CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Julkari

STUK

STUK-A219 / SEPTEMBER 2006

RELATIONSHIPS BETWEEN PHYSICAL MEASUREMENTS AND USER EVALUATION OF IMAGE QUALITY IN MEDICAL RADIOLOGY – A REVIEW

M. Tapiovaara

STUK • **SÄTEILYTURVAKESKUS**
STRÅLSÄKERHETSCENTRALEN
RADIATION AND NUCLEAR SAFETY AUTHORITY

Osoite/Address • Laippatie 4, 00880 Helsinki
Postiosoite/Postal address • PL/P.O.Box 14, FI-00881 Helsinki, FINLAND
Puh./Tel. +358 9 759 881 • Fax +358 9 759 88 500 • www.stuk.fi

The conclusions presented in the STUK report series are those of the authors and do not necessarily represent the official position of STUK.

ISBN 952-478-160-3 (print)
ISBN 952-478-161-1 (pdf)
ISSN 0781-1705

Dark Oy, Vantaa/Finland, 2006

Sold by:
STUK – Radiation and Nuclear Safety Authority
P.O.Box 14, FI-00881 Helsinki, Finland
Phone: +358 9 759 881
Fax: +358 9 7598 8500

TAPIOVAARA Markku. Relationships between Physical Measurements and User Evaluation of Image Quality in Medical Radiology – a Review. STUK-A219. Helsinki 2006, 62 pp.

Keywords: medical imaging, x-ray imaging, image quality, test methods, optimisation, quality assurance, SENTINEL

Abstract

There are many tasks in radiology departments which involve assessment of image quality. Equipment purchasing is partly based on performance specifications, acceptance testing verifies that the system fulfils the specified performance criteria, constancy testing attempts to notice any changes in the imaging system, clinical testing concentrates on the fulfilment of clinical needs, and optimisation processes attempt to find best ways to use the imaging system for clinical purposes. These different tasks are best performed by different assessment methods and the outcome is often referred to as technical (or physical) image quality or clinical image quality, according to the method used. Although establishing the link between physical image quality measures and clinical utility has been pursued for decades, the relationship between the results of physical measurements, phantom evaluations and clinical performance is not fully understood. This report shortly discusses various assessment methods, points out factors that may influence the interpretation of their results, and reviews recent studies that have explored the relationships between them.

TAPIOVAARA Markku. Relationships between Physical Measurements and User Evaluation of Image Quality in Medical Radiology – a Review. STUK-A219. Helsinki 2006, 62 s.

Avainsanat: lääketieteellinen kuvantaminen, röntgenkuvantaminen, kuvanlaatu, testausmenetelmät, optimointi, laadunvarmistus, SENTINEL

Tiivistelmä

Röntgenosastoilla on monia tehtäviä, joissa tarvitaan kuvanlaadun arviointia. Laitteiden hankinta perustuu osaltaan suorituskykyspesifikaatioihin, vastaanottotarkastuksessa varmennetaan näiden spesifikaatioiden täyttymistä, vakioisuusmittauksin pyritään mahdollisimman varhaisessa vaiheessa huomaamaan kuvantamislaitteessa tapahtuneet muutokset, kliininen testaus keskittyy lääketieteellisten tarpeiden täyttymiseen ja optimoinnilla yritetään löytää laitteen parhaat käyttötavat käytännön työhön. Näihin erilaisiin tarkoituksiin soveltuvat parhaiten toisistaan poikkeavat kuvanlaadun arviointimenetelmät. Niiden perusteella mitattua tai arvioitua kuvanlaatua kutsutaan usein tekniseksi (tai fysikaaliseksi) kuvanlaaduksi tai kliiniseksi kuvanlaaduksi arvioinnissa käytetyn menetelmän mukaan. Fysikaalisen kuvanlaadun mittaukseen käytettävien suureiden ja kliinisen käytettävyyden välistä yhteyttä on yritetty selvittää jo kauan, mutta fysikaalisten mittausten, erilaisista kappaleista otettujen testikuvien ja kliinisen suorituskyvyn välistä yhteyttä ei vieläkään kunnolla tunneta. Tässä raportissa käsitellään lyhyesti erilaisia kuvanlaadun arviointimenetelmiä, tarkastellaan tekijöitä, jotka vaikuttavat niiden avulla saatavien tulosten tulkintaan ja tehdään kirjallisuuskatsaus viimeaikaisiin julkaisuihin, joissa on tutkittu eri arviointimenetelmien antamien tulosten välistä yhteyttä.

Contents

1 Introduction

There are many tasks in radiology departments which involve assessment of image quality. Equipment purchasing is partly based on performance specifications, acceptance testing verifies that the system fulfils the specified performance criteria, constancy testing attempts to notice any changes in the imaging system, clinical testing concentrates on the fulfilment of clinical needs, and optimisation processes attempt to find best ways to use the imaging system for clinical purposes. These different tasks are best performed by different assessment methods and the outcome is often referred to as technical (or physical) image quality or clinical image quality, according to the method used. Sometimes (e.g., ICRU 2003), the term image quality is devoted mainly to the technical aspects of the image: primarily contrast, sharpness and noise.

Even when one speaks of clinical image quality, the actual point of view and the definition of image quality are often left unspecified. Most often one just refers to a subjective judgement of quality in the clinical radiographs and/or fluoroscopic image. If the opinion is just based on an impression of quality, the usefulness of the assessment may be questionable (Vucich 1979, Barrett and Myers 2004, Månsson 2000). When judged by task-based criteria – for example by the opinion of the radiologist relating to his ability to perceive certain anatomical details or features in the image and his/her confidence on the perception of these details, the assessment is more relevant. However, even then the outcome may be uncertain: the subjectivity of the evaluation leaves notable variability and bias in the results (ICRU 1996, Krupinski 2000, Barrett and Myers 2004).

In medical radiology, images are used to diagnose patients (diagnostic radiology) or to treat them (interventional radiology). Therefore, image quality in radiology is most meaningfully defined through the usefulness of the images in accomplishing these tasks. The present consensus for defining diagnostic image quality is based on such a task-based approach (ICRU 1996, Barrett and Myers 2004). This approach differs from subjective assessment by setting a specified task for the image and actually measuring the performance achieved. This controlling of the outcome is not done in a subjective assessment, and often even the task is left unspecified.

Because of the limited value of subjective assessment and the difficulty of performance-based measurement of image quality from clinical images, other, more precise and analytical means are needed for such purposes as equipment design, performance specification and acceptance and constancy testing (Tapiovaara 2005). The methods used include the measurement of the physical characteristics of the images (and the imaging system) and/or evaluation of image quality from phantom images. These various methods have been reviewed,

e.g., by ICRU (1996), Martin et al. (1999a) and Dobbins (2000) for radiology in general, and by Bosmans et al. (2005) for digital mammography in particular. Although establishing the link between physical image quality measures and clinical utility has been pursued for decades, the relationship between the results of physical measurements, phantom evaluations and clinical performance is not fully understood (Wagner 1977 and 1987, Wagner et al. 2001).

The purpose of this paper is to shortly discuss various assessment methods, to point out factors that may influence the interpretation of their results, and to review recent studies that have explored the relationships between them. Further references can be found in the cited papers. The work has been limited to projection imaging. Computed tomography (CT) has not been considered, in spite of its importance in radiology, as it is outwith the remit of the SENTINEL project, within which this review was made.

2 Hypothetical relationship between diagnostic performance and physical image quality

The relationship between diagnostic performance and physical image quality (mainly contrast, sharpness and noise) has been discussed, for example, in ICRU (1986) where it was depicted by a simplified graph, similar to that shown in Fig. 1. If the level of physical image quality is extremely low, the image can provide no information for the diagnosis and diagnostic accuracy is unaffected by the x-ray examination. When the physical image quality improves, important radiological patterns become recognizable and diagnostic performance improves. Beyond a certain level of physical image quality – where all the important features are already visible and no additional image information that would be useful for the radiologist can be brought in the image – the performance will saturate. On the other hand, there are also examples with improved image quality decreasing the performance of radiologists. For example, this could happen in detecting small nodules or pulmonary masses in chest radiographs, because of the increased normal background visibility and complexity (ICRU 2003). In digital imaging this is not contradictory to the curve in Fig. 1, however: it is more a question of image processing and display than a question of hardware performance. In digital imaging, sharp images can always be processed to be unsharp, if it is deemed useful.

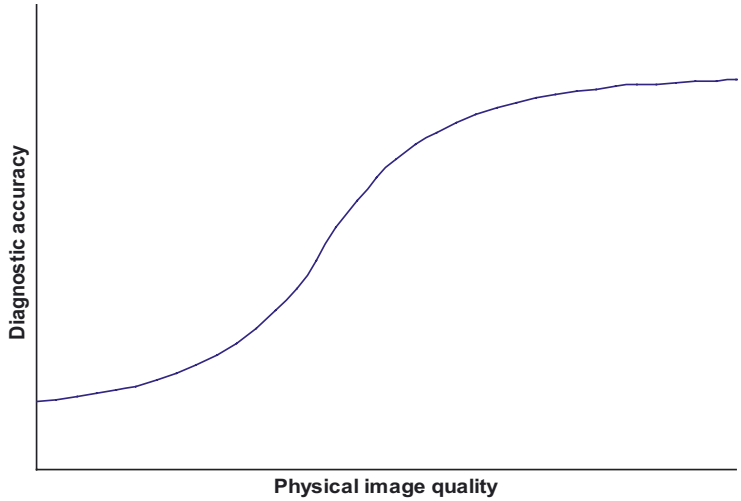

Figure 1. Simplified qualitative relationship between physical image quality and diagnostic performance (Adapted from ICRU 1986).

Diagnostic performance, and the curve depicted in Fig. 1, depend of course also on other factors than physical image quality. Among such factors are for example 1) the diagnosed disease and the nature, subtlety and specificity of the radiological features it causes, 2) the anatomical structures and their variability in both the positive and the negative patients, 3) the prior information of the patient given to the radiologist, and 4) the skill, medical knowledge and experience of the radiologist (Nodine and Mello-Thoms 2000). Of course, such curves may then differ widely if they are drawn to depict the performance of a given radiologist, a specific group of radiologists or the profession in general; they would also differ if drawn for a given patient, a group of patients with certain examination indications or the whole ensemble of patients.

Clinical performance does not improve along with physical image quality when the operating point is already on the saturated region of the curve. However, it is not always known where one is working on such a curve and whether clinical performance would be affected by moderate changes in physically or technically assessed quality. Similar considerations apply also to the relationship between subjectively assessed image quality and clinical performance: it is not always granted that diagnostic performance would be better when the images give an impression of improved quality (Leitz et al. 1993, Krupinski 2000, Tingberg et al. 2005a). If it would be known that all features of interest are confidently seen in all the evaluated images, the value of saying that the features are "better" visible in some of them would be of questionable or no significance. On the other hand, such sufficiency of certain features may be seldom granted, and their better visibility is likely to indicate that the image would also be able to reveal fainter features that cannot be seen in the other images. Then the improved system could also be more effective clinically.

Striving for the best possible physical or technical image quality is not optimal in medical radiology, because image quality is intimately related to the radiation dose to the patient and this would result to unacceptably high doses. In film-screen imaging this would correspond to choosing excessively slow screens and film and in digital imaging working at an unnecessary high image receptor air kerma level. Therefore, the optimum is presently often defined such that one obtains an image quality that is sufficient for the intended purpose, with the lowest possible dose to the patient. For example, when bone angle measurements are made the dose to the patient can in digital x-ray imaging be reduced to a small fraction of what is needed in film-based imaging (Sanfridsson et al. 1999) as this examination would not benefit from improving the visualisation of faint details. Similarly, Kotre et al (2004) demonstrated that low dose fluoroscopy is acceptable in cardiac pacing and electrophysiology, where the clinical requirement is that the high contrast wires are adequately distinguished. The idea of being

content to work with sufficient image quality should not, however, be taken too literally to mean that one should work too close to the limit of perceiving important details. The risks from an inadequate examination could then easily exceed the health benefits from a reduced dose (ICRP 1996, Martin et al. 1999b). Although one would work on the shoulder region of the average-patient curve, where benefits to the patients from improved physical image quality would be small on the average, the benefits to some individual patients could be large. Defining the actual optimal operating point on such a graph would require a detailed comparison of costs, health detriments from the patients' radiation dose and expected health benefits from the examination. Suggestions for choosing a sufficient dose level (corresponding to high, medium or low image quality) for different examination indications and different types of digital x-ray equipment have been given by Busch (2004) and discussed by Busch and Faulkner (2005).

Båth et al. (2005a) have discussed optimisation strategies for digital x-ray imaging. They note that digital imaging has provided new freedom for choosing the exposure and image contrast, and suggest that imaging technique optimisation should: a) include the anatomical background during the optimisation, b) be performed at a constant effective dose, and c) separate the image display stage from the image collection stage. During the actual optimisation, one should first determine the optimal settings of technique parameters, then determine the optimal processing and display settings, and finally determine the optimal (necessary) dose level using the best combination of imaging technique and processing/display parameters. In this paper (Båth et al. 2005a) they note that traditional physical optimisation has been based on determining detectability of faint details in uniform phantoms, although growing evidence shows that in digital radiography at presently typical dose levels detection is often more limited by the anatomical background than by quantum noise.

3 Assessment of clinical image quality

3.1 Controlled patient studies (ROC and AFC)

As already stated above, the purpose of medical images is to provide information on the health state of the patient or to enable treating the patient. Therefore, in principle, image quality should optimally be measured by methods that address clinical performance. The measurement of receiver operating characteristics (ROC) is presently considered being the best method of quantifying and reporting diagnostic performance, and the ROC curve can then also be considered as a good descriptor of clinical image quality.

ROC analysis is based on the fact that the diagnostician can adapt to different "critical confidence levels" for calling an image normal or abnormal. Therefore, simple measurement of the sensitivity and specificity of the diagnosis is not sufficient, but the sensitivity and specificity pairs need be evaluated for various critical confidence levels: these sensitivity and specificity pairs then define the ROC graph. For references on the ROC methodology see, e.g., ICRU (1996) or Metz (2000). There are also variants of the ROC methodology that take the locations of the lesions into account and thereby increase the statistical power of the evaluation (Chakraborty 2000, Månsson 2000).

If sensitivity and specificity data at different critical confidence levels are not needed, the results of ROC experiments are often summarized in terms of a single number, such as the area under the ROC curve (A_z) or the detectability index d_a. The same measure can also be obtained by two-alternative forced-choice (2-AFC) (or multiple alternative forced choice (M-AFC)) methods (ICRU 1996, Burgess 1995, Metz 2000), where the observer is presented two images (or a larger number of images in M-AFC) of which only one is actually abnormal and the task of the observer is to choose the abnormal image. The result of the experiment is the probability of the correct response. It can be transformed to the detectability index d', which can be interpreted as the decision stage signal-to-noise ratio of the observer for the signal detection task in question (in practice, d' and d_a are same quantities, see e.g. ICRU 1996, Burgess 1995 or Metz 2000). AFC experiments are simpler and quicker to perform than complete ROC studies, but require more images to achieve equal statistical power (Burgess 1995).

