Radiation Protection Dosimetry (2008), Vol. 129, No. 1–3, pp. 244–248
Advance Access publication 5 February 2008
doi:10.1093/rpd/ncn009

REVIEW OF RELATIONSHIPS BETWEEN PHYSICAL MEASUREMENTS AND USER EVALUATION OF IMAGE QUALITY

M. J. Tapiovaara*
STUK – Radiation and Nuclear Safety Authority, PO Box 14, 00881 Helsinki, Finland

Some of the findings of a review of the relationship between physical measurements and clinical image quality have been summarised. Mixed results were found: some studies had no relationship at presently typical dose levels, whereas others had a clear correlation between them. It is concluded that the various image quality evaluation tasks in an X-ray department are best done by different methods. Presently, exact physical measurements cannot supersede subjective evaluation in judging the acceptability of clinical images, whereas they are indispensable in specification and testing of technical performance.

INTRODUCTION

There are many tasks in radiology departments that involve the assessment of image quality: (1) equipment purchasing is partly based on performance specifications, (2) acceptance testing verifies that the system fulfils the specified performance criteria, (3) constancy testing attempts to notice any changes in the imaging system, (4) clinical testing concentrates on the fulfilment of clinical needs and (5) optimisation attempts 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 physical (or technical) image quality or clinical image quality, according to the method used in the evaluation. It would be desirable to understand the relationship between these various evaluations and the usefulness of various assessment methodologies for the various purposes. A literature survey[1] was conducted within the SENTINEL project to clarify what has been reported on these matters. There were 184 papers included in the review, with a majority of them published in the last six years. This paper summarises some of the review's findings on the relationship between various image quality assessment methods.

MEASUREMENT OF PHYSICAL IMAGE QUALITY

The most fundamental factors of physical image quality are image sharpness, noise and contrast. Their assessment has been classically based on visual observation of images of test objects, such as resolution gratings, low-contrast details or contrast-detail phantoms. All these methods suffer from the subjectivity of the evaluation: there is no testing of the critical confidence level of the observer and this level may vary between various observers and between various observations made by a given observer. An improvement to this may be achieved in digital imaging where the image data are readily available for computer analysis[1].

In field work, image sharpness is commonly measured by observing a resolution test object image. Although useful for many purposes, this measurement gives an incomplete description of sharpness because it only considers the highest spatial frequency that is resolved. No emphasis is given to signal transfer at lower frequencies—which are known to be important in clinical images. This situation can be improved by measuring the modulation transfer function (MTF) of the imaging system.

Somewhat similarly, image noise is often measured in field work as the standard deviation of pixel values. Although useful for many purposes, this measurement suffers from not taking into account the fact that noise is often correlated within nearby pixels. When the image noise is stationary, it can be analysed properly by measuring the noise power spectrum (Wiener spectrum, NPS).

The large-area signal transfer, MTF and NPS can be combined with a given detail model to calculate the ideal observer's signal-to-noise ratio (SNR) for detecting that signal. This quantity depicts the best possible detection performance for the given detail. The large-area signal transfer, MTF and NPS can also be combined with the actual number of X-ray quanta that are used in forming the image to calculate the detective quantum efficiency (DQE). It expresses the ability of an image receptor to transfer information to the image from the radiation beam impinging on the image receptor. Currently, DQE is considered as the best way to specify the performance of the image receptor. In principle, the above concepts can also be generalised to fluoroscopic

*Corresponding author: markku.tapiovaara@stuk.fi

imaging by including the temporal frequency in the analysis.

These exact measures of physical image quality have been thoroughly discussed in many books on the subject[2–5]. However, these measures consider only the image formation stage. The display stage needs a separate evaluation. The less-exact visual image quality evaluations (resolution, threshold contrast, contrast-detail performance) take into account the whole imaging chain.

There are many studies concerning the relationship between human performance and the performance of the ideal observer and other related computational observers in detecting exactly known signals embedded in noise[3,5–9]. This detection task, the signal-known-exactly/background-known-exactly (SKE/BKE) task, corresponds closely to that of flat-background test phantoms. The general finding in these experiments is that the performance of human observers can be well predicted from the performance of these observers. 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, or if the image noise is strongly coloured.

