic-Type Imaging Processes* (London: Academic Press) (1974).
3. Dobbins, J. T. III. *Image quality metrics for digital systems.* In: Handbook of Medical Imaging. **Vol. 1** Physics and Psychophysic, Beutel, J., Kundel, H. L. and Van Metter, R. L. Eds. (Bellingham: SPIE Press) pp. 161–222 (2000).
4. Tapiovaara, M. J. and Wagner, R. F. *SNR and DQE analysis of broad spectrum x-ray imaging.* Phys. Med. Biol. **30**, 519–529 (1985).
5. Dobbins, J. T. III, Ergun, D. L., Rutz, L., Hinshaw, D. A., Blume, H. and Clark, D. C. *DQE(f) of four generations of computed radiography acquisition devices.* Med. Phys. **22**, 1581–1593 (1995).
6. Granfors, P. R. and Aufrichtig, R. *Performance of a 41 × 41-cm² amorphous silicon flat panel x-ray detector for radiographic imaging appplications.* Med. Phys. **27**, 1324–1331 (2000).
7. Båth, M., Sund, P. and Månsson, L. G. *Evaluation of the imaging properties of two generations of a CCD-based system for digital chest radiography.* Med. Phys. **29**, 2286–2297 (2002).
8. Tward, D. J. and Siewerdsen, J. H. *Cascaded systems analysis of the 3D noise transfer characteristics of flat-panel cone-beam CT.* Med. Phys. **35**, 5510–5529 (2008).
9. Starck, S.-Å., Båth, M. and Carlsson, S. *The use of detective quantum efficiency (DQE) in evaluating the performance of gamma camera systems.* Phys. Med. Biol. **50**, 1601–1609 (2005).
10. Moy, J. P. *Signal-to-noise ratio and spatial resolution in x-ray electronic imagers: is the MTF a relevant parameter?* Med. Phys. **27**, 86–93 (2000).
11. Sund, P., Båth, M., Kheddache, S. and Månsson, L. G. *Comparison of visual grading analysis and determination of detective quantum efficiency for evaluating system*

performance in digital chest radiography. Eur. Radiol. **14**, 48–58 (2004).