All these methods require a large number of patient images and that the true health state of each patient is known. Further, the patient cases need to be so difficult that errors will be made in the image interpretation. These requirements make the measurement of clinical performance difficult in practice. The methods are most easily used for comparing different imaging modalities

or imaging techniques with images that are taken from the same patients: comparison of patient image-based ROC results between different clinics, for example, is not reasonable unless one can be sure that the patient material is similar in all the clinics of the study. Somewhat similarly, the results depend on the skills of the participating radiologists. Also, if a new imaging method with vastly different image characteristics is being compared to a commonly used old imaging method, it must be considered whether the results from the novel system could have suffered from the unusual appearance of its images compared to the old modality, with which the radiologists have worked a long time. Radiological expertise is much based on image-reading experience (Nodine and Mello-Thoms 2000) thus the results with the new system might unfairly suffer from the lack of experience.

The ROC method suffers from its inability to deal simultaneously with multiple diagnostic alternatives in a completely adequate way. This may be particularly cumbersome in chest radiology where a wide range of image features are of interest and a large number of different pathological conditions can be found (ICRU 2003). Metz (2000) also points out a potential problem with the difference in actually-positive image prevalence between ROC experiments (typically around 50%) and the clinical situation (of the order of 0.5% in mammography screening): the extent to which human strategy and performance may change at very low disease prevalence is not known. A further concern of the clinical relevance of ROC analyses is the need of the ROC experiment to focus on very subtle lesions and the fact that much of the ROC curve depicts false positive diagnosis probabilities that are not acceptable in most radiological practice (ICRU 2003).

A recent example of ROC-based image quality comparisons is the paper of Eisenhuber et al (2003) where the aim was to measure the visibility of low-contrast catheter fragments in bedside chest radiographs obtained with film-screen (400-speed) and storage phosphor based image receptors at various dose levels (corresponding to 200-, 400- and 800-speed systems). The problem of patient material variability was avoided by imaging the same patients in an intensive care unit with the various systems on consecutive days, and the problem of establishing the truth was not present because the experimenters knew the actual position of each catheter fragment. It was found that the detectability of the catheters was significantly better in the CR images of all dose levels than in the film-screen image. The detectability was significantly poorer in the 800-speed CR image than in the 200- and 400-speed images; no significant difference was found between the latter two. Another recent ROC experiment compared observer performance in detection of wrist fractures with a common PC display and a dedicated diagnostic display (Doyle et al. 2005): surprisingly, no differences in

performance were seen in spite of the much more modest technical specifications of the PC display.

ROC or AFC studies need not necessarily be made based on actual patient images. Instead, one can use images of a suitable phantom in which a relevant signal detail can be added or removed. Then the experimenter has no difficulty in knowing the actual truth state in any image. In principle it is possible to use such phantoms and details that the images represent most of the physical aspects of the clinical images, but it is difficult to obtain similar variability in the images that would result from the variability in real patients' anatomy (background) and features of pathology (signal). In practice, the phantoms are often homogeneous attenuating and scattering blocks or anatomical phantoms in which the observer knows (or may get familiar with) the background structure. Also, usually the observer knows the shape and location of the signal detail, which is very often a simple object like a circular or square disk. While such studies can be totally adequate to take into account all the physical factors affecting the visibility of the detail in the test setting, it is often difficult to know how well the results would agree with data that would be obtained with actual patient imaging. However, it seems plausible to expect that the results of such phantom studies would be closely related to clinical performance if all the factors that are important for the detectability of the diagnostic details in the patients are taken into account in the phantom study. This is not always given the emphasis that it would require in planning the phantom study. For example, experiments trying to optimise the radiation quality for detecting a certain detail in the phantom may result in incorrect conclusions if the energy dependence of the contrast of the test detail differs from that of the relevant signals in actual patients. Therefore, for example, it is not reasonable to use test details made of aluminium or plastic material for optimising examination conditions when iodine contrast material in the patient needs be seen. Similarly, the conclusions of a phantom study are doubtful if the experiments are made using a homogeneous phantom, while the detectability of the actual detail in the clinical images would be mainly restricted by the varying and non-homogeneous background in the patient images – which the phantom study does not include. Recently, an example of poorly simulated anatomical details in attempting to find the optimal x-ray tube voltage for digital radiography was pointed out by Venema et al. (2005). They criticized a paper where anatomical soft tissue details in chest radiography were mimicked by using organic material soaked with iodine contrast medium.

Realistic, variable radiological backgrounds can be obtained by using images of verified normal patients (Metz 2000), and artificially adding signals that represent pathological changes in the images. The main difficulty in such an approach is that it is imperative that the computationally added signal

images are similar to images that would have been obtained if the lesion had actually been in the patient (Sorenson et al. 1980). This requires at a least thorough consideration of the factors influencing the contrast and sharpness of the signal details, and therefore, among other things, accurate and realistic modelling of the lesions and the imaging system. For the latter of these factors, the modulation transfer function (MTF) and the signal transfer of the imaging system need to be measured. This is not a trivial task: e.g., consideration needs to be given to unsharpness resulting from the actual shape and size of the x-ray tube focal spot, the unsharpness from signal spread in the image receptor, and even motion unsharpness. One practical alternative, properly taking into account many of the factors, might be to take images of actual detail objects embedded in homogeneous, patient simulating phantoms and to extract the signals from these phantom images. Using mathematically defined details and modifying them by factors related to the radiation contrast of the detail and the sharpness and sensitometric properties of the imaging system may be more challenging in practice.

Routine conduction of controlled observer studies (ROC, 2-AFC or M-AFC) is not practical because of the large number of images and the significant observer time and overall amount of work that the evaluation requires. Therefore, simpler methods to assess clinical image quality are needed. However, in principle, controlled observer studies with clinical images can be considered being the gold standard against which other assessment methods should be validated. In practice, the results of ROC studies are specific for the type of signal/background combination studied and a generalised measure of clinical image quality is difficult to obtain (Båth et al. 2005a). Therefore, Båth et al. (2005a) suggest that subjective evaluation methods should be more suitable than ROC methods for studies attempting to optimise imaging technique parameters.

3.2 Subjective assessment of clinical image quality

As noted in the ICRU report (1996) “subjective” refers to individual human judgement, so all methods employing human observers are subjective. However, the above mentioned methods (ROC and AFC) can, in a broader sense, be thought to provide objective measures of image quality – they provide numerical results that are not dependent on the opinion of the observer, but on the performance that the observer achieves. Here, subjective methods refer to methods which rely on the observers’ preference and where the correctness of their opinion is not controlled (and cannot even be defined).

Subjective evaluation of image quality is of course useful – and often it is also the only practical alternative for assessing clinical image quality (ICRU

1996, Dobbins 2000): some kind of verification of the clinical acceptability of patient images is necessary (Dobbins 2000, ICRU 2003). However, it should be understood that there are many possible sources of bias and uncertainty in these evaluations. The sources of bias range from preference for the aesthetically pleasing to prejudice against the unfamiliar (ICRU 1996); according to Krupinski (2000), basing her view on a number of papers, the unspoken assumption of better performance in image preference studies translating to better clinical performance may not always be true.

There is also notable variability in subjective assessment results. The sources of variability include between-observer variability, within-observer variability and case variability (variability of individual patient images). These factors will limit the use of these methods only to finding large differences in image quality (ICRU 1996, Krupinski 2000), and may make comparisons of results obtained in different clinics virtually impossible unless image qualities clearly below or above standard clinical practice are involved (Verdun et al. 1993). These methods are more easily used for comparisons of, e.g., various imaging techniques within an x-ray department (e.g., Almén et al. 2000 and 2004, Tingberg et al. 2004, Tingberg and Sjöström 2005).

Some reduction of variability and bias may be obtained by giving the observers instructions to focus their attention upon specific features in the image (Thornbury et al. 1977, Vucich 1979, ICRU 1996, Dobbins 2000). Such an approach has been adopted in the image criteria which the Commission of European Communities (CEC) and others have published (Commission of the European Communities 1996a, 1996b and 2000, European CT Study Group 2004, Busch 2004, Bernardi et al. 2005a); these criteria are intended to be used to characterise acceptable radiographs of normal, healthy patients. The use of anatomical landmarks in assessing the quality of radiographs has been discussed in more detail by Vucich (1979).

Methods for subjective quantitative assessment of clinical images have also been discussed by Dobbins (2000). These tests typically fall into two categories: side-by-side comparative studies and independent assessments on an absolute scale of merit. These are sometimes referred to as relative or absolute visual grading analysis (VGA), respectively: most often, however, the term VGA refers only to the relative method using a reference radiograph. The special case of fulfilling the European image criteria has been referred to as image criteria scoring (ICS). For references on such studies see e.g. Börjesson et al. (2005) and Tingberg et al. (2004 and 2005a). In the study of Tingberg et al. (2004) it was found that in repeated readings of the same radiographs the radiologists changed their opinion on the visibility of a structure in about one out of four times, on the average. It is noted that the relative VGA method, where the same reference

radiograph is compared against each other image, is able to provide much more consistent results than the absolute VGA method (Tingberg et al. 2000).

The relevance of the image criteria for diagnostic performance has not been strictly proved, but they are based on the professional judgement of a group of radiologists. It is not self-evident that the fulfilment of all or some of the criteria is necessary or sufficient for accurate diagnostics – either in general or for specified examination indications in particular. There are few publications ascertaining this relationship by using the traditional ROC approach (which, of course, would be very tedious and still only applicable to the specific examination and pathologic condition chosen). An example where a strong link between performance and subjective evaluation was found is the study on the diagnostics of craniosynostosis by Pilgram et al. (1989). Comparisons of subjective evaluation and free-response forced error (FFE) experiments with simulated pathology have provided mixed results. In an earlier experiment where images were manipulated to correspond to three different screen-film combinations the subjective evaluation and FFE results were in good agreement (Tingberg et al. 2000), whereas in a later paper Tingberg et al. (2005a) found that images with low noise were preferred in the VGA study although the noise level did not affect detectability in the FFE experiment. (One should note however that the result of the FFE experiment relates only to the type of detail used in the study.) However, it seems clear that also the visibility of normal anatomy is of great diagnostic significance to radiologists. Demonstration of expected structures and patterns is reassuring to the interpreting physician and allows pathology to be excluded more reliably. The trade-offs that arise in the demonstration of normal anatomy are largely the same that arise when depiction of pathology is considered (ICRU 2003). Vañó et al. (1995) have analysed the fulfilment of the image criteria in both accepted and rejected chest radiographs, and found that rejected radiographs failed to meet the CEC image criteria much more often than accepted radiographs did.

It seems evident that the European image criteria are not always unequivocal and radiologists may interpret them differently (Maccia et al. 1997, Almén et al. 2000, Tingberg et al. 2004 and 2005b); modification of the European image criteria has been suggested in several papers (e.g., Vañó et al. 1995, Almén et al. 2004, Hemdal et al. 2005). If the criteria are used for comparing different imaging systems, inter-observer variability is often larger than the difference between the systems (Tingberg 2005b). Tingberg et al. (2005b) attempted to reduce the inter-observer variability in image quality assessments by masking the images to show only the parts of the images that were necessary for the image criteria evaluation. However, it was found that this increased the variability, against expectations, and therefore they recommended not using the masking technique.

Also in the trial evaluation of the quality criteria for cardiac angiographic images (Bernardi et al. 2005b) notable variability was seen in the assessment of many patient studies, although the variability was small for the majority of them (which obtained nearly the maximum score in the evaluation). It also seems to be a common finding that overall image quality is judged sufficient even if not all CEC image criteria are fulfilled (Vañó et al. 1995, Maccia et al. 1997, Martin et al. 1999a, Offiah and Hall 2003) and it has been suggested that the importance of different image criteria should be identified and scored. Otherwise, not fulfilling a single but important criterion could easily result to wrongly accepting an inadequate radiograph (Vañó et al. 1998).

As already mentioned above, the CEC image criteria have been developed as a means to demonstrate acceptable image quality in film-screen radiography. They cannot be easily used to alert for excessively high (and therefore non-optimal) image quality. This was also pointed out in the discussion notes following Vucich's paper (1979), where the merit of "optimally visualised" over "adequately visualised" was questioned.

A recent example of using subjective evaluation methods is the paper of Uffmann et al. (2005) where chest images were taken with a CsI-based flat panel system from the same patients at three different x-ray tube voltages but with equal effective dose to the patient at each tube voltage. The tube voltage dependence of the visibility of various anatomic structures was evaluated by both absolute and relative visual grading analysis. In the paper these were referred to as visibility of anatomic structures and preference of acquisition technique, respectively. The radiologists in the study ranked images taken at 90 kV significantly superior to images taken at 121 kV and 150 kV. Geijer and Persliden (2005) have compared lumbar spine radiographs of an anthropomorphic phantom obtained at various x-ray tube voltages and equal effective dose with a flat-panel detector. They concluded that the examination would be optimised with lower than their present clinical x-ray tube voltage (77 kV); best overall evaluation was obtained at 52 kV. They found similar results also in contrast-detail measurements with a CDRAD 2.0 phantom placed in the middle of a 20 cm thick PMMA slab phantom.

Sometimes, subjective comparisons of different imaging systems or techniques may be performed on patients with pathological changes by imaging the same patient with various methods. This possibility is restricted to examinations where the dose to the patient is low and even then only a small number of imaging conditions can be considered. A practical difficulty with this method is that it is hard to find patients where the pathological condition is only marginally detectable so that differences in diagnostic performance could be reliably inferred. Of course, if a sufficient number of patients could be found

it would be better to make the evaluation using the ROC methodology. For references of this approach in chest radiology see ICRU (2003).

Many of the difficulties of the patient image-based evaluations can be avoided by using mathematically simulated pathological details or phantoms with added details instead of actual patients (section 4.2). Båth et al. (2005b) have suggested an approach to optimise the dose level in radiography by using computational methods to manipulate clinical images to correspond to images being taken with a lower dose. Also Monte Carlo models with voxel phantoms have been suggested for image optimisation studies (e.g., Sandborg et al. 2001b, McVey et al. 2003 and Ullman et al. 2006). More often, the Monte Carlo method has been used for similar purposes with simple phantom models.

4 Assessment of physical or technical image quality

4.1 Physical approach

Imaging is basically a process consisting of two distinct stages: image recording and image display (Wagner 1983, ICRU 1996). This division is especially important in digital imaging, where these stages are clearly separate. In digital imaging the image recording stage (or the image data stage) determines the information that has been captured in the image data and can be analysed in terms of the pixel values. Performing actual physical measurements of the display stage is cumbersome and its evaluation is for the most part done mainly visually. In screen-film imaging the analysis is made using a microdensitometer and sensitometry of the film response. Below, mainly the assessment of the image data stage in digital imaging is considered. The post-processing and display stage should be designed and arranged so that a human observer can perceive the image information efficiently. In spite of the importance of the display stage it has been given here only a minimal consideration.

The most fundamental quality-related factors in imaging are contrast, sharpness and noise. Clearly defined quantitative measures exist for these factors. For example, the sharpness of imaging systems can be described by the modulation transfer function (MTF), contrast gain by the characteristic curve of the system, and noise by the noise power spectrum (NPS, also referred to as the Wiener spectrum) (for these factors see, e.g., Dainty and Shaw 1974, Barrett and Swindell 1981, Cunningham 2000, Dobbins 2000 and Barrett and Myers 2004. For recent intercomparisons of different measurement methods see Neitzel et al. 2004, Dobbins et al. 2006 and Samei et al. 2006).