EVALUATION OF CLINICAL IMAGE QUALITY

Most often clinical image quality just refers to a subjective judgement of quality in the clinical radiographs or the fluoroscopic image. Such an impression-based assessment does not necessarily relate to clinical usefulness and, therefore, its utility has been seriously questioned[3,10,11]. The situation is improved if the evaluation focuses on the visibility of specified details, but even then the subjectivity of the evaluation may leave notable variability and bias in the results[3,5,12]. The sources of bias range from preference for the aesthetically pleasing to prejudice against the unfamiliar[5]. According to Krupinski[12], 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. For discussions on the methodology and use of subjective evaluation methods, see also the works of Dobbins, Börjesson *et al.* and Tingberg *et al.*[13–16]

The most meaningful method of evaluating clinical image quality is to actually measure the usefulness of the image for the intended diagnostic task by the receiver-operating characteristic (ROC) or multiple-alternative forced-choice (M-AFC) method[5,17,18]. However, these methods require a large number of both normal and abnormal verified patient images and are therefore often impossible to be used in everyday work. Digital imaging may offer some relief in obtaining the required image set, if normal patient images can be manipulated to include artificially added realistic-appearing details, but even then the required observer time is prohibitive because a large number of images need be observed in order to obtain precise results.

In spite of its weaknesses, subjective evaluation of image quality is of course useful—and often it is also the only practical alternative for assessing clinical image quality[5,13]: some kind of verification of the clinical acceptability of patient images is anyway necessary[13,19].

RELATIONSHIPS BETWEEN PHYSICAL AND CLINICAL IMAGE QUALITY ASSESSMENTS

It is more difficult to detect details against radiographic backgrounds of patients than against the uniform background of homogeneous phantoms[20,21]. Anatomical background complicates search operations by making the scene busy and full of visible structures, and even if the possible signal location is known, the signal may be masked or mimicked by overlying anatomical background. These are obvious factors deteriorating human performance. However, they do not affect the ideal observer in the SKE/BKE task and are not considered within the computational signal-detection theoretic (SDT) approach. 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 also suffer from the variability of the signal and the background. However, the statistics of these variabilities are not known and ideal observer performance in such tasks cannot be calculated.

The mechanisms of how anatomical background decreases human observers' detail detectability are not well understood. One effect is that 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[22]. Håkansson *et al.*[23] 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. This distinction between anatomical noise and background is not always made, and the disturbing anatomic background variability is often also called anatomical noise. Also others have found it to deteriorate detectability to a much higher degree than actual stochastic noise in typical cases[24–28].

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 considered 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[24,26,28,29]. The background may act partly as random noise and partly as a recognisable structure in detecting mass objects. 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[28–30].

Sandborg *et al.*[31,32] have 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[31] and in digital chest and pelvis radiography[32]. They found significant correlation between blood vessel contrast and subjective evaluations in film-screen based 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 limited more by contrast than by noise. In film-screen-based lumbar spine imaging, the best predictors of clinical image quality were the contrast and SNR of small soft tissue cavities in bone (trabecular structures). In digital imaging the SNR measures were found to be strongly related to the radiologists' grading of the images.

CONCLUSIONS

Generally, it seems clear that physical image quality is monotonically related to clinical performance by enabling the detection and identification of the necessary details. There can be two regions in X-ray imaging thought to exist, characterised by asymptotes corresponding to either quantum-limited or quantum-saturated imaging. In the former, detection is mainly limited by the system noise (quantum and electronic noise) and can be improved by, e.g. 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. Then, simple test phantoms and physical assessment methods are not sufficient to optimise the noise level in clinical X-ray images when anatomical background is an issue[11]. This is not always the case; for some features detectability is mainly limited by random noise[26]. Anatomical background may not be as disturbing in all projection radiography as it is in mammography and chest imaging, where anatomical background is remarkable. 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.

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 for image quality assessment one must generally be content with subjective, opinion-based evaluations instead of a truly quantitative measurement. Then, the significance to actual clinical performance is often left unclear.