12. Samei, E., Flynn, M. J. and Eyler, W. R. *Detection of subtle lung nodules: relative influence of quantum and anatomic noise on chest radiographs.* Radiology **213**, 727–734 (1999).
13. Bochud, F. O., Valley, J.-F., Verdun, F. R., Hessler, C. and Schnyder, P. *Estimation of the noisy component of anatomical backgrounds.* Med. Phys. **26**, 1365–1370 (1999).
14. Burgess, A. E., Jacobson, F. L. and Judy, P. F. *Human observer detection experiments with mammograms and power-law noise.* Med. Phys. **28**, 419–437 (2001).
15. Båth, M., Håkansson, M., Börjesson, S., Kheddache, S., Grahn, A., Ruschin, M., Tingberg, A., Mattsson, S. and Månsson, L. G. *Nodule detection in digital chest radiography: introduction to the RADIUS chest trial.* Radiat. Prot. Dosimetry **114**, 85–91 (2005).
16. Håkansson, M., Båth, M., Börjesson, S., Kheddache, S., Flinck, A., Ullman, G. and Månsson, L. G. *Nodule detection in digital chest radiography: effect of nodule location.* Radiat. Prot. Dosimetry **114**, 92–96 (2005).
17. Håkansson, M., Båth, M., Börjesson, S., Kheddache, S., Allansdotter Johnsson, Å. and Månsson, L. G. *Nodule detection in digital chest radiography: effect of system noise.* Radiat. Prot. Dosimetry **114**, 97–101 (2005).
18. Båth, M., Håkansson, M., Börjesson, S., Kheddache, S., Grahn, A., Bochud, F. O., Verdun, F. R. and Månsson, L. G. *Nodule detection in digital chest radiography: part of image background acting as pure noise.* Radiat. Prot. Dosimetry **114**, 102–108 (2005).
19. Båth, M., Håkansson, M., Börjesson, S., Hoeschen, C., Tischenko, O., Kheddache, S., Vikgren, J. and Månsson, L. G. *Nodule detection in digital chest radiography: effect of anatomical noise.* Radiat. Prot. Dosimetry **114**, 109–113 (2005).
20. Håkansson, M., Båth, M., Börjesson, S., Kheddache, S., Grahn, A., Ruschin, M., Tingberg, A., Mattsson, S. and Månsson, L. G. *Nodule detection in digital chest radiography: summary of the RADIUS chest trial.* Radiat. Prot. Dosimetry **114**, 114–120 (2005).
21. Richard, S., Siewerdsen, J. H., Jaffray, D. A., Moseley, D. J. and Bakhtiar, B. *Generalized DQE analysis of radiographic and dual-energy imaging using flat panel detectors.* Med. Phys. **32**, 1397–1413 (2005).
22. Cunningham, I. A. *Use of the detective quantum efficiency in a quality assurance program.* Proc. SPIE **6913**, p69133I (2007).
23. Schegerer, A. A., Schlatt, H., Renger, B., Dietz, W., Brunner, C. and Hoeschen, C. *Quality control of CT system: a new, objective approach.* Radiat. Prot. Dosim. **139**, 439–442 (2010).
24. Illers, H., Buhr, E., Günther-Kohfahl, S. and Neitzel, U. *Measurement of the modulation transfer function of digital X-ray detectors with an opaque edge-test device.* Radiat. Prot. Dosimetry **114**, 214–219 (2005).
25. Neitzel, U., Günther-Kohfahl, S., Borasi, G. and Samei, E. *Determination of the detective quantum efficiency of a digital x-ray detector: comparison of three evaluations using a common image data set.* Med. Phys. **31**, 2205–2211 (2004).
26. International Electrotechnical Commission. *Medical electrical equipment—characteristics of digital X-ray imaging devices—Part 1: determination of the detective quantum efficiency.* IEC 62220-1 (Geneva: IEC) (2003).
27. International Electrotechnical Commission. *Medical electrical equipment—characteristics of digital X-ray imaging devices—Part 1-2: determination of the detective quantum efficiency—detectors used in mammography.* IEC 62220-1-2 (Geneva: IEC) (2007).
28. International Electrotechnical Commission. *Medical electrical equipment—characteristics of digital X-ray imaging devices—Part 1-3: determination of the detective quantum efficiency—detectors used in dynamic imaging.* IEC 62220-1-3 (Geneva: IEC) (2008).
29. Illers, H., Buhr, H. and Hoeschen, C. *Measurement of the detective quantum efficiency (DQE) of digital X-ray detectors according to the novel standard IEC 62220-1.* Radiat. Prot. Dosimetry **114**, 39–44 (2005).
30. Rose, A. *The sensitivity performance of the human eye on an absolute scale.* J. Opt. Soc. Am. **38**, 196–208 (1948).
31. Burgess, A. E. *The Rose model, revisited.* J. Opt. Soc. Am. A **16**, 633–646 (1999).
32. Green, D. M. and Swets, J. A. *Signal Detection Theory And Psychophysics* (New York: Wiley) (1966).
33. Dorfman, D. D., Berbaum, K. S. and Metz, C. E. *ROC characteristic rating analysis: generalization to the population of readers and patients with the jackknife method.* Invest. Radiol. **27**, 723–731 (1992).
34. Bunch, P. C., Hamilton, J. F., Sanderson, G. K. and Simmons, A. H. *A free-response approach to the measurement and characterization of radiographic-observer performance.* J. Appl. Photogr. Eng. **4**, 166–171 (1978).
35. Chakraborty, D. P. and Berbaum, K. S. *Observer studies involving detection and localization: modeling, analysis and validation.* Med. Phys. **31**, 2313–2330 (2004).
36. Chakraborty, D. P. *A status report on free-response analysis.* Radiat. Prot. Dosimetry. Epub ahead of print January 18 doi:10.1093/rpd/ncp305 (2010).
37. Penedo, M. *et al. Free-response receiver operating characteristic evaluation of lossy JPEG2000 and object-based set partitioning in hierarchical trees compression of digitzed mammograms.* Radiology **237**, 450–457 (2005).
38. Brennan, P. C., McEntee, M., Evanoff, M., Phillips, P., O'Connor, W. T. and Manning, D. J. *Ambient lighting: effect of illumination on soft-copy viewing of radiographs of the wrist.* Am. J. Roentgenol. **188**, W177–W180 (2007).
39. Ruschin, M. *et al. Dose dependence of mass and microcalcification detection in digital mammography: free response human observer studies.* Med. Phys. **34**, 400–407 (2007).
40. Svahn, T., Hemdal, B., Ruschin, M., Chakraborty, D. P., Andersson, I., Tingberg, A. and Mattsson, S. *Dose reduction and its influence on diagnostic accuracy and radiation risk in digital mammography: an observer study using an anthropomorphic breast phantom.* Br. J. Radiol. **80**, 557–562 (2007).
41. Brennan, P. C., Ryan, J., Evanoff, M., Toomey, R., O'Beirne, A., Manning, D., Chakraborty, D. P. and McEntee, M. *The impact of acoustic noise found within clinical departments on radiology performance.* Acad. Radiol. **15**, 472–476 (2008).

42. Vikgren, J., Zachrisson, S., Svalkvist, A., Johnsson, Å. A., Boijsen, M., Flinck, A., Kheddache, S. and Båth, M. *Comparison of chest tomosynthesis and chest radiography for detection of pulmonary nodules: human observer study of clinical cases.* Radiology **249**, 1034–1041 (2008).

43. Zachrisson, S., Vikgren, J., Svalkvist, A., Johnsson, Å. A., Boijsen, M., Flinck, A., Månsson, L. G., Kheddache, S. and Båth, M. *Effect of clinical experience of chest tomosynthesis on detection pulmonary nodules.* Acta Radiol. **50**, 884–891 (2009).

44. Chakraborty, D. P. *Validation and statistical power comparison of methods for analyzing free-response observer performance studies.* Acad. Radiol. **15**, 1554–1566 (2008).