It is pointed out that some often-used simple descriptors of the above factors are not sufficient to characterise them. For example, the measurement of resolution by a line pair test object attempts to express sharpness by finding the maximum frequency that is seen in the image. It does not describe signal transfer at lower frequencies, which are in practice responsible for giving the diagnostic information (Cowen et al. 1997). Somewhat similarly, a measurement of noise by just obtaining the variance of pixel values is not sufficient. The noise in medical images is typically not independent from one pixel to another, and therefore the noise power spectrum (or, equivalently, the noise autocovariance function in the spatial domain) is required for full characterisation of noise. Even these descriptors are not sufficient if the noise is not stationary. Pixel variance is

a useful quantity, of course, if it is known that the shape of the NPS is constant in all cases being compared. Using pixel variance may therefore be reasonable and sufficient in constancy testing, but it is absolutely not an appropriate quantity for comparing the noise in imaging systems with different NPS shapes.

Digital images can be altered easily and, therefore, the MTF and NPS are not of equal importance alone by themselves as they are in film-based imaging: they must be combined together to express the quality of imaging (e.g., Moy 2000). This combining is based on statistical decision theory (SDT) (Wagner and Brown 1985, ICRU 1996, Beutel et al. 2000, Myers 2000, Barrett and Myers 2004) where image quality is evaluated by the observer's performance in a specified imaging task. The central concept in this approach is the ideal observer, which uses all available image information and all available prior information about the task in an optimal way to make its decision. The ideal observer achieves the best performance that is possible (statistically, in a large number of trials). Therefore, its performance in a specified imaging task with specified prior information is a measure of the amount of image information that is relevant to that task. This approach differs from the controlled human observer studies (section 3.1) basically only by not using human observers but the best possible observer. One should note that the ideal observer's performance cannot be improved by image post-processing without actually bringing new information in the image (although the opposite is of course possible). Image post-processing may be very useful for a human observer, but cannot improve human performance above the level that the ideal observer would achieve in the non-processed images: post-processing can be seen just as a means to try to improve the interface between the image data and the human observer.

To define the ideal observer, one needs to first define optimality of detection performance. Various definitions exist: e.g., minimum error probability, Bayes criterion for minimum expected cost and the Neyman-Pearson criterion for maximising detection probability for a given probability of false alarm. It can be shown that all these definitions of optimal detection can be accomplished by basing the decision on the likelihood ratio of the data (Whalen 1971). The likelihood ratio may be extremely complicated in most detection tasks but it is well-known and simple in some special cases; this allows the measurement of the ideal observer's performance in such simple signal detection tasks. The best understood and most widely considered imaging task in this approach is the so-called signal-known-exactly/background-known-exactly (SKE/BKE) detection task with additive (i.e., signal-independent), normally distributed, stationary noise and a linear (or linearizable) imaging system. The ideal observer's performance in this task can be summarised by the ideal-observer's signal-to-noise ratio (SNR_{ideal}) which for the continuous (analogue) imaging case can be written as

$$SNR^2_{ideal} = \iint \frac{G^2 MTF^2(f_x,f_y)}{W(f_x,f_y)} \cdot \Delta S^2(f_x,f_y)\, df_x df_y \ . \quad (1)$$

This depicts the ideal observer's SNR at the decision stage and is an equal concept to the detectability index d' (and d_a) mentioned earlier. Here $\Delta S(f_x,f_y)$ is the Fourier transform of $\Delta s(x,y)$, the signal to be detected, and G takes care of treating the signal and noise in the same units. $MTF(f_x,f_y)$ and $W(f_x,f_y)$ are the modulation transfer function of the system and the noise power spectrum, respectively. This equation shows how the system's large area signal transfer, sharpness and noise combine to an overall quantity expressing image quality for a particular task, the detection of the signal $\Delta s(x,y)$. When the signal and noise are expressed in relative (contrast) units the first factor in the integrand is called the noise equivalent quanta (NEQ) and expresses the effective fluence that the image is worth at each spatial frequency. When NEQ is divided by the actual fluence used for forming the image, one obtains the detective quantum efficiency (DQE), which expresses the efficiency of information transfer in the imaging system. Presently, imaging systems are often specified by their DQE for specified x-ray spectra (IEC 2003). In principle one can then calculate the SNR_{ideal} for signal details of various sizes and shapes and different image receptor dose levels (provided that the imaging system is quantum noise limited in the dose range being considered). A difficulty in obtaining SNR_{ideal} by this approach, in addition to the measurement of G, $MTF(f_x,f_y)$ and $W(f_x,f_y)$, is that one needs to accurately model the radiation contrast of the detail. Also, DQE measurements are usually intended to characterise only the image receptor part of the imaging system. The influence of other factors, e.g., scattered radiation and the finite size of the x-ray tube focal spot are minimised in the measurements.

The SDT approach is used also in digital imaging. For a precise treatment using the vector notation of discrete imaging and a thorough discussion of the concepts, see Barrett and Myers (2004). It is noted, however, that digital imaging systems do not generally fulfil the assumption of shift-invariance (which is needed for applying the MTF concept) and this makes the applying of the approach cumbersome. This has been circumvented by introducing the concept of presampling MTF (Fujita et al. 1989, Dobbins 2000). When SNR_{ideal} is calculated, the signal needs be first filtered with the presampling MTF and then sampled according to the pixel pitch before it can be compared to the NPS. SNR_{ideal} is then dependent of the accurate location of the detail with respect to the pixel sampling points. In the case of undersampling, NEQ and DQE are not strictly valid concepts. In spite of this, the DQE of digital imaging systems is often calculated using the presampling MTF in the formula for DQE (e.g.,

IEC 2003). This makes the interpretation of the resulting quantity challenging (Dobbins 1995 and 2000, Albert and Maidment 2000, Pineda and Barrett 2001, Gagne et al. 2001a and b).

The SNR measurement in digital imaging systems can also be done more directly, by not going through the transfer function analysis, but by estimating the expected signal (i.e., the numerator of eq. 1) from the difference of averaged signal and background images (Tapiovaara and Wagner 1993, Tapiovaara 1993, Chakraborty 1996, 1997a and 1999, Gagne et al. 2001a and 2006). Then, shift-invariance needs not be assumed, G and MTF(f_x,f_y) need not be measured, and one does not need a model for $\Delta s(x,y)$: an actual signal detail object is used instead. The practical difficulty with this approach is in obtaining a sufficient number of signal and background images for the averaging, so that the error from residual noise in the estimate of $\Delta s(x,y)$ is small, or correcting for the bias from the small number of sample images (Gagne and Wagner 1998, Tapiovaara 2003, Gagne et al. 2006). Linearity of the imaging system is not a prerequisite of this measurement. However, in order for the noise being multivariate normally distributed and for being able to extrapolate the SNR to different imaging conditions, it is useful to verify that the system is linear in the pixel value range of the images.

In measuring the SNR, it may not always be necessary to consider the strict ideal observer. Other sub-optimal computational observers, such as the non-prewhitening matched filter (NPWMF, Wagner and Brown 1985) or the DC-suppressing observer (Tapiovaara and Wagner 1993) perform often almost equally well and may be more easily employed. If the NPS is constant in the frequency range of the signal, the NPWMF is in fact the ideal observer. In cases where the noise is typical to projection x-ray imaging (low-pass noise) and the signal is of reasonable size the penalty of not prewhitening the data is small (Tapiovaara and Sandborg 1995). The DC-suppressing observer is otherwise similar to the NPWMF, but it does not use the DC component (average image brightness) in its decision. This suppression, or equivalently, normalizing images to the same average brightness, is often necessary for the computational observer to tolerate small variability in the average brightness of real radiographs: this suppression then brings the NPWMF observer closer to the ideal observer by filtering out the noisy DC-channel (Tapiovaara and Wagner 1993). Other observer models, which include some features of human vision, will be discussed later in section 5.1.

In addition to the under-sampling problems discussed earlier, the direct SNR-measurement method provides also a solution to a further problem in NEQ- and DQE-like quantities, which inherently apply only to imaging where a detail object changes only the intensity of the radiation and leaves the x-ray spectrum behind the detail unchanged (Tapiovaara and Wagner 1985 and 1993, Cahn et al. 1999). In x-ray imaging the detail of interest modifies also the x-ray spectrum

shape. Therefore, when optimising the x-ray imaging conditions (for example the x-ray spectrum), it is not sufficient to consider only NEQ or DQE, but one must consider the spectral dependence of radiation contrast as well, and include it in the factor $\Delta s(x,y)$ above. Spectral dependence is properly and automatically taken care of by the direct SNR measurement methods.

The above methods apply for static x-ray images. To measure the information relevant to the detectability of a static detail in fluoroscopy, one must determine the accumulation rate of SNR^2 (SNR^2_{rate}, see Tapiovaara and Wagner 1993, Tapiovaara 1993, 1997 and 2003, Tapiovaara and Sandborg 2004). This quantity is the live-image analogy of SNR^2 in static imaging, and is required in fluoroscopy because the information obtained depends on the length of the image sequence; in fluoroscopy, the ideal observer's SNR^2 for detecting a static signal is equal to SNR^2_{rate} times the imaging time (Tapiovaara 1993, see also Cunningham et al. 2001 and Siewerdsen et al. 2002). A full characterisation of fluoroscopic systems would require measuring the spatiotemporal noise power spectrum $W(f_x,f_y,f_t)$ and the spatiotemporal modulation transfer function $MTF(f_x,f_y,f_t)$. These could then be combined with the fluence rate and the characteristic curve to obtain the spatiotemporal DQE. Similarly, the SNR could be measured for time-dependent signals $\Delta s(x,y,t)$.

We make here one further note considering DQE. The main application of this quantity is to describe the efficiency of the image receptor. Therefore, in the calculation of DQE, the number of noise equivalent quanta is compared to the actual number of quanta impinging on the image receptor. This is not directly the optimisation problem that is of greatest interest in x-ray imaging. X-ray imaging efficiency is better described by comparing the achieved image quality (SNR^2) to the dose (D) of the patient. Therefore, in many papers discussing optimal imaging conditions optimisation is based on maximising the efficiency of radiation use in terms of the dose-to-information conversion factor SNR^2/D which is a nearly dose-independent quantity when quantum noise dominates. The imaging parameters which result to the maximum SNR^2/D are the most efficient parameters for the detection task considered, and the optimisation is concluded by deciding on the image quality level (or dose level) required. This final result is then the identification of the optimal imaging conditions, in the physical sense of image formation. The DQE can also be generalized to include other components of the imaging system than the image receptor alone. Kyprianou et al. (2005) have presented a DQE generalization to take into account the effect of the focal spot unsharpness and x-ray scatter from the patient.

However, all the above approaches are based on the SKE/BKE task and do not consider variability in (or prior uncertainty of) the signal and background. Good performance in SKE/BKE detection tasks does not always guarantee good

performance in clinically relevant detection tasks where there is less a-priori information available. It is known, that also in the ideal observer formalism it is necessary in some cases to include variability in the detection task in order to obtain results that would be reasonable to more realistic detection tasks (Myers et al. 1990, Myers 2000). However, when the prior information on the detection task is not complete, the ideal observer becomes mathematically complicated. An example of this is given in Brown et al. (1995), where the case of unknown signal position was considered. In the detection task of that paper a monotonic relationship was found for the performance in the SKE/BKE and unknown signal position cases, however. Wagner et al. (1990) have discussed also other detection tasks involving uncertainty of the signal and background. The relationship of the performance of the ideal observer and human observers is considered in more detail in chapter 5.

4.2 Phantom based approaches

The clinical imaging task can be made better defined and the case variability reduced by evaluating images of phantoms and test objects instead of real patients (chapter 3). X-ray technique optimisation studies are often based on simple phantoms with suitable test objects. Then, it is important that the results are also verified by actual patient studies before being put into use (e.g., Vassileva 2004).

As discussed in section 3.1 phantoms could be constructed to mimic patients within arbitrary detail. For example, Månsson et al. (1999) have compared various chest radiography systems by using an antropomorphic chest phantom and test details simulating pathology. Further, they did not rely on subjective measures of detectability but measured the detection performance of their observers using a modified free-response receiver operating characteristic (FROC) analysis, the free-response forced error experiment (FFE). In most phantom-based studies, however, detectability refers to a subjective judgement of detail visibility and the phantoms are stylised homogeneous blocks of material, and mimic the patient mainly just by scattering, filtering and attenuating radiation essentially similarly as a patient would do. Even then, the appearance of the test detail in the final processed images may be very different to what it would have been in actual patient images if images are digitally post-processed according to the image contents.

Various phantoms and test objects are extensively discussed in volume 49 of Radiation Protection Dosimetry (1993), which consists of papers presented at the workshop “Test Phantoms and Optimisation in Diagnostic Radiology and Nuclear Medicine”. After that, for example Guibelalde et al. (2001) have described the use

of dynamic phantoms and Balter (2001) has described a phantom intended to be used for dose and image quality measurements in cardiovascular fluoroscopy. Various quality control protocols for digital fluorography and digital subtraction angiography have been reviewed by Kotre and Marshall (2001) and Schreiner-Karoussou (2005) has reviewed image quality measurement standards for digital x-ray systems. Test methods that are intended for evaluating sharpness, contrast and noise only in general terms, without referring to particular test objects or test methods are discussed below. Other important image quality related factors, e.g., image distortions, artefacts and image homogeneity issues (with respect to brightness, sharpness or dead pixels) will not be reviewed either.

It is not always tried to simulate patient imaging in image quality measurements. For example, when the resolution of an x-ray image intensifier-television system is measured, it is customary to focus only to the performance of the image intensifier and television part of the whole x-ray system. The lead-bar test object is then placed as close to the image receptor as possible in order to minimise the effect of the x-ray tube focal spot, the contrast is kept maximum by using a low x-ray tube voltage and no phantom, and noise is minimised by setting the dose rate high. This is done in order to minimise the influence of any other factor than the sharpness of the x-ray image intensifier-television chain. The measurement result expresses the maximum line-pair frequency seen in the image. Similar measurements are also done for radiographic systems.

Such measurements are useful for equipment constancy testing purposes and for indicating the resolution of the imaging system, but they are not sufficient to fully characterise sharpness: as already mentioned earlier, they give no emphasis for lower frequencies than the maximum. For a more complete characterisation of image sharpness the MTF is needed. If the sharpness of the whole imaging system needs to be evaluated, the test object needs to be located further away from the image receptor and also movement unsharpness (resulting from movements of the patient or mechanical vibrations in the imaging system) may need to be taken into account. In dynamic imaging movement unsharpness is related to temporal lag and may, e.g., result to different amounts of blurring in continuous and pulsed fluoroscopy of moving objects.

Another common test of the performance of the image receptor is the measurement of the contrast threshold (also referred to as low contrast resolution), which is the lowest contrast detail (usually a circular disk) of a given size that can be seen in the image. The result is mainly dependent on the contrast of the detail and the noise of the image, but other factors (such as image sharpness, image display settings and the lag in a fluoroscopic system) contribute, too. Typically also this measurement is aimed to characterise just the image receptor part of the x-ray system and the influence of other factors is then minimised.