Various ways of assessing image quality—in the clinical, technical and physical sense of the concept—have been discussed and studies of the relationships between various assessment results have been reviewed[1]. In the review it was seen that the relationship between the SDT-based image quality measures and the performance of human observers in simple SKE/BKE 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. 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 be easily done 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 deterioration cannot be handled by simple corrective actions, but requires expensive investment in equipment, it may be more reasonable to make decisions based on some kind of clinical evaluation than by simplistic limits of measured parameters. Establishing the relationship

between technical image quality parameters and clinical performance has proved to be difficult, at least. 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 anatomical 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); here, SNR denotes the ideal observer's SNR and D the dose in the patient. The quantity D can be chosen from a variety of dose quantities (e.g. entrance skin dose or effective dose) according to the optimisation strategy chosen. If resolution-related things are not of interest, one may even use the more simple contrast-to-noise ratio (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. However, Månsson *et al.*[33] and Båth *et al.*[34] criticise 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; therefore, optimisation studies need be done with actual patient images or with 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[35] 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[36]. Although such data are not directly related to clinical requirements, they should be useful for indicating typical and/or acceptable X-ray system performance, much in analogy with the approach using diagnostic reference doses. 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 using SDT-based computational observers instead of humans, but the display stage then needs separate consideration.

ACKNOWLEDGEMENT

This paper has been prepared as part of the SENTINEL project.

FUNDING

The SENTINEL project, contract FP6-012909, was partially supported and received funding from the EC-Euratom Sixth Framework Programme.

REFERENCES

1. Tapiovaara, M. *Relationships between physical measurements and user evaluation of image quality in medical radiology—a review.* Report STUK-A219 (Helsinki: STUK—Radiation and Nuclear Safety Authority) (2006). Available on http://www.stuk.fi/julkaisut/stuk-a/stuk-a219.pdf
2. Beutel, J., Kundel, H. and Van Metter, R., Eds. *Handbook of Medical Imaging, Vol 1: Medical Physics and Psychophysics* (Bellingham: SPIE) (2000).
3. Barrett, H. H. and Myers, K. *Foundations of Image Science* (Hoboken: John Wiley and Sons) (2004).
4. Dainty, J. C. and Shaw, R. *Image Science* (London: Academic Press) (1974).
5. ICRU. *Medical imaging—the assessment of image quality.* Report 54 (Bethesda: International Commission on Radiation Units and Measurements) (1996).
6. Tapiovaara, M. J. and Sandborg, M. *How should low-contrast detail detectability be measured in fluoroscopy*? Med. Phys. **31**, 2564–2576 (2004).
7. Marshall, N. W. *A comparison between objective and subjective image quality measurements for a full field digital mammography system*. Phys. Med. Biol. **51**, 2441–2463 (2006).
8. Myers, K. J. *Ideal observer models of visual signal detection*. In: Handbook of Medical Imaging, Vol 1: Medical Physics and Psychophysics, Beutel, J., Kundel, H. and Van Metter, R., Eds. (Bellingham: SPIE) (2000).
9. Abbey, C. K. and Bochud, F. O. *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. and Van Metter, R., Eds. (Bellingham: SPIE) (2000).
10. Vucich, J. J. *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, A. G. Ed. (New York: American Association of Physicists in Medicine) (1979).
11. Månsson, L. G. *Methods for the evaluation of image quality: a review.* Rad. Prot. Dosim. **90**, 89–99 (2000).
12. Krupinski, E. A. *Practical applications of perceptual research*. In: Handbook of Medical Imaging, Vol 1: Medical Physics and Psychophysics, Beutel, J., Kundel,