45. Gur, D., Bandos, A. I., Furhman, C. R., Klym, A. H., King, J. L. and Rockette, H. E. *The prevalence effect in a laboratory environment: changing the confidence ratings.* Acad. Radiol. **14**, 49–53 (2007).

46. Sund, P., Herrmann, C., Tingberg, A., Kheddache, S., Månsson, L. G., Almén, A. and Mattsson, S. *Comparison of two methods for evaluating image quality of chest radiographs.* Proc. SPIE **3981**, 251–257 (2000).

47. Tingberg, A., Herrmann, C., Lanhede, B., Almén, A., Besjakov, J., Mattsson, S., Sund, P., Kheddache, S. and Månsson, L. G. *Comparison of two methods for evaluation of the image quality of lumbar spine radiographs.* Radiat. Prot. Dosimetry **90**, 165–168 (2000).

48. Sandborg, M., McVey, G., Dance, D. R. and Alm Carlsson, G. *Comparison of model predictions of image quality with results of clinical trials in chest and lumbar spine screen-film imaging.* Radiat. Prot. Dosimetry **90**, 173–176 (2000).

49. Sandborg, M. *et al. Demonstration of correlations between clinical and physical image quality measures in chest and lumbar spine screen-film radiography.* Br. J. Radiol. **74**, 520–528 (2001).

50. Sandborg, M., Tingberg, A., Ullman, G., Dance, D. R. and Alm Carlson, G. *Comparison of clinical and physical measures of image quality in chest and pelvis computed radiography at different tube voltages.* Med. Phys. **33**, 4169–4175 (2006).

51. Tingberg, A., Båth, M., Håkansson, M., Medin, J., Besjakov, J., Sandborg, M., Alm-Carlsson, G., Mattsson, S. and Månsson, L. G. *Evaluation of image quality of lumbar spine images: a comparison between FFE and VGA.* Radiat. Prot. Dosimetry **114**, 53–61 (2005).

52. Månsson, L.G. *Evaluation of radiographic procedures—investigations related to chest imaging.* Thesis, Göteborg, Göteborg University (1994).

53. Månsson, L. G. *Methods for the evaluation of image quality: a review.* Radiat. Prot. Dosimetry **90**, 89–99 (2000).

54. Tingberg, A. *Quantifying the quality of medical X-ray images—an evaluation based on normal anatomy for lumbar spine and chest images.* Thesis, Lund, Lund University (2000).

55. Båth, M. *Imaging properties of digital radiographic systems—development, application and assessment of evaluation methods based on linear-systems theory.* Thesis, Göteborg, Göteborg University (2003).

56. Båth, M. and Månsson, L. G. *Visual grading characteristics (VGC) analysis: a non-parametric rank-invariant statistical method for image quality evaluation.* Br. J. Radiol. **80**, 169–176 (2007).

57. Chakraborty, D. P. *Problems with the differential receiver operating characteristic (DROC) method.* Proc. SPIE **5372**, 138–143 (2004).

58. Vucich, J., Goodenough, D. J., Lewicki, A., Briefel, E. and Weaver, K.E. *Use of anatomical criteria in screen/film selection for portable chest x-ray procedures.* In: Optimization of Chest Radiography. HHS Publication 80-8124. Cameron, J. Ed. (Rockville, MD: FDA) pp. 237–248 (1980).

59. Kundel, H. L. *Images, image quality and observer performance.* Radiology **132**, 265–271 (1979).

60. Commission of the European Communities. *European guidelines on quality criteria for diagnostic radiographic images.* Report EUR 16260 EN (Luxembourg: Office for official publications of the European Communities) (1996).

61. Commission of the European Communities. *European guidelines on quality criteria for diagnostic radiographic images in paediatrics.* Report EUR 16261 EN (Luxembourg: Office for official publications of the European Communities) (1996).

62. Commission of the European Communities. *European guidelines on quality criteria for computed tomography.* Report EUR 16262 EN (Luxembourg: Office for official publications of the European Communities) (1996).

63. Geijer, H., Verdonck, B., Beckman, K. -W., Andersson, T. and Persliden, J. *Digital radiography of scoliosis with a scanning method: initial evaluation.* Radiology **218**, 402–410 (2001).

64. Smedby, Ö. and Fredriksson, M. *Visual grading regression–analysing data from visual grading experiments with regression models.* Br. J. Radiol. (accepted for publication 2009) [DOI: 10.1259/bjr/35254923]

65. Altman, D. G. *Practical Statistics For Medical Research* (London: Chapman&Hall) (1991).

66. Stevens, S. S. *On the theory of scales of measurement.* Science **103**, 677–680 (1946).

67. Tapiovaara, M. J. and Sandborg, M. *How should low-contrast detail detectability be measured in fluoroscopy?* Med. Phys. **31**, 2564–2576 (2004).

68. Thilander-Klang, A., Ledenius, K., Hansson, J., Sund, P. and Båth, M. *Evaluation of subjective assessment of the low-contrast visibility in constancy control of computed tomography.* Radiat. Prot. Dosim. **139**, 449–454 (2010).

69. McKee, M. *Not everything that counts can be counted; not everything that can be counted counts.* BMJ **328**, 153 (2004).