For example, the commonly used Leeds fluoroscopic test objects (Cowen et al. 1992) are intended to be used under specified conditions (x-ray tube voltage, additional filtration, no scatter) that allow the inherent radiation contrast of the details to be known and which facilitate reproducible testing. The results for a given x-ray system depend, among other things, on the dose rate (because of quantum noise), which should then be unchanged in constancy testing and be at least accounted for if the test is used for checking fulfilment of performance specifications or for comparing different x-ray systems against each other. There are various designs of test objects for contrast threshold measurement, both for fluoroscopic and radiographic imaging.

A variant of the contrast threshold measurement is the measurement of contrast-detail detectability (or threshold contrast detail detectability, C-D or TCDD), where the lowest visible contrast is measured and reported for a range of detail sizes (for further references, see e.g. Martin et al. 1999a, Dobbins 2000 or Evans et al. 2004). The measurement can then be considered combining the system's sharpness, contrast transfer and noise properties. Also in contrast-detail testing the result depends on the dose (radiographic imaging) or dose rate (fluoroscopic imaging), which then needs be considered along with the results (Dobbins 2000, Gallacher et al. 2003). Evans et al. (2004) concluded that image receptor structure noise may dominate over quantum noise in image-intensifier based digital radiography. Therefore, they suggest not making a dose effect correction to such C-D results although recommend correction for the effect of dose rate in fluoroscopy.

As already pointed out, the above measurements are often intended for characterising just the image receptor, not the imaging performance of the whole system. In order to better resemble patient imaging the measurement setup is sometimes modified: a soft tissue equivalent phantom is used to attenuate the x-ray beam and to produce scattered radiation in the image, and the test object is inserted within the phantom to properly include the unsharpness from the x-ray tube focal spot size (e.g., Geleijns et al. 1993 and van Engen et al. 2003). Such a measurement setup is often used especially in contrast-detail testing, and the results are then given in terms of the thickness of the last detectable detail (instead of contrast), as a function of the detail diameter. As in section 3.1, it is again noted that if the aim is to find optimal imaging conditions one needs to carefully consider the relevance of the phantom and the test details to the actual clinical task.

Test object details that are made of metal (e.g., aluminium, copper or gold) may not always be appropriate for comparing different imaging systems or imaging techniques, because their contrast may behave differently to the contrast of clinically interesting details with a changing radiation quality. One

may end up optimising the system to good performance in the phantom test instead of good performance in clinical tasks (as pointed out in, e.g., Bosmans et al. 2005). However, in mammography aluminium details have been reported to give reasonably similar energy dependence of contrast as would be obtained for microcalcifications or mass lesions in average breast material (Jennings et al. 1981, Brettle and Cowen 1994, the approximate similarity of the energy dependence of microcalcification and mass lesion signals is also noted in Dance et al. 2000). Aluminium is also often used to approximate the contrast of bony structures. Test details that are made of plastic disks or by drilling holes of different depths in a plastic plate should evidently be appropriate for simulating soft tissue details in chest imaging (Honey et al. 2005). However, care is required in using a contrast-detail phantom with plastic details because the diameter of the smallest details may be several times less than the thickness of the detail. The detail is not imaged as intended if the radiation is not directed accurately along the axis of the detail.

A serious problem in the above visual measurements is that one is trying to measure something that in fact does not exist: the detection threshold. In reality the transition from not visible to clearly visible is smooth and goes through various levels of certainty in perceiving the detail, somewhat similarly to the graph in Fig. 1. The observer needs to try to adapt to a certain critical confidence level in order to obtain consistent results. It is difficult to define, communicate and keep such a criterion. Therefore, notable variability between the results of different observers and the results in repeated observations of the same observer are constantly found both in radiography (Loo et al. 1983, Cohen et al. 1984) and fluoroscopy (Marshall et al. 1992, Launders et al. 1995, Tapiovaara and Sandborg 2004). Barrett and Myers (2004) criticize contrast detail measurements by the lack of any control of the decision threshold and particularly by the lack of a control of false positive detection (wishful thinking). They note that contrast detail testing is frequently used in quality assurance but caution against using such methods as a quantitative tool in the assessment of imaging systems. The precision of the measurements can be improved by increasing the number of observers, image samples and repeated readings (Cohen et al. 1984). In order to fully improve the precision all such readings should be independent, which requires careful planning of the image reading sessions. However, even if a large number of images and observers were used, it is not certain that the obtained average results would be always the same. The results are likely to depend on the instructions given for the observers and it may be that different observer groups tend to adopt different confidence criteria. This was seen in the experiments of Loo et al. (1983), where radiologists adopted more strict confidence criteria than other observer groups. Designing the test as a M-AFC or ROC experiment

may make the measurement results of various assessments better comparable (Loo et al. 1983), because the critical confidence level of declaring the detail visible is being controlled. However, much of the practicality of the test is then lost. A practical method to improve the detection of image quality changes with the subjective measurement method is to compare images against a previously taken reference image.

It is a natural attempt to try to achieve objectivity in test image evaluation by employing systems that are based on a computer analysis of images, instead of using human observers (Chakraborty and Eckert 1995). Many computer-based methods have been suggested for measuring the sharpness and low-contrast test detail detectability in phantom images. Often also other quality related features (e.g., image uniformity, image aspect ratio, artifacts) are evaluated in addition to them. Such automatic phantom image evaluation systems are available from some x-ray system manufacturers, to be used in the quality control of their equipment.

Some of the computer-based methods to measure low-contrast detail detectability are based on statistical decision theory (e.g., Desponds et al. 1991, Chakraborty 1996 and 1997a, Gagne et al. 2001a and 2006, Moeckli et al. 2000, Tapiovaara and Wagner 1993, Tapiovaara 1993 and 2003, Verdun et al. 1996). Other methods exist, too (Brok and Slump 1989, Chakraborty and Eckert 1995, Brooks et al. 1997, Castellano Smith et al. 1998, Dougherty 1998, Jansen and Zoetelief 2000, Wang et al. 2001, Kwan et al. 2003, Schiltz 2004a and 2004b, Jahnen 2004, Pascoal et al. 2005, Rampado et al. 2006, among others). Typically they measure the signal difference that a test object causes and treat the noise just by calculating the pixel variance, but other analysis methods have been used as well (e.g., Schiltz 2004a, where the assessment is based on edge extraction and Hough transformation). The methods are commonly reported to have yielded test phantom image quality scores which have been comparable to but more consistent than scores given by human observers. It seems clear that such systems should be very useful for constancy testing. However, these ad hoc methods cannot be confidently used for more general image assessment purposes, e.g., for setting image quality standards or for comparing different imaging systems: this would require SDT-based evaluation methods which properly take into account the characteristics of signal transfer and noise, and whose results relate directly to the information in the images.

The above discussion applies also for measurements of the contrast-to-noise ratio (CNR), where the pixel value difference between a test object (typically a thin aluminium disk) and its neighbourhood is measured and compared to the pixel value standard deviation in the background (van Engen et al 2003, Bosmans et al 2005). [CNR is equal to the concept of signal-difference-to-noise

ratio (SdNR, Samei et al. 2005).] Such measurements are useful for constancy testing and may also be used, e.g., for optimising the radiation quality in a given x-ray system if: (1) the contrasting detail material is reasonably chosen, (2) the dose-pixel value relationship is linear, (3) the NPS shape can be assumed to stay constant and (4) it is known that the modifications of the imaging parameters do not change the MTF of the system. Similarly as was concluded with other pixel value variance-based methods above, CNR is clearly not a suitable quantity for comparing different imaging systems to each other or for setting performance limits.

It is stressed that the methods using the variance of pixel values as a measure of noise are also sensitive to any deterministic trend or inhomogeneity in the area used: all deviations from a constant value are interpreted as being random noise. The method of Brok and Slump (1989) to evaluate image noise from the difference image between two similar radiographs or the method suggested by Samei et al. (2005) to obtain two images, one with and one without the signal detail, should be useful in computer based QA constancy testing because they remove the trends and other deterministic image inhomogeneities (structure noise) assuming that the repeatability of the x-ray generator and the image detector is good. The importance of this correction for quantitative measurements of actual random noise has also been pointed out by Burgess (2004). On the other hand, it may frequently be difficult for a human observer to make a distinction between deterministic but random-appearing image inhomogeneity and truly random image noise; their effect on detail detectability is often similar (Kume et al. 1986, Marshall et al. 2001). Then one should also take into account this type of structure noise in the image receptor in the evaluation of image quality. Even then, removal of large scale trends and artefacts should be made.

5 Relationships between the various assessment methods

5.1 Statistical decision theory and phantom imaging

There are many studies concerning the relationship between human performance (as measured by the AFC or ROC methods) and the performance of the ideal observer in detecting signals embedded in noise (for references, see ICRU 1996, Myers 2000, Barrett and Myers 2004). This detection task corresponds closely to that with flat-background test phantoms. The general finding in these experiments is that the detection performance of human observers is reasonably close to the performance of the ideal observer (typically human observer efficiencies of the order of 30–50%, and even up to 80%, have been found). The performance of human observers is therefore not far from the best achievable, and can be well predicted from the performance of the ideal observer. For equal performance the human observer just needs a somewhat lower noise or a somewhat higher contrast signal. Humans have been found to fall farther away from the ideal observer if the contrast of the displayed image is low, the signal extends to a large area (or is otherwise complicated (Burgess 1985)), or if the image noise is strongly coloured: some of these cases leave room for improving human performance by image post-processing. See Myers (2000) for a more thorough discussion on the relationship between the human and the ideal observer and references on the original publications.

The above results are often interpreted such that a human observer may base his/her decision on a matched filter operation, but with some inconsistencies, such as internal noise (both additive and induced, Burgess and Colborne 1988), a reduced sensitivity to both high and low spatial frequencies, an inability to use exactly the correct signal shape and position information, an inability to fully integrate over a large signal area, and an inability to accurately perform the prewhitening operation that the ideal observer would apply before match filtering the data. Some of these features have been included in various observer models, which then predict human signal detection performance even better than the ideal observer does. Such models include the perceived statistical decision theory model (Loo et al. 1984), the NPWE model (Burgess et al. 2001b) and the channelized Hotelling observer (Myers and Barrett 1987). For a more detailed discussion and references on observer models, see ICRU (1996), Myers (2000), Abbey and Bochud (2000), Eckstein et al (2000) and Barrett and Myers (2004). A large number of publications have shown the close relationship between the ideal

observer, the above observer models and the performance of human observers for various types of images (for references, see, e.g., ICRU 1996, Myers 2000, Abbey and Bochud 2000 and Barrett and Myers 2004). A recent paper (Marshall 2006) compared the NPWMF and NPWE models with human observer C-D data. The results agreed within approximately 15% for all dose levels studied and for all but the smallest detail diameter (0.1 mm). As noted in 4.1 such failing is not surprising because this detail size is near the pixel size of the system used (0.07 mm).

It can be noted that the observers above base their decision on a linear combination of the image data, and are therefore incapable to handle efficiently higher order classification tasks, involving for example detection of symmetry, noise level, or signals with unknown position. Humans, instead, are known to be able to classify images also in such cases. Therefore, the above models cannot be comprehensive models of human visual performance although humans may use the template-matching strategy in some visual tasks (Burgess 1985). Barrett and Myers (2004) suggest that human visual processing may consist of a combination of linear and logic operations, but it is also possible that humans are able to do higher-order than linear processing of image data: in fact visual perception is not a well understood process (Hawkins and Blakeslee 2004).

Although it is not actually known how visual information is processed in the eye-brain system, the above models are useful and successful in predicting human performance. In projection x-ray imaging of test phantoms human performance can be predicted with reasonable accuracy by calculating the performance of the ideal observer or the simple sub-optimal computational observers mentioned in 4.1. The channelized observer models provide a better agreement with human observers in strongly coloured noise and may be therefore preferred in such cases. The NPWE observer should result to improved human result agreement when very small or large details are involved and the viewing distance is known. It is not, however, always necessary or even reasonable to try to find the observer model that best matches with human performance. The ideal observer's performance is an absolute, and conceptually also the most fundamental measure of image information.

It would seem likely that even better agreement with human observer performance would be achieved by incorporating further characteristics of human vision in the observer models. This has been attempted in computational prediction of subjectively assessed image quality (Eckert and Bradley 1998, Martens and Meesters 1998, Pons et al. 1999, Winkler 1999; see also Wang et al. 2004). However, such refinements have not resulted in significant improvements over more simple computational methods, such as the measurement of the root mean squared error between the source image and processed image (Eskicioglu

and Fisher 1995, Martens and Meesters 1998, Eckert and Bradley 1998, Rohaly et al. 2000).

Human observer signal-detection performance has also been compared to the performance of various observer models in fluoroscopic imaging. Marshall (2001) found a reasonable match between the perceived non-prewhitening matched filter model and human observers in contrast-detail measurements in fluoroscopy when the threshold SNR was set at 3.5 and the image integration time in the noise measurement was 0.64 s. Tapiovaara and Sandborg (2004) made human observer AFC experiments of static low-contrast detail detectability and found that the results were well predicted by the relationship

$$d'_{human} = \sqrt{SNR^2_{rate} \cdot t_{eff}} \quad , \tag{2}$$

where t_{eff} is the effective image information integration time. Depending on the test setup, values of 0.3 s and 0.6 s were found for t_{eff} : the lower value applied to the 16-AFC test and the higher value to the 2-AFC test. This is different to results obtained in M-AFC experiments with static images, where d'_{human} has been found not to depend on the number of alternatives (M) for M values ranging from 2 to 1800 (Burgess and Ghandeharian 1984). This difference was suggested to result from the higher requirement on memory in the fluoroscopic M-AFC experiments (Tapiovaara and Sandborg 2004). It was further suggested that t_{eff} may also depend on the frame rate and temporal lag in the fluoroscopic imaging system. The relationship between contrast-detail performance and image lag has been studied by Marshall and Kotre (2002) for the case of a stationary test object, and by Kotre and Guibelalde (2004) for the case of a moving test object. The former of these papers demonstrated that the performance of human observers in detecting static details improves with an increasing lag, whereas the latter paper found that for abdomen examinations the persistence time constant should be approximately 0.15 s and for cardiac studies as low as possible.

Wilson et al. (2000) have compared human observer and model observer performance in low-acquisition rate fluoroscopy and digital temporal and spatial filtering of image data using computer generated white-noise images. The model observer considered was the NPWMF, modified by adding a spatio-temporal contrast sensitivity function and a temporal window function to describe the limited capacity of humans to use information in a sequence of images. This model was successful in predicting the performance of human observers to detect both stationary and moving targets in the image sequence.

Above, the human observer performance has been commonly measured by using M-AFC or ROC methods, which are able to provide objective and accurate performance measures. These methods are not practical for most

image evaluation purposes because of their labour-intensity. Therefore, phantom images are usually evaluated by subjective assessment of the lowest contrast and/or smallest details that are seen in the image. Any comparison of SDT-based measures and phantom scoring then suffers from the variability in the visual assessments. Tapiovaara and Sandborg (2004) also demonstrated that observers were not able to keep their confidence criteria constant in cases where the noise level was changed. Their change of subjectively judged threshold contrast was notably less than was expected from detectability tests with AFC methods.