H. and Van Metter, R., Eds. (Bellingham: SPIE) (2000).
13. Dobbins, J. T. III. *Image quality metrics for digital systems.* In: Handbook of Medical Imaging, Vol 1: Medical Physics and Psychophysics, Beutel, J., Kundel, H. and Van Metter, R., Eds. (Bellingham: SPIE) (2000).
14. Börjesson, S., et al. *A software tool for increased efficiency in observer performance studies in radiology.* Rad. Prot. Dosim. **114**, 45–52 (2005).
15. Tingberg, A., et al. *Influence of the characteristic curve on the clinical image quality of lumbar spine and chest radiographs.* Br. J. Radiol. **77**, 204–215 (2004).
16. Tingberg, A., Båth, M., Håkansson, M., Medin, J., Besjakov, J., Sandborg, M., Alm-Carlsson, G., Mattson, S. and Månsson, L. G. *Evaluation of image quality of lumbar spine images: a comparison between FFE and VGA*. Rad. Prot. Dosim. **114**, 53–61 (2005).
17. Metz, C. E. *Fundamental ROC analysis.* In: Handbook of Medical Imaging, Vol 1: Medical Physics and Psychophysics, Beutel, J., Kundel, H. and Van Metter, R., Eds. (Bellingham: SPIE) (2000).
18. Burgess, A. E. *Comparison of receiver operating characteristic and forced choice observer performance methods.* Med. Phys. **22**, 643–655 (1995).
19. ICRU. *Image quality in chest radiography.* Report 70. J. ICRU **3**(2), 1–165 (2003).
20. Kundel, H. L., Nodine, C. F., Thickman, D., Carmody, D. and Lawrence, T. *Nodule detection with and without a chest image.* Invest. Radiol. **20**, 25–29 (1985).
21. Samei, E., Flynn, M. J., Peterson, E. and Eyler, W. R. *Subtle lung nodules: influence of local anatomic variations on detection*. Radiology **228**, 76–84 (2003).
22. 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).
23. Håkansson, M., Båth, M., Börjesson, S., Kheddache, S., Grahn, A., Ruschin, M., Tingberg, A., Mattson, S. and Månsson, L. G. *Nodule detection in digital chest radiography: summary of the RADIUS chest trial*. Rad. Prot. Dosim. **114**, 114–120 (2005).
24. Ruttimann, U. E. and Webber, R. L. *A simple model combining quantum noise and anatomical variation in radiographs.* Med. Phys. **11**, 50–60 (1983).
25. Kotre, C. J. *The effect of background structure on the detection of low contrast objects in mammography.* Br. J. Radiol. **71**, 1162–1167 (1998).
26. Bochud, F. O., Valley, J.-F., Verdun, F. R., Hessler, C. and Schnyder, P. *Estimation of the noisy component of anatomical backgrounds.* Med. Phys. **26**, 1365–1370 (1999).
27. Samei, E., Flynn, M. J. and Eyler, W. R. *Detection of subtle lung nodules: relative influence of quantum and anatomic noise on chest radiographs.* Radiology **213**, 727–734 (1999).
28. Burgess, A. E., Jacobson, F. L. and Judy, P. F. *Human observer detection experiments with mammograms and power-law noise.* Med. Phys. **28**, 419–437 (2001).
29. Båth, M., Håkansson, M., Börjesson, S., Kheddache, S., Grahn, A., Bochud, F. O., Verdun, F. R. and Månsson, L. G. *Nodule detection in digital chest radiography: part of image background acting as pure noise.* Rad. Prot. Dosim. **114**, 102–108 (2005).
30. 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. and Van Metter, R., Eds. (Bellingham: SPIE) (2000).
31. Sandborg, M., et al. *Demonstration of correlations between clinical and physical image quality measures in chest and lumbar spine screen-film radiography.* Br. J. Radiol. **74**, 520–528 (2001).
32. Sandborg, M., Tingberg, A., Ullman, G., Dance, D. R. and Alm Carlsson, G. *Comparison of clinical and physical measures of image quality in chest and pelvis computed radiography at different tube voltages.* Med. Phys. **33**, 4169–4175 (2006).
33. Månsson, L. G., 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).
34. Båth, M., Håkansson, M., Hansson, J. and Månsson, L. G. *A conceptual optimisation strategy for radiography in a digital environment*. Rad. Prot. Dosim. **114**, 230–235 (2005).
35. Busch, H. P. and Faulkner, K. *Image quality and dose management in digital radiography: a new paradigm for optimisation*. Rad. Prot. Dosim. **117**, 143–147 (2005).
36. Evans, D. S., MacKenzie, A., Lawinski, C. P. 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).