Chakraborty has demonstrated (Chakraborty 1997b) that SNR is related to subjective ranking of phantom detail visibility also when the details are clearly seen. In his experiments SNR agreed well with observers' preference ranking of images in the SNR range of 5–81. ROC and AFC methodologies are applicable only at lower SNR levels, because they are based on the frequency of the observers' detection errors.

As in section 4.1 it is again noted that there is potential for misinterpreting the ideal-observer results in a SKE/BKE task, if the task is specified such that the features used by the ideal observer are not available in practical imaging tasks. An example of this is the problem of aperture-size optimisation in emission imaging. In a SKE/BKE task, the performance of the ideal observer improves with an increasing aperture size, while a statistically varying background leads to an aperture size that is matched to the signal (Myers et al. 1990). In this case the SKE/BKE task allows the observer using such features in the detection strategy that are not available in clinical work, and the conclusions based on the simplified optimisation are not valid in clinical tasks.

5.2 Computational analysis of test phantom images

The above discussion concerned the relationship between human observers and ideal (or sub-optimal) observers. As discussed in 4.2, also other types of computer programs are used for evaluating phantom images, mainly for equipment constancy testing purposes. Then it may not always be required that the evaluation is based on actual measurement of detectability as in SDT, but it often suffices that the results correlate with visual judgement and are able to verify whether the imaging system has deteriorated or not. Computational evaluation of the images will then usually outperform visual evaluation methods because the large variability in visual evaluation by humans can be avoided. For example, Chakraborty and Eckert (1995) found a high correlation between the computational image evaluation measurements they described and mean human observer results. A variability analysis showed that the largest components of variability in evaluating images were the between-reader and within-reader

variances of human observers. Such variability practically disappeared when the computer analysis system was used, and variability was then dominated by the case-sample variance (variability in the images). The precision was better by almost a factor of ten when compared with the American College of Radiology (ACR) method of scoring by three observers. Dougherty (1998) reported measurements of the contrast and a type of simple SNR-measure (contrast divided by the pixel standard deviation of digitised images) of these details. Of these measures, the contrast of the mass object was found to correlate best to the opinion of experts who visually ranked the images for overall quality and according to ACR scoring of detail visibility. Pascoal et al. (2005) evaluated a software package developed for automated scoring of CDRAD test object images and compared the software results human observer results. The precision of the software results was seen to be better than the precision of human observers and allowed detection of smaller low-contrast variations in QA measurements, although the variability between replicas of similarly obtained test object images was notable. However, the C-D curves produced by the software deviated somewhat from average human observer results: the software does not fully imitate an average observer. Similar conclusions of the usefulness of the computational methods are reached in many of the papers discussing them.

5.3 Physical image quality and clinical image quality

It is well known that it is more difficult to detect details against radiographic backgrounds of patients than against the uniform backgrounds of homogeneous phantoms. It is frequently found (Kundel et al. 1985, Samei et al. 1999, 2000 and 2003, Håkansson et al. 2005b) that detectability is not limited by system noise (such as quantum noise, film granularity and electronic noise), but often by normal anatomic structure: subtle abnormalities are frequently missed. Manning et al. (2004) conclude by eye tracking study results that the majority of missed lesions in chest radiology can be classified as failures of decision rather than detection. Therefore, because performance does not appear to be system noise-limited in many diagnostic tasks, it is often concluded that there may be notably room for dose reduction in radiology (e.g., Månsson et al. 2005). Similar thoughts have been also presented regarding the fixed pattern noise of the image receptor: it is not optimal to work at such a high dose level where fixed pattern noise limits detectability (Marshall et al. 2001).

The degree with which the anatomic background disturbs detail detection depends on the modality, among other factors. In cross-sectional imaging, such as CT, the background is simpler than in projection imaging where the patient's structures at different depths are superimposed over each other. Various methods,

such as classical tomography, digital tomosynthesis, digital subtraction imaging and dual-energy imaging can be used to reduce the influence of anatomic structure in projection x-ray images.

Anatomical background complicates search operations by making the scene busy and full of visible structures. Signals may be masked or mimicked by overlying anatomical background. These result to signal misses and false alarms and are obvious factors deteriorating human performance. However, they do not affect the ideal observer in the SKE/BKE task and are therefore not considered within the mathematical formalism described in 4.1. Of course, the signals and backgrounds in radiology are not fully a-priori known, and the ideal observer in the actual clinical detection task would then suffer from the variability of the signal and the background as well. However, the statistics of these variabilities are not known and ideal observer performance cannot be calculated. Also other abnormalities than the one being sought may influence detection by the satisfaction-of-search effect. The observer may stop the search prematurely after finding one abnormality in the image (Krupinski 2000, Samei et al. 2000).

The mechanisms of how anatomical details decrease human observers' detail detectability are not well understood. It has been proposed that in addition to truly random system noise and deterministic image receptor non-uniformities (structure noise or fixed pattern noise) also the overlaying of many small anatomical structures along the x-ray beam leads to a noise-like pattern without distinguishable structures in the projected image (Tischenko et al. 2003, Hoeschen et al. 2005). These authors make a strong distinction between this anatomical noise and anatomical background, where the latter in their terminology refers to distinguishable anatomical structures. Håkansson et al. (2005a) concluded that the detectability of nodules in chest radiography is limited more because of such anatomical noise than the technical noise at the dose levels used today. However, the disturbing effect of anatomical background was found to be larger than that of anatomical noise. Similar conclusion of present-day chest images not being quantum-noise limited has been made by Sund et al. (2004) who measured the presampling MTF and DQE of four digital chest units and compared the results with subjective clinical evaluation of images of 23 volunteers. The clinical evaluation results could not be wholly explained by the physical image quality measurements; instead they believed that the differences in the clinical evaluation results were related to anatomical background and differences in image processing.

The above distinction between anatomical noise and background is not always made, and the disturbing anatomic background variability is often also called anatomical noise. It has been shown to deteriorate detectability to a

much higher degree than actual stochastic noise in some cases (Ruttimann and Webber 1983, Kotre 1998, Bochud et al. 1999, Samei et al. 1999, Burgess et al. 2001a and 2001b).

The effect of anatomical structures on detection might sometimes be expected to be almost insignificant. This could happen if the signal features would not interfere with the background features and the observer could be assumed to be able to mentally subtract the background. At the other end of expectation, when one might not be able to infer and mentally ignore the background structure, anatomical background variability could be treated as being random noise. The degree to which these alternative expectations apply depends at least partly on the anatomical region and detection task in question (Ruttimann and Webber 1983, Bochud et al. 1999, Burgess et al. 2001b, Båth et al. 2005c). For example, it would be clearly not appropriate to treat ribs in chest radiology purely as noise (with a random phase between the various spatial frequencies): such structures have a strong spatial correlation and treating them by measuring their spatial frequency spectrum alone, without considering the phase information, would overestimate their effect on detail detectability (Båth et al. 2005c). In mammography the effects of the anatomical background have been found to be different for mass objects and microcalcifications and to depend also on the strength of the visible anatomical background in the experiments of Bochud et al. (1999). The background acted partly as random noise and partly as a recognizable structure in detecting mass objects, but the effect depended on the strength of the anatomical background in microcalcification detection: a strong anatomical background acted as pure noise while a lower anatomical background disturbed the detection only slightly. Burgess et al. (2001b) came to similar conclusions about mass lesion detectability in mammographic backgrounds. In general, neither of the above extreme expectations seems to be valid for human observers, who often operate somewhere between these two interpretations: background variability appears to be a mixture of noise and deterministic components. For a more detailed discussion on this matter, see e.g., Burgess et al. (2001b), Samei et al. (2000) and Båth et al. (2005c) and the references in these papers.

Burgess et al. (2001b) studied the detectability of nodules in mammographic backgrounds and random noise backgrounds with the same average NPS ($1/f^3$ noise). They found that the nodule was easier to find in the true mammographic backgrounds in a signal-known-exactly experiment, but in a search experiment human performance was equal for both backgrounds, suggesting that mammographic background can be considered to be pure random noise for this task. The required contrast for equal detectability in both background cases increased with increasing nodule size, conversely to common C-D diagrams obtained in typical system noise cases. Human observer results were reasonably

well predicted by the channelized Fisher-Hotelling observer model, and the efficiency of human observers against the ideal observer was about 40% in the simulated noise case – of the same order that has been found in experiments with white noise. They concluded that although mammographic backgrounds are not stationary, SDT -based observer models are useful to predict human performance in mammography.

Sandborg et al. (2001a) studied the correlations between image criteria-based subjective evaluation of radiographs and physical image quality measures (contrast and SNR of specified details) in chest and lumbar spine film-screen radiography. Their results show significant correlation of blood vessel contrast, especially in the retrocardiac area, and the subjective evaluations in chest imaging. The correlation of the SNR of the blood vessel and subjective evaluation was lower: this was suggested to indicate that in film-screen chest radiography clinical image quality is more limited by contrast than by noise. In lumbar spine imaging the best predictors of clinical image quality were the contrast and SNR of small soft tissue cavities in bone (trabecular structures).

Ullman et al. (2004) have studied the effect of x-ray tube voltage on digital chest and pelvis radiography. Clinical image quality was evaluated by the relative VGA method and slightly modified CEC image criteria from radiographs of an anthropomorphic phantom, obtained at the same effective dose at each x-ray tube voltage. Physical image quality was described by the average value of the ideal observer's SNR for a number of small details at various locations in the phantom. In chest imaging small details of blood, soft tissue and bone embedded in the lung tissue were used; in pelvis imaging the details were small details of bone embedded in soft tissue. The SNR values were obtained by Monte Carlo simulation. Both the clinical and the physical evaluation resulted to image quality decreasing monotonically with an increasing x-ray tube voltage in the range studied (70–150 kV in chest radiography, 50–102 kV in pelvis radiography). They found a positive linear relationship between the results of the two evaluation methods (chest PA: r^2=0.91, pelvis AP: r^2=0.94), indicating that the SNR is strongly related to the radiologists' grading of the images.

Tingberg et al. (2004) studied how an altered contrast, obtained by simulating different characteristic curves of films (H&D curves), would change clinical evaluation of chest and lumbar spine film-screen radiographs. They found that steep curves were preferred, but cautioned that this could be a consequence of their masking of the films to show only small local areas of interest; their earlier tests with unmasked radiographs suggested that the overall impression of high gradient films was poor. They further noted that lumbar spine images taken at 90 kV were significantly worse than those taken at 70 independent of the film gradient. Tingberg et al. (2002) studied the relative importance of spatial

resolution and noise by altering digitized lumbar spine radiographs. Clinical image quality was assessed by the visual grading analysis method, and it was seen that added noise was more deteriorating than reduced spatial resolution. They also noted that the appearance of a noisy image can sometimes be improved by reducing the spatial resolution.

Redlich et al. (2005) have assessed several chest radiography systems by measuring their DQE, performing a VGA study using clinical images and a ROC study using images of an anthropomorphic phantom with details simulating pathology. It was noted that the ranking of the image quality of the systems was nearly the same with all these assessment systems. This is not surprising because the images of the phantom and the images of each patient were taken with the same dose and technique factors for all the x-ray systems being compared. The effects of other influencing factors than physical image quality (and image post-processing) were then largely removed, and the effect of physical image quality was highlighted. Anyway, also this study supports the expected result that physical image quality is monotonically related to clinical image quality.

Somewhat more qualitative evidence of the relationship between technical and clinical image quality is given by Vañó et al. (1995) who noted that their Leeds TOR(CDR) and scattering phantom images that were highly scored by physicists coincided in time with clinical images fulfilling technique-related image criteria. They concluded that the correct status of the x-ray system can be verified by the evaluation of phantom or clinical images.

Walsh et al. (2005) made measurements with Leeds test objects on several fluoroscopic systems and compared these results with the users' subjective opinion of image quality. In their paper they mainly reported on the clinicians' judgment of image contrast and compared it to the threshold contrast measured by the test object approach. Practically no correlation was found between these two evaluations. They explained this by the small differences between the qualities of the systems. (Both the test object results and the mean clinical judgements were typically average to good. Only one system was found that was classified as poor; this judgement was reached in both assessments). Overall, only a low correlation was found between the rank order by these two assessment methods. Walsh et al. point out that because of the spread of the clinical judgements of different patient cases it is necessary to use the average score of several cases in the clinical assessment. A clear trend of perceived image quality decreasing with an increasing patient size was seen in the results (Walsh, personal communication). This same effect is more easily seen when the thickness of the scattering phantom is increased in test object imaging; the magnitude of the effect may, of course, vary from one imaging system to another, for example because of different technique

adjustment methods used in various automatic brightness control systems or anti-scatter grids of different types.

Metz et al. (1995) have reviewed the assessment of medical image quality, and noted that there exists a wide consensus in measuring the sensitometric quantities, the MTF, and the NPS of radiological systems. They also agreed that the combined measures NEQ, DQE and SNR_{ideal} are useful for normalising the measurements on an absolute scale and for relating those measurements to the decision performance of the ideal observer. However, they stress that in the two-stage (recording and display) description of the imaging process, SNR_{ideal} describes image quality at the stage of image recording. This can be considered an advantage for understanding the steps through which images are formed, but cannot be used alone to predict the ranking of images that a human might make on basis of the displayed images if the characteristics between the images are too different. In many cases, however, such as projection radiography using simple phantoms, the human and ideal observer results show a good correlation. Human performance is not well understood for many clinically relevant tasks, and the relevance of the objective measurements to human observer performance is not clear in all cases. Metz et al. (1995) stress that the assessment of medical imaging systems requires going also beyond phantom/laboratory measurements into the clinical setting, where clinical performance can be assessed by ROC-studies, for example.

This conclusion is accepted also in ICRU (1996), which recommends characterising an imaging system firstly by using physical measurements of the large area signal transfer, MTF and Wiener spectrum and combining these to the NEQ and/or calculating the ideal observer's SNR for reasonably clinical-related tasks. In addition to these measurements image quality should be assessed visually by using well-controlled tests, such as ROC or 2-AFC methods, with images as close to the clinical situation as possible.

ICRU (2003) contains a description of various laboratory and field methods for assessment of image quality related factors in chest x-ray imaging. It is suggested that the whole imaging chain is tested with additional separate tests on the various components of the chain. In the report it is noted that field assessments of sharpness and noise by test object detail visibility involve highly subjective judgements and test results should be interpreted with caution. As a complementary approach to test object imaging, constant evaluation of the visibility of normal anatomy is also recommended in the report and inclusion of a few further image features to the European image criteria are suggested.

In summary of the above, there can be thought to exist two regions in x-ray imaging, characterised by asymptotes corresponding to either quantum-limited or quantum-saturated imaging. In the former, detection is mainly limited

by the imaging system noise and can be improved by decreasing the noise, e.g. by increasing the dose to the patient. In the latter, detectability is limited by anatomical variability (and image receptor fixed pattern noise) and will not be notably improved along with a further reduction of the random noise in the imaging system; this latter zone corresponds to the saturation level of Fig.1. Evaluation of image quality by simple test phantoms is not sufficient to optimise the noise level in clinical x-ray images when anatomical background is an issue (Månsson et al. 2005). This is not always the case; for some features detectability is mainly limited by random noise (Bochud et al. 1999). However, it seems that much of the present day x-ray imaging is performed in the quantum-saturated zone and would then leave room for lowering the patient's dose. This could be done much more easily in present-day digital x-ray equipment than in screen-film based x-ray systems.

Removing anatomical background often requires the use of special imaging techniques instead of traditional projection imaging. Therefore, being largely an uncontrollable factor in a given imaging system, it may not always be a central issue in image quality considerations: performance specification, quality assurance and some optimisation tasks. Also, it is noted that the detectability of some clinical features is limited by technical image quality issues instead of anatomical complexity and that anatomical background may not be as disturbing in all projection radiography as it is in mammography and chest imaging, where anatomical background is remarkable. It seems clear that keeping all other factors equal, improving the technical quality of images (in the SNR_{ideal} or NEQ(f) sense, for example) will result to an improvement in performance – although, as discussed above, the improvement may sometimes be minor.

6 Conclusions

A generally accepted principle is that image quality is most meaningfully defined and measured in relationship with the intended task of the image. Therefore, the best way of evaluating the quality of medical imaging should be to measure clinical performance by quantitative methods, such as the ROC analysis. This is not usually a practical option, however: if clinical images are used, one must generally be content with subjective, opinion-based evaluations instead of a truly quantitative measurement. Subjective evaluation suffers from inter-observer, intra-observer and case-sample variability, which restrict its use to reliably finding only large image quality differences. The precision can be improved significantly if the evaluation is done in a relative way, by comparing images side-by-side. Anyway, the significance to actual clinical performance remains often unclear. We further note that no patient image evaluation method can be considered as a measurement in the common meaning of the word: the results will be dependent on the diagnosticians and patient material in the study. It is difficult to see any method of calibrating clinical image quality measurement so that results obtained at different laboratories could be directly compared.

Case-sample variability can be reduced and a better transportability of the results introduced by using phantoms instead of patients. Technical image quality is frequently measured using simple uniform phantoms and various test objects and is reported in terms of visibility limits, such as contrast-detail curves and limiting resolution. However, it is still difficult to achieve equal results with different observers – and repeatable results with the same observer – in subjective threshold visibility tests. Controlled detectability measurements (such as ROC or AFC) allow for a well-defined measure of detectability, but suffer from the large number of observations required for precise results.

Digital x-ray imaging has made the division to the image data capture and display stages even more evident than it was in film-screen imaging. The data capture stage determines the information content of the image and can be analysed in detail, although this is presently limited to simple detection tasks in practice. The display stage should attempt to build an efficient interface to the human visual system, such that the captured image information is well perceived. The data capture stage can be considered as being the more fundamental of these two: it sets the performance that the ideal observer will achieve and no other observer can exceed even by utilizing any available post-processing or display improvements. In the physical sense, image quality is best specified by the ideal observer's SNR if a particular imaging task can be specified, or by NEQ(f) or DQE(f) when one wishes to evaluate the image receptor for a range of tasks. These measures can be related to the results of simple test object imaging.

The main difference between this approach and the evaluation of test object visibility is that the image quality is evaluated at the data stage by the ideal (or a sub-optimal) observer instead of a human observer doing the evaluation at the display stage.

Of course, the display stage is of great significance, because image information which is not perceived is not useful. In visual image quality assessment the two stages are evaluated together. This allows the testing of the whole imaging chain at once. On the other hand, this may sometimes obscure matters because it cannot always be known which of the two stages is limiting performance.

Various ways of assessing image quality – in the clinical, technical and physical sense of the concept – have been discussed above and studies of the relationships between various assessment results have been reviewed. In the review it was seen that the relationship between the SDT-based image quality measures and the performance of human observers in simple detection tasks is reasonably well understood. However, this does not extend to clinical imaging where the masking effects of anatomical background and the prior uncertainty of the signal and background complicate the situation. Which of the image quality evaluation methods should be used is clearly dependent on the purpose of the image quality evaluation task and the resources that can be used for accomplishing it.

It seems that equipment specification is best done in terms of the objective SDT-based quantities (NEQ(f) and DQE(f)). They relate directly to the information content in the images, the measurement methods can be standardised, and the measurements can be repeated to see whether specifications have been met. This cannot easily be done by using visually evaluated descriptors of technical image quality because the critical confidence level of detail visibility is not controlled.

Quality control constancy testing requires methods that are not too labour-intensive and expensive; instead, they must be sensitive to detect changes in the imaging system. To fulfil these objectives, it may be reasonable to relax requirements of the results being directly descriptive of diagnostic performance, although diagnostic performance should be kept in mind when deciding on actions on deteriorated imaging performance. If such changes cannot be handled by simple corrective actions, but require expensive investment in equipment, it may be more reasonable to make decisions based on some kind of clinical evaluation than by simple technical limits of measured parameters. Establishing the relationship between technical image quality parameters and clinical performance has proved to be difficult or impossible. For example, the resolution limits commonly set to

film-screen mammography have not been met in digital mammography. In spite of this digital systems have generally been found to be clinically acceptable.

There are several approaches that can be used for optimising x-ray imaging techniques. If the anatomic background is not an issue, it seems credible that optimal imaging conditions can be identified by finding the technique factors where SNR^2/D is maximum for the detail type of interest (e.g. iodine contrast material in a phantom). If resolution-related things are not of interest, one may even use the CNR instead of the ideal (or sub-optimal) observer's SNR. Of course, such results must be verified by clinical experiments and finally the dose level must be set such that image noise does not compromise clinical performance. On the other hand, Månsson et al. (2005) criticize the use of contrast-detail phantoms (and other test methods that are based on homogeneous patient simulating phantoms) for optimisation studies, and suggest that their use should be limited to constancy checks. They argue that optimisation by such methods is not relevant to the actual tasks in diagnostic radiology, where lesion detectability is frequently much more limited by anatomical background than by system noise (e.g., quantum and electronic noise); therefore, optimisation studies need be done with actual patient images or high-quality anthropomorphic phantoms. They note that this approach enables one to reduce radiation doses in cases where the diagnosis is not quantum-limited. Also Busch and Faulkner (2005) reach the same conclusion that optimisation must be based on clinical studies instead of using test phantoms, whereas test phantom imaging is useful for, e.g., quality control and standardisation purposes. Test object performance data have been collected in a number of x-ray departments (e.g., Evans et al. 2004). Although such data are not directly related to clinical requirements, they should be useful for indicating typical and/or acceptable x-ray system performance (e.g., Cowen 1993). Contrast-detail testing is tempting because it considers the whole imaging chain and the results are straightforward to interpret. The transportability of test results is difficult to ensure, however, and the relatively high variability makes the testing often insensitive to small or moderate changes in the imaging system. This could be improved by using SDT-based computational observers instead of humans, but the display stage then needs separate consideration.

Acknowledgements

This report has been prepared as part of the SENTINEL project. The SENTINEL project, contract FP6- 012909, was partially supported and has received funding from the EC-Euratom Sixth Framework Programme. Keith Faulkner, Paula Pöyry, Virginia Tsapaki and Hans Zoetelief are thanked for their helpful commenting on the manuscript.

References

Abbey CK and Bochud FO, Modeling visual detection tasks in correlated image noise with linear model observers, in *Handbook of Medical Imaging, Vol 1: Medical Physics and Psychophysics*, Beutel J, Kundel H, Van Metter R (eds.), SPIE, Bellingham 2000.

Albert MA and Maidment DA, Linear response theory for detectors consisting of discrete arrays, Med. Phys. 27, 2417–2434, 2000.

Almén A, Tingberg A, Mattson S, Besjakov J, Kheddache S, Lanhede B, Månsson LG and Zankl M, The influence of different technique factors on image quality of lumbar spine radiographs as evaluated by established CEC image criteria, Br. J. Radiol. 73, 1192–1199, 2000.

Almén A, Tingberg A, Besjakov J and Mattson S, The use of reference image criteria in x-ray diagnostics: an application for the optimisation of lumbar spine rediographs, Eur. Radiol. 14, 1561–1567, 2004.

Balter S, A new tool for benchmarking cardiovascular fluoroscopes, Rad. Prot. Dosim. 94, 161–166, 2001.

Barrett HH and Swindell W, *Radiological Imaging, Volumes I and II*, Academic Press, New York, 1981.

Barrett HH and Myers K, *Foundations of Image Science*, John Wiley and Sons, 2004.

Bernardi G, Padovani R, Spedicato L, Desmet W, Malisan MR, Giannuleas JD et al, Image quality criteria in cardiology, Rad. Prot. Dosim 117, 102–106, 2005a.

Bernardi G, Padovani R, Desmet W, Peterzol A, Giannuleas JD, Neofotisou E et al, A study to validate the method based on DIMOND quality criteria for cardiac angiographic images, Rad. Prot. Dosim 117, 263–268, 2005b.

Bochud FO, Valley J-F, Verdun FR, Hessler C and Schnyder P, Estimation of the noisy component of anatomical backgrounds, Med. Phys. 26, 1365–1370, 1999.

Bosmans H, Carton A-K, Rogge F, Zanca F, Jacobs J, Van Ongeval C et al, Image quality measurements and metrics in full field digital mammography : an overview, Rad. Prot. Dosim. 117, 120–130, 2005.

Brettle DS and Cowen AR, Dual-energy digital mammography utilizing stimulated phosphor computed radiography, Phys. Med. Biol. 39, 1989–2004, 1994.

Brok M and Slump CH, Automatic determination of image quality parameters in digital radiographic imaging systems, SPIE Vol. 1090, 246–256, 1989.

Brooks KW, Trueblood JH, Kearfott KJ and Lawton DT, Automated analysis of the American College of Radiology mammographic accreditation phantom images, Med. Phys. 24, 709–723, 1997.

Brown DG, Insana MF and Tapiovaara M, Detection performance of the ideal decision function and its McLaurin expansion: Signal position unknown, J. Acoust. Soc. Am. 97, 379–398, 1995.

Burgess AE, Visual signal detection III: On Bayesian use of prior knowledge and cross correlation, J. Opt. Soc. Am. A 2, 1498–1507, 1985.

Burgess AE, Comparison of receiver operating characteristic and forced choice observer performance methods, Med. Phys. 22, 643–655,1995.

Burgess AE, On the noise variance of a digital mammography system, Med. Phys. 31, 1987–1995, 2004.

Burgess AE and Ghandeharian H, Visual signal detection II: Signal-location identification, J. Opt. Soc. Am. A 1, 906–910, 1984.

Burgess AE and Colborne B, Visual signal detection IV: observer inconsistency, J. Opt. Soc. Am. A 5, 617–627, 1988.

Burgess AE, Jacobson FL and Judy PF, Lesion detection in digital mammograms, Proc. SPIE 4320, 555–560, 2001a.

Burgess AE, Jacobson FL and Judy PF, Human observer detection experiments with mammograms and power-law noise, Med. Phys. 28, 419–437, 2001b.

Busch HP, DIMOND III – Image Quality and Dose Management for Digital Radiography – Final report, Trier 2004.
Available at http://www.dimond3.org (→Reports→WP1).

Busch HP and Faulkner K, Image quality and dose management in digital radiography: a new paradigm for optimisation. Rad. Prot. Dosim. 117, 143–147, 2005.

Båth M, Håkansson M, Hansson J and Månsson LG, A conceptual optimisation strategy for radiography in a digital environment, Rad. Prot. Dosim. 114, 230–235, 2005a.

Båth M, Håkansson M, Tingberg A and Månsson LG, Method of simulating dose reduction for digital radiographic systems, Rad. Prot. Dosim. 114, 253–259, 2005b.

Båth M, Håkansson M, Börjesson S, Kheddache S, Grahn A, Bochud FO, Verdun FR and Månsson LG, Nodule detection in digital chest radiography: part of image background acting as pure noise, Rad. Prot. Dosim. 114, 102–108, 2005c.

Börjesson S, Håkansson M, Båth M, Kheddache S, Svensson S, Tingberg A, Grahn A, Ruschin M, Hemdal B, Mattson S and Månsson LG, A software tool for increased efficiency in observer performance studies in radiology. Rad. Prot. Dosim. 114, 45–52, 2005.

Cahn RN, Cederström B, Danielsson M, Hall A, Lundqvist M, Nygren D, Detective quantum efficiency dependence on x-ray energy weighting in mammography, Med. Phys. 26, 2680–2683, 1999.

Castellano Smith AD, Castellano Smith IA and Dance DR, Objective assessment of phantom image quality in mammography: a feasibility study, Br. J. Radiol. 71, 48–58, 1998.

Chakraborty DP, Physical measures of image quality in mammography, Proc. SPIE 2708, 179–185, 1996.

Chakraborty DP, Computer analysis of mammography phantom images (CAMPI): An application to the measurement of microcalcification image quality of directly acquired digital images, Med. Phys. 24, 1269–1277, 1997a.

Chakraborty DP, Comparison of computer analysis of mammography phantom images (CAMPI) with perceived image quality of phantom targets in the ACR phantom, Proc. SPIE 3036, 160–167, 1997b.

Chakraborty DP, Effect of antiscatter grid and target/filters in full-field digital mammography, Proc. SPIE 3659, 878–885, 1999.

Chakraborty DP, The FROC, AFROC and DROC variants of the ROC analysis, in *Handbook of Medical Imaging, Vol 1: Medical Physics and Psychophysics*, Beutel J, Kundel H, Van Metter R (eds.), SPIE, Bellingham 2000.

Chakraborty DP and Eckert MP, Quantitative versus subjective evaluation of mammography accreditation phantom images, Med. Phys. 22, 133–143, 1995.

Cohen G, McDaniel DL and Wagner LK, Analysis of variations in contrast-detail experiments, Med. Phys. 11, 469–473, 1984.

Commission of the European Communities (CEC), European guidelines and quality criteria for diagnostic radiographic images. EUR 16260 EN, CEC, Brussels 1996a.

Commission of the European Communities (CEC), European guidelines and quality criteria for diagnostic radiographic images in paediatrics. EUR 16261 EN, CEC, Luxembourg 1996b.

Commission of the European Communities (CEC) European guidelines on quality criteria for computed tomography. EUR 16262 EN, CEC, Luxembourg 2000.

Cowen AR, The application of image quality test objects in support of clinical X ray service: physical and technical considerations, Rad. Prot. Dosim. 49, 27–33, 1993.

Cowen AR, Clarke OF, Coleman NJ, Craven DM, McArdle S and Hay GA, Leeds x-ray test objects, Instruction manual, The University of Leeds, 1992.

Cowen AR, Launders JH, Jadav M and Brettle DS, Visibility of microcalcifications in computed and screen-film mammography, Phys. Med. Biol. 42, 1533–1548, 1997.

Cunningham IA, Applied linear-systems theory, in *Handbook of Medical Imaging, Vol 1: Medical Physics and Psychophysics*, Beutel J, Kundel H, Van Metter R (eds.), SPIE, Bellingham 2000.

Cunningham IA, Moschandreou T, Subotic V, Detective quantum efficiency of fluoroscopic systems: the case for a spatial-temporal approach (or, does the ideal observer have infinite patience?), Proc. SPIE 4320, 479–488, 2001.

Dainty JC and Shaw R, *Image Science*, Academic Press. London, 1974.

Dance DR, Thilander Klang A, Sandborg M, Skinner CL, Castellano Smith IA and Alm Carlsson G, Influence of anode/filter material and tube potential on contrast, signal-to-noise ratio and average absorbed dose in mammography: a Monte Carlo study, Br. J. Radiol. 73, 1056–1067, 2000.

Desponds L, Depeursinge C, Grecescu M, Hessler C, Samiri A and Valley JF, Image quality index (IQI) for screen-film mammography, Phys. Med. Biol. 36, 19–33, 1991.

Dobbins JT III, Effects of undersampling on the proper interpretation of modulation transfer function, noise power spectra, and noise equivalent quanta of digital imaging systems, Med. Phys. 22, 171–181, 1995.

Dobbins JT III, Image quality metrics for digital systems, in *Handbook of Medical Imaging, Vol 1: Medical Physics and Psychophysics*, Beutel J, Kundel H, Van Metter R (eds.), SPIE, Bellingham 2000.

Dobbins JT III, Samei E, Ranger NT and Chen Y, Intercomparison of methods for image quality characterization. II. Noise power spectrum, Med. Phys. 33, 1466–1475, 2006.

Dougherty G, Computerized evaluation of mammographic image quality using phantom images, Computerized Medical Imaging and Graphics 22, 365–373, 1998.

Doyle AJ, Le Fevre J and Anderson GD, Personal computer versus workstation display: observer performance in detection of wrist fractures on digital radiographs, Radiology 237, 872–877, 2005.

Eckert MP and Bradley AP, Perceptual quality metrics applied to still image compression. Signal Processing 70, 177–200, 1998.

Eckstein MP, Abbey CK and Bochud FO, A practical guide to model observers for visual detection in synthetic and natural noisy images, in *Handbook of Medical Imaging, Vol 1: Medical Physics and Psychophysics*, Beutel J, Kundel H, Van Metter R (eds.), SPIE, Bellingham 2000.

Eisenhuber E, Stadler A, Prokop M, Fuchsjäger M, Weber M and Schaefer-Prokop C, Detection of monitoring materials on bedside chest radiographs with the most recent generation of storage phosphor plates: dose increase does not improve detection performance, Radiology 227, 216–221, 2003.

Eskicioglu AM, Fisher PS, Image quality measures and their performance, IEEE Transactions on Communications 43, 2959–2965, 1995.

European CT study group, 2004 CT Quality Criteria. Available at http://www.msct.info/CT_Quality_Criteria.htm, 2004.

Evans DS, MacKenzie A, Lawinski CP and Smith D, Threshold contrast detail detectability curves for fluoroscopy and digital acquisition using modern image intensifier systems, Br. J. Radiol. 77, 751–758, 2004.

Fujita H, Ueda K, Morishita J, Fujikawa T,Ohtsuka A and Sai T, Basic imaging properties of a computed radiographic system with photostimulable phosphors, Med. Phys. 16, 52–59, 1989.

Gagne RM and Wagner RF, Prewhitening matched filter: practical implementation, SNR estimation and bias reduction, Proc. SPIE 3336, 231–242, 1998.

Gagne RM, Myers KJ and Quinn PW, Effect of shift invariance and stationarity assumptions on simple detection tasks: spatial and spatial frequency domains, Proc. SPIE 4320, 373–380, 2001a.

Gagne RM, Boswell JS, Myers KJ and Peter G, Lesion detectability in digital radiography, Proc. SPIE 4320, 316–325, 2001b.

Gagne RM, Gallas BD and Myers KJ, Toward objective and quantitative evaluation of imaging systems using images of phantoms, Med. Phys. 33, 83–95, 2006.

Gallacher DJ, Mackenzie A, Batchelor S, Lynch J, Saunders JE, Use of a quality index in threshold contrast detail detection experiments in television fluoroscopy. Br. J. Radiol. 76, 464–472, 2003.

Geijer H and Persliden J, Varied tube potential with constant effective dose at lumbar spine radiography using a flat-panel digital detector, Rad. Prot. Dosim. 114, 240–245, 2005.

Geleijns J, Schultze Kool LJ, Zoetelief J, Zweers D and Broerse JJ, Image quality and dosimetric aspects of chest x ray examinations: measurements with various types of phantoms, Rad. Prot. Dosim. 49, 83–88, 1993.

Guibelalde E, Vañó E, Kotre CJ, Faulkner K, Fernández JM, Ten JI and Rawlings DJ, The use of dynamic phantoms in interventional radiology, Rad. Prot. Dosim. 94, 155–159, 2001.

Hawkins J and Blakeslee S, *On intelligence*, Times Books, New York, 2004.

Hemdal B, Andersson I, Grahn A, Håkansson M, Ruschin M, Thilander-Klang A, Båth M, Börjesson S, Medin J, Tingberg A, Månsson LG and Mattson S, Can the average glandular dose in routine digital mammography screening be reduced? A pilot study using revised image quality criteria, Rad. Prot. Dosim. 114, 385–390, 2005.

Hoeschen C, Tischenko O, Buhr E and Illers H, Comparison of technical and anatomical noise in digital thorax x-ray images, Rad. Prot. Dosim. 114, 75–80, 2005.

Honey ID, MacKenzie A and Evans DS, Investigation of optimum energies for chest imaging using film-screen and computed radiography, Br. J. Radiol. 78, 422–427, 2005.

Håkansson M, Båth M, Börjesson S, Kheddache S, Grahn A, Ruschin M, Tingberg A, Mattson S and Månsson LG, Nodule detection in digital chest radiography: summary of the RADIUS chest trial, Rad. Prot. Dosim. 114, 114–120, 2005a.

Håkansson M, Båth M, Börjesson S, Kheddache S, Johnsson ÅA and Månsson LG, Nodule detection in digital chest radiography: effect of system noise, Rad. Prot. Dosim. 114, 97–101, 2005b.

ICRP Publication 73. Radiological protection and safety in medicine. Annals of the ICRP 26(2), 1996.

ICRU Report 41 Modulation transfer function of screen-film systems, International Commission on Radiation Units and Measurements, 1986.

ICRU Report 54 Medical imaging - the assessment of image quality, International Commission on Radiation Units and Measurements, 1996.

ICRU Report 70 Image quality in chest radiography, Journal of the ICRU Vol. 3 No.2, 2003.

IEC standard 62220-1, Medical electrical equipment - Characteristics of digital X-ray imaging devices – Part 1: Determination of the detective quantum efficiency, IEC Central office, Geneva, 2003.

Jahnen A, Optimage program homepages, http://santec.tudor.lu/projects/optimage, 2004

Jansen JTM and Zoetelief J, Computer aided assessment of image quality for mammography using a contrast detail phantom, Rad. Prot. Dosim. 90, 181–184, 2000.

Jennings RJ, Eastgate RJ, Siedband MP and Ergun DL, Optimal x-ray spectra for screen-film mammography, Med. Phys. 8, 629–639, 1981.

Kotre CJ, The effect of background structure on the detection of low contrast objects in mammography, Br. J. Radiol. 71, 1162–1167, 1998.

Kotre CJ and Marshall NW, A review of image quality and dose issues in digital fluorography and digital subtraction angiography, Rad. Prot. Dosim. 94, 73–76, 2001.

Kotre CJ and Guibelalde E, Optimisation of variable temporal averaging in digital fluoroscopy, Br. J. Radiol. 77, 675–678, 2004.

Kotre CJ, Charlton S, Robson KJ, Birch IP, Willis SP and Thornley M, Application of low dose rate pulsed fluoroscopy in cardiac pacing and electrophysiology: patient dose and image quality implications, Br. J. Radiol. 77, 597–599, 2004.

Krupinski EA, Practical applications of perceptual research, in *Handbook of Medical Imaging, Vol 1: Medical Physics and Psychophysics*, Beutel J, Kundel H, Van Metter R (eds.), SPIE, Bellingham 2000.

Kume Y, Doi K, Ohara K and Giger ML, Investigation of basic imaging properties in digital radiography. 10. Structure mottle of II-TV digital imaging systems, Med. Phys. 13, 843–849, 1986.

Kundel HL, Nodine CF, Thickman D, Carmody D and Lawrence T, Nodule detection with and without a chest image, Invest. Radiol. 20, 25–29, 1985.

Kwan ALC, Filipow LJ and Le LH, Automatic quantitative low contrast analysis of digital chest phantom radiographs, Med. Phys. 30, 312–320, 2003.

Kyprianou IS, Rudin S, Bednarek DR and Hoffmann KR, Generalizing the MTF and DQE to include x-ray scatter and focal spot unsharpness: application to a new microangiographic system, Med. Phys. 32, 613–626, 2005.

Launders JH, McArdle S, Workman A and Cowen AR, Update on the recommended viewing protocol for FAXIL threshold contrast detail detectability test objects used in television fluoroscopy, Br. J. Radiol. 68, 70–77, 1995.

Leitz WK, Månsson LG, Hedberg BRK and Kheddache S, In search of optimum chest radiography techniques, Br. J. Radiol. 66, 314–321, 1993.

Loo L-ND, Doi K, Ishida M and Metz CE, An empirical investigation of variability in contrast-detail measurements, Proc SPIE 419, 68–76, 1983.

Loo L-ND, Doi K, and Metz CE, A comparison of physical image quality indices and observer performance in the radiographic detection of nylon beads, Phys. Med. Biol. 29, 837–856, 1984.

Maccia C, Moores BM and Wall BF, The 1991 CEC trial on quality criteria for diagnostic radiographic images: detailed results and findings. EUR 16635 EN, CEC, Luxembourg 1997.

Manning DJ, Ethell SC and Donovan T, Detection or decision errors? Missed lung cancer from the posteroanterior chest radiograph, Br. J. Radiol. 77, 231–235, 2004.

Marshall NW, The practical application of signal detection theory to image quality assessment in x-ray image intensifier-TV fluoroscopy, Phys. Med. Biol. 46, 1631–1649, 2001.

Marshall NW, A comparison between objective and subjective image quality measurements for a full field digital mammography system, Phys. Med. Biol. 51, 2441–2463, 2006.

Marshall NW and Kotre CJ, Measurement and correction of the effects of lag on contrast-detail test results in fluoroscopy, Phys. Med. Biol. 47, 947–960, 2002.

Marshall NW, Faulkner K, Kotre CJ and Robson K, Analysis of variations in contrast-detail measurements performed on image intensifier-television systems, Phys. Med. Biol. 37, 2297–2302, 1992.

Marshall NW, Kotre CJ, Robson KJ and Lecomber AR, Receptor dose in digital fluorography: a comparison between theory and practice, Phys. Med. Biol. 46, 1283–1296, 2001.

Martens J-B and Meesters L, Image dissimilarity, Signal Processing 70, 155–176, 1998.

Martin CJ, Sharp PF and Sutton DG, Measurement of image quality in diagnostic radiology, Applied Radiation and Isotopes 50, 21–38, 1999a.

Martin CJ, Sutton DG and Sharp PF, Balancing patient dose and image quality, Applied Radiation and Isotopes 50, 1–19, 1999b.

McVey G, Sandborg M, Dance DR and Alm Carlsson G, A study and optimization of lumbar spine X-ray imaging systems, Br. J. Radiol. 76, 177–188, 2003.

Metz CE, Fundamental ROC analysis, in *Handbook of Medical Imaging, Vol 1: Medical Physics and Psychophysics*, Beutel J, Kundel H, Van Metter R (eds.), SPIE, Bellingham 2000.

Metz CE, Wagner RF, Doi K, Brown DG, Nishikawa RM, Myers KJ, Toward consensus on quantitative assessment of medical imaging systems, Med. Phys. 22, 1057–1061, 1995.

Moeckli R, Verdun FR, Fiedler S, Pachoud M, Schnyder P and Valley J-F, Objective comparison of image quality and dose between conventional and synchrotron radiation mammography, Phys. Med. Biol. 45, 3509–3523, 2000.

Moy JP, Signal-to-noise ratio and spatial resolution in x-ray electronic imagers: is the MTF a relevant parameter?, Med. Phys. 27, 86–93, 2000.

Myers KJ, Ideal observer models of visual signal detection, in *Handbook of Medical Imaging, Vol 1: Medical Physics and Psychophysics*, Beutel J, Kundel H, Van Metter R (eds.), SPIE, Bellingham 2000.

Myers KJ and Barrett HH, Addition of a channel mechanism to the ideal-observer model, J. Opt. Soc. Am. A 4, 2447–2457, 1987.

Myers KJ, Rolland JP, Barrett HH and Wagner RF, Aperture optimization for emission imaging: effect of a spatially varying background, J. Opt. Soc. Am. A 7,1279–1293, 1990.

Månsson LG, Methods for the evaluation of image quality: a review, Rad. Prot. Dosim. 90, 89–99, 2000.

Månsson LG, Kheddache S, Lanhede B and Tylén U, Image quality for five modern chest radiography techniques: a modified FROC study with an anthropomorphic chest phantom, Eur. Radiol. 9, 1826–1834, 1999.

Månsson LG, Båth M and Mattson S, Priorities in optimisation of medical x-ray imaging–a contribution to the debate, Rad. Prot. Dosim. 114, 298–302, 2005.

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 data set, Med. Phys. 31, 2205–2211, 2004.

Nodine CF and Mello-Thoms C, The nature of expertise in radiology, in *Handbook of Medical Imaging, Vol 1: Medical Physics and Psychophysics*, Beutel J, Kundel H, Van Metter R (eds.), SPIE, Bellingham 2000.

Offiah AC and Hall CM, Evaluation of the Commission of the European Communities quality criteria for the paediatric lateral spine, Br. J. Radiol. 76, 885–890, 2003.

Pascoal A, Lawinski CP, Honey I and Blake P, Evaluation of a software package for automated quality assessment of contrast detail images–comparison with subjective visual assessment, Phys. Med. Biol. 50, 5743–5757, 2005.

Pilgram TK, Vannier MW, Hildebolt CF, Marsh JL, McAlister WH, Shackelford GD, Offutt CJ and Knapp RH, Craniosynostosis: Image quality, confidence, and correctness in diagnosis, Radiology 173, 675–679, 1989.

Pineda AR and Barrett HH, What does DQE say about lesion detectability in digital radiography?, Proc. SPIE 4320, 561–569, 2001.

Pons AM, Malo J, Artigas JM, Capilla P, Image quality metric based on multidimensional contrast perception models, Displays 20, 93–110, 1999.

Rampado O, Isoardi P and Ropolo R, Quantitative assessment of computed radiography quality control parameters, Phys. Med. Biol. 51, 1577–1593, 2006.

Redlich U, Hoeschen C and Doehring W, Assessment and optimisation of the image quality of chest-radiography systems, Rad. Prot. Dosim. 114, 264–268, 2005.

Rohaly AM, Libert J, Corriveau P, Webster A (eds.), Final report from the video quality experts group on the validation of objective models of video quality assessment, March 2000. (Available on the web-pages of the Video Quality Experts Group http://www-ext.crc.ca/vqeg.).

Ruttimann UE and Webber RL, A simple model combining quantum noise and anatomical variation in radiographs, Med. Phys. 11, 50–60, 1983.

Samei E, Flynn MJ and Eyler WR, Detection of subtle lung nodules: relative influence of quantum and anatomic noise on chest radiographs, Radiology 213, 727–734, 1999.

Samei E, Eyler W and Baron L, Effects of anatomical structure on signal detection, in *Handbook of Medical Imaging, Vol 1: Medical Physics and Psychophysics*, Beutel J, Kundel H, Van Metter R (eds.), SPIE, Bellingham 2000.

Samei E, Flynn MJ Peterson E and Eyler WR, Subtle lung nodules: influence of local anatomic variations on detection, Radiology 228, 76–84, 2003.

Samei E, Dobbins JT III, Lo JY and Tornai MP, A framework for optimising the radiographic technique in digital x-ray imaging, Rad. Prot. Dosim. 114, 220–229, 2005.

Samei E, Ranger NT, Dobbins JT III and Chen Y, Intercomparison of methods for image quality characterization. I. Modulation transfer function, Med. Phys. 33, 1454–1465, 2006.

Sandborg M, Tingberg A, Dance DR, Lanhede B, Almén A, McVey G, Sund P, Kheddache S, Besjakov J, Mattson S, Månsson LG and Alm Carlsson G, Demonstration of correlations between clinical and physical image quality measures in chest and lumbar spine screen-film radiography, Br. J. Radiol. 74, 520–528, 2001a.

Sandborg M, McVey G, Dance DR and Alm Carlsson G, Schemes for the optimization of chest radiography using a computer model of the patient and x-ray imaging system, Med. Phys. 28, 2007–2019, 2001b.

Sanfridsson J, Holje G, Svahn G, Ryd L and Jonsson K, Radiation dose and image information in computed radiography. A phantom study of angle measurements in the weight-bearing knee, Acta Radiol. 39, 642, 1999.

Schilz C, JAnalyser-CDRAD quality testing of digital projection radiography, Appendix 2 in Busch HP, DIMOND III Final Report Image quality and dose management for digital radiography. 2004a.
Available at http://www.dimond3.org/Reports/WP%201/part_j_Annex2.pdf

Schilz C, JAnalyser constancy testing of digital projection radiography, Appendix 3 in Busch HP, DIMOND III Final Report Image quality and dose management for digital radiography, 2004b.
Available at http://www.dimond3.org/Reports/WP%201/part_k_Annex3.pdf

Schreiner-Karoussou A, Review of image quality standards to control digital X-ray systems, Rad. Prot. Dosim. 117, 23–25, 2005.

Siewerdsen JH, Cunningham IA and Jaffray DA, A framework for noise-power spectrum analysis of multidimensional images, Med. Phys. 29, 2655–2671, 2002.

Sorenson JA, Nelson JA, Niklason LA and Klauber MR, Simulation of lung nodules for dodule detection studies, Invest. Radiol. 15, 490–495, 1980.

Sund P, Båth M, Kheddache S and Månsson LG, 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.

Tapiovaara MJ, SNR and noise measurements for medical imaging: II. Application to fluoroscopic x-ray equipment, Phys. Med. Biol. 38, 1761–1788, 1993.

Tapiovaara MJ, Efficiency of low-contrast detail detectability in fluoroscopic imaging, Med. Phys. 24, 655–664, 1997.

Tapiovaara M, Objective measurement of image quality in fluoroscopic x-ray eaquipment: FluoroQuality, Report STUK-A196, Radiation and Nuclear Safety Authority, Helsinki, 2003.
Available at http://www.stuk.fi/julkaisut/stuk-a/stuk-a196.pdf

Tapiovaara MJ, Image quality measurements in radiology, Rad. Prot. Dosim. 117, 116–119, 2005.

Tapiovaara MJ and Wagner RF, SNR and DQE analysis of broad spectrum x-ray imaging, Phys. Med. Biol. 30, 519–529, 1985 (Corrigendum : Phys. Med. Biol. 31, 195).

Tapiovaara MJ and Wagner RF, SNR and noise measurements for medical imaging: I. A practical approach based on statistical decision theory, Phys. Med. Biol. 38, 71–92, 1993.

Tapiovaara MJ and Sandborg M, Evaluation of image quality in fluoroscopy by measurements and Monte Carlo calculations, Phys. Med. Biol. 40, 589–607, 1995.

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

Thornbury JR, Fryback DG, Patterson FE and Chiavarini RL, A methodology for comparison of quality of radiologic images from different screen/film combinations based on radiologists' subjective judgments, Proc. SPIE 127, 24–29, 1977.

Tingberg A and Sjöström D, Optimisation of image plate radiography with respect to tube voltage, Rad. Prot. Dosim. 114, 286–293, 2005.

Tingberg A, Herrmann C, Lanhede B, Almén A, Besjakov J, Mattson S, Sund P, Kheddache S and Månsson LG, Comparison of two methods for evaluation of the image quality of lumbar spine radiographs, Rad. Prot. Dosim. 90, 165–168, 2000.

Tingberg A, Herrmann C, Besjakov J, Almén A, Sund P, Adliene D, Mattson S, Månsson LG and Panzer W, What is worse: decreased spatial resolution or increased noise?, Proc. SPIE 4686, 338–346, 2002.

Tingberg A, Herrmann C, Lanhede B, Almén A, Sandborg M, Mc Vey G, Mattson S, Panzer W, Besjakov J, Månsson LG, Kheddache S, Alm Carlsson G, Dance DR, Tylén U and Zankl M, Influence of the characteristic curve on the clinical image quality of lumbar spine and chest radiographs, Br. J. Radiol. 77, 204–215, 2004.

Tingberg A, Båth M, Håkansson M, Medin J, Besjakov J, Sandborg M, Alm-Carlsson G, Mattson S and Månsson LG, Evaluation of image quality of lumbar spine images: a comparison between FFE and VGA, Rad. Prot. Dosim. 114, 53–61, 2005a.

Tingberg A, Eriksson F, Medin J, Besjakov J, Båth M, Håkansson M, Sandborg M, Almén A, Lanhede B, Alm-Carlsson G, Mattson S and Månsson LG, Inter-observer variation in masked and unmasked images for quality evaluation of clinical radiographs, Rad. Prot. Dosim. 114, 62–68, 2005b.

Tischenko O, Hoeschen C, Effenberger O, Reissberg S, Buhr E and Döhring W, Measurement of the noise components in the medical X-ray intensity pattern due to overlaying, nonrecognizable structures, Proc. SPIE 5030, 422–432, 2003.

Uffmann M, Neitzel U, Prokop M, Kabalan N, Weber M, Herold CJ and Schaefer-Prokop C, Flat-panel-detector chest radiography: effect of tube voltage on image quality, Radiology 235, 642–650, 2005.

Ullman G, Sandborg M, Tingberg A, Dance DR, Hunt R and Alm Carlsson G, Comparison of clinical and physical measures of image quality in chest PA and pelvis AP views at varying tube voltages, Report 98, Linköping University, 2004. Available at http://huweb.hu.liu.se/inst/imv/radiofysik/pdfs/Rep98.pdf

Ullman G, Malusek A, Sandborg M, Dance DR and Alm Carlsson G, Calculation of images from an anthropomorphic chest phantom using Monte Carlo methods, Proc. SPIE 6142, 2006.

van Engen R, Young K, Bosmans H, Thijssen M, Addendum on digital mammography. European guidelines for quality assurance in mammography screening, Third edition, 2003.

Vañó E, Guibelalde E, Morillo A, Alvarez-Pedrosa CS and Fernández JM, Evaluation of the European image quality criteria for chest examinations, Br. J. Radiol. 68, 1349–1355, 1995.

Vañó E, Gonzales L and Oliete S, The relevance of quality criteria for optimisation in conventional radiology, Rad. Prot. Dosim. 80, 39–44, 1998.

Vassileva J, A phantom approach to find the optimal technical parameters for plain chest radiography, Br. J. Radiol. 77, 648–653, 2004.

Venema HW, van Straten M and den Heeten GJ, Digital radiography of the chest: reassessment of the high-voltage technique? Letters to the editor. Radiology 235, 336–337, 2005.

Verdun FR, Bochud F, Depeursinge C, Desponds L, Grecescu M, Hessler C, Raimondi S and Valley J-F, Subjective and objective evaluation of chest imaging systems, Rad. Prot. Dosim. 49, 91–94, 1993.

Verdun FR, Moeckli R, Valley J-F, Bochud F and Hessler C, Survey on image quality and dose levels used in Europe for mammography, Br. J. Radiol 69, 762–768, 1996.

Vucich JJ, The role of anatomic criteria in the evaluation of radiographic images. In: *The physics of Medical Imaging: Recording system measurements and techniques. Medical Physics Monograph No. 3*, Haus AG, editor. American Association of Physicists in Medicine, 1979.

Wagner RF, The laboratory/clinical interface in image evaluation, SPIE 127, 2, 1977.

Wagner RF, Low-contrast sensitivity of radiologic, CT, nuclear medicine, and ultrasound medical imaging systems, IEEE Transactions on Medical Imaging MI-2, 105–121, 1983.

Wagner RF, Characteristic images emerging from recent SPIE medical image symposia, Proc. SPIE 767, 138–141, 1987.

Wagner RF and Brown DG, Unified SNR analysis of medical imaging systems, Phys. Med. Biol. 30, 489–518, 1985.

Wagner RF, Myers KJ, Tapiovaara MJ, Brown DG, Burgess AE, Maximum a posteriori detection and figures of merit for detection under uncertainty, Proc. SPIE 1231, 195–204, 1990.

Wagner RF, Beiden SV and Campbell G, Multiple-reader studies, digital mammography, computer-aided diagnosis – and the Holy Grail of imaging physics (I), Proc. SPIE 4320, 611–618, 2001.

Walsh C, Dowling A, Meade A and Malone J, Subjective and objective measures of image quality in digital fluoroscopy, Rad. Prot. Dosim. 117, 34–37, 2005.

Wang X, Van Metter RL, Foos DH and Steklenski D, Comprehensive and automated image quality performance measurement of computed radiography systems, Proc. SPIE 4320, 308–315, 2001.

Wang Z, Bovik AC, Sheikh HR and Simoncelli EP, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Processing 13, 1–14, 2004.

Whalen AD, *Detection of Signals in Noise*, Academic Press, New York, 1971.

Wilson DL, Jabri KN and Manjeshwar RM, Quantitative image quality studies and the design of x-ray fluoroscopy systems, in *Handbook of Medical Imaging, Vol 1: Medical Physics and Psychophysics*, Beutel J, Kundel H, Van Metter R (eds.), SPIE, Bellingham 2000.

Winkler S, Issues in vision modeling for perceptual video quality assessment, Signal Processing 78, 231–252, 1999.

STUK-A reports/STUK-A-sarjan julkaisuja

STUK-A219 Tapiovaara M. Relationships between physical measuments and user evaluation of image quality in medical radiology – a review. Helsinki 2006.

STUK-A218 Ikäheimonen TK, Klemola S, Ilus E, Varti V-P, Mattila J. Monitoring or radionuclides in the vicinities of Finnish nuclear power plants in 1999–2001. Helsinki 2006.

STUK-A217 Ikäheimonen TK (toim.). Ympäristön radioaktiivisuus Suomessa – 20 vuotta Tshernobylista. Symposium Helsingissä 25.–26.4.2006. Helsinki 2006.

STUK-A216 Pastila R. Effect of long-wave UV radiation on mouse melanoma: An in vitro and in vivo study. Helsinki 2006.

STUK-A215 Rantavaara A. Elintarvikeketjun suojaustoimenpiteet laskeumatilanteiden varalle. Helsinki 2005.

STUK-A214 Sinkko K, Ammann M, Hämäläinen RP, Mustajoki J. Facilitated workshop on clean-up actions in inhabited areas in Finland after an accidental release of radionuclides. Helsinki 2005.

STUK-A213 Vesterbacka P. ^{238}U-series radionuclides in Finnish groundwater-based drinking water and effective doses. Helsinki 2005.

STUK-A212 Kantala T. Elintarviketeollisuuslaitosten ja niiden ympäristön puhdistustoimenpiteet säteilytilanteessa. Helsinki 2005.

STUK-A211 Muikku M, Arvela H, Järvinen H, Korpela H, Kostiainen E, Mäkeläinen I, Vartiainen E, Vesterbacka K. Annoskakku 2004 – suomalaisten keskimääräinen efektiivinen annos. Helsinki 2005.

STUK-A210 Salomaa S, Ikäheimonen TK (eds.). Research activities of STUK 2000–2004. Helsinki 2005.

STUK-A209 Valmari T, Rantavaara A, Hänninen R. Radioaktiivisten aineiden siirtyminen päästöpilven kulkeutumisen aikana tuotettaviin elintarvikkeisiin. Helsinki 2004.

STUK-A208 Kiuru A. Molecular biology methods in assessing radiation-induced hereditary risks in humans. Helsinki 2004.

STUK-A207 Sinkko K. Nuclear emergency response planning based on participatory decision analytic approaches. Helsinki 2004.

STUK-A206 Hämäläinen K, Vesterbacka P, Mäkeläinen I, Arvela H. Vesilaitosten vedenkäsittelyn vaikutus luonnon radionuklidipitoisuuksiin (VEERA). Helsinki 2004.

STUK-A205 Klemola S, Ilus E, Ikäheimonen TK. Monitoring of radionuclides in the vicinities of Finnish nuclear power plants in 1997 and 1998. Helsinki 2004.

STUK-A204 Kettunen A. Radiation dose and radiation risk to foetuses and newborns during x-ray examinations. Helsinki 2004.

STUK-A203 Rahola T, Etherington G, Bérard P, Le Guen B, Hurtgen C, Muikku M, Pusa S. Survey of Internal Dose Monitoring Programmes for Radiation Workers. WP 1 in the project OMINEX (Optimisation of Monitoring for Internal Exposure). Helsinki 2003.

STUK-A202 Salomaa S (ed.). Research projects of STUK 2003–2005. Helsinki 2004.

STUK-A201 Mäkeläinen I (toim.). Säteilyn ja kemiallisten aineiden riskifilosofiat ja suojeluperusteet. Helsinki 2003.

STUK-A200 Vetikko V, Valmari T, Oksanen M, Rantavaara A, Klemola S, Hänninen R. Energiateollisuudessa syntyvän puuntuhkan radioaktiivisuus ja sen säteilyvaikutukset. Helsinki 2004.

STUK-A199 Vesterbacka P, Mäkeläinen I, Tarvainen T, Hatakka T, Arvela H. Kaivoveden luonnollinen radioaktiivisuus – otantatutkimus 2001. Helsinki 2004.

STUK-A198 Eloranta E. Geofysiikan kenttäteoria. Helsinki 2003.

STUK-A197 Vesterbacka P, Turtiainen T, Hämäläinen K, Salonen L, Arvela H. Talousveden radionuklidien poisto. Helsinki 2003.

STUK-A196 Tapiovaara M. Objective Measurement of Image Quality in Fluoroscopic X-ray Equipment: FluoroQuality. Helsinki 2003.

STUK-A195 Paile W (ed.). Radiation Protection in the 2000s – Theory and Practice. Nordic Society for Radiation Protection. Proceedings of the XIII ordinary meeting, Turku/Åbo, Finland, August 25–29, 2002. Helsinki 2003.

STUK-A194 Ikäheimonen TK. Determination of transuranic elements, their behaviour and sources in the aquatic environment. Helsinki 2003.

A full list of publications is available from

Radiation and Nuclear Safety Authority (STUK)
P.O.Box 14, FI-00881 Helsinki
FINLAND
Tel +358 9 759 881