Physica Medica 31 (2015) 823–843

Contents lists available at ScienceDirect

Physica Medica

journal homepage: http://www.physicamedica.com

Review Paper

Image quality in CT: From physical measurements to model observers

F.R. Verdun [a,*,1], D. Racine [a,1], J.G. Ott [a], M.J. Tapiovaara [b], P. Toroi [b], F.O. Bochud [a], W.J.H. Veldkamp [c,d], A. Schegerer [e], R.W. Bouwman [d], I. Hernandez Giron [c], N.W. Marshall [f], S. Edyvean [g]

[a] Institute of Radiation Physics, Lausanne University Hospital, 1 Rue du Grand-Pré, 1007 Lausanne, Switzerland
[b] STUK-Radiation and Nuclear Safety Authority, PO Box 14, FIN-00881 Helsinki, Finland
[c] Department of Radiology, Leiden University Medical Center, C2-S, PO Box 9600, 2300RC Leiden, The Netherlands
[d] Dutch reference Centre for Screening (LRCB), Radboud University Medical Centre, PO Box 6873, 6503 GJ Nijmegen, The Netherlands
[e] Department for Radiation Protection and Health External and Internal Dosimetry, Biokinetics, Ingolstädter Landstr, 1, 85764 Neuherberg, Germany
[f] Katholieke Universiteit Leuven, Oude Markt 13, 3000 Leuven, Belgium
[g] Medical Dosimetry Group, Centre for Radiation Chemicals and Environmental Hazards, Public Health England, Didcot, UK

ARTICLE INFO

Article history:
Received 8 May 2015
Received in revised form 4 August 2015
Accepted 23 August 2015
Available online 12 October 2015

Keywords:
Computed tomography
Image quality
Patient dose optimisation
Model observer

ABSTRACT

Evaluation of image quality (IQ) in Computed Tomography (CT) is important to ensure that diagnostic questions are correctly answered, whilst keeping radiation dose to the patient as low as is reasonably possible. The assessment of individual aspects of IQ is already a key component of routine quality control of medical x-ray devices. These values together with standard dose indicators can be used to give rise to 'figures of merit' (FOM) to characterise the dose efficiency of the CT scanners operating in certain modes. The demand for clinically relevant IQ characterisation has naturally increased with the development of CT technology (detectors efficiency, image reconstruction and processing), resulting in the adaptation and evolution of assessment methods. The purpose of this review is to present the spectrum of various methods that have been used to characterise image quality in CT: from objective measurements of physical parameters to clinically task-based approaches (i.e. model observer (MO) approach) including pure human observer approach. When combined together with a dose indicator, a generalised dose efficiency index can be explored in a framework of system and patient dose optimisation. We will focus on the IQ methodologies that are required for dealing with standard reconstruction, but also for iterative reconstruction algorithms. With this concept the previously used FOM will be presented with a proposal to update them in order to make them relevant and up to date with technological progress. The MO that objectively assesses IQ for clinically relevant tasks represents the most promising method in terms of radiologist sensitivity performance and therefore of most relevance in the clinical environment.

Introduction

Diagnostic x-rays contribute to nearly 50% of the total annual collective effective dose of radiations from man-made and natural sources to the general population in western countries; computed tomography (CT) is the largest single source of this medical exposure.

The contribution of CT to collective dose has significantly increased in recent years and a considerable effort is required to control this trend and ensure that the benefits from the use of this technology outweigh the risks [1]. For example, in 2007–2008 the average dose per inhabitant, due to CT, was about 0.8 mSv in France and Switzerland, and about 0.7 mSv in Germany (as part of an average for all x-ray imaging of about 1.2 mSv and 1.7 mSv, respectively) [2–4]. An update of the French and German data showed that in 2012 the contribution of CT exposure had increased to approximately 1.15 mSv, with a similar increase shown in the last Swiss survey performed for 2013 [5].

In this context the radiation protection requirements in diagnostic radiology (justification of the examination and optimisation of the imaging protocol) need to be re-enforced. Justifying a CT scan is a clinical consideration and therefore will not be addressed in this work. However, the optimisation of a CT examination is achieved when image quality enables the clinical question to be answered whilst keeping patient radiation dose as low as reasonably possible. For this purpose the clinical question needs to be formulated as concretely as possible to enable a clear description of the image quality level required. To achieve this, appropriate and clinically

* Corresponding author. Institute of Radiation Physics, Lausanne University Hospital, 1 Rue du Grand-Pré, 1007 Lausanne, Switzerland. Tel.: +41 21 314 82 50; fax: +41 21 314 8299.
E-mail address: Francis.Verdun@chuv.ch (F.R. Verdun).
[1] Both authors contributed equally to this work.

http://dx.doi.org/10.1016/j.ejmp.2015.08.007

relevant image quality parameters and radiation dose indices must be defined, described, and used. This paper concentrates on image quality parameters.

The first step of the optimisation process should ensure that x-ray conversion into image information is performed as efficiently as possible. In projection radiology such as radiology or mammography one can use the DQE (Detective Quantum Efficiency as described in IEC 62220-1/2) as a global figure of merit. Unfortunately, due to the geometry and data processing required for CT, the use of such a quantity is not feasible. In general, one will assess the amount of radiation required to achieve a certain level of image quality. As a surrogate of the radiation received by the detector one uses the standardised CT dose index ($CTDI_{vol}$). This quantity represents the average dose delivered in PMMA phantoms of 16 and 32 cm in diameter and is related to the amount of noise present in an image. According to its definition $CTDI_{vol}$ is different from the actual average dose delivered in a slice of a patient, and the latter should be estimated using the Size Specific Dose Estimator (SSDE) proposed by the AAPM (American Association of Physics in Medicine) [6]. For a given $CTDI_{vol}$ level, image quality parameters are generally assessed using the signal detection theory that considers the imaging system linear and shift invariant.

The next step of the optimisation process should be done with the clinical applications in mind. Direct determination of clinical performance is, however, difficult, expensive, and time-consuming. Furthermore, the results in these studies can be strongly dependent on the patient sample and on the radiologists involved. As an alternative, one can assess image quality using task-oriented image quality criteria. They will necessarily be simplistic in comparison to the clinical situations but make it possible to predict the perception of simple structures within an image. The phantoms available for this type of study remain quite simple whilst trying to mimic important disease-related structures in actual patients. It is likely that 3D printing techniques will improve phantom and task realism in the future [7–9]. To seek optimisation, task-oriented image quality metrics could be studied as a function of $CTDI_{vol}$ or SSDE. Figure 1 summarises this optimisation process.

Part 1 of this review focuses on signal detection theory and summarises the methods used to assess image quality in an objective way. When CT images are reconstructed using the standard filtered back-projection (FBP), these methods are commonly used to characterise a CT unit. The objective image quality metrics assess separate aspects of the features of the image, and therefore need to be combined to give an overall representation of the image quality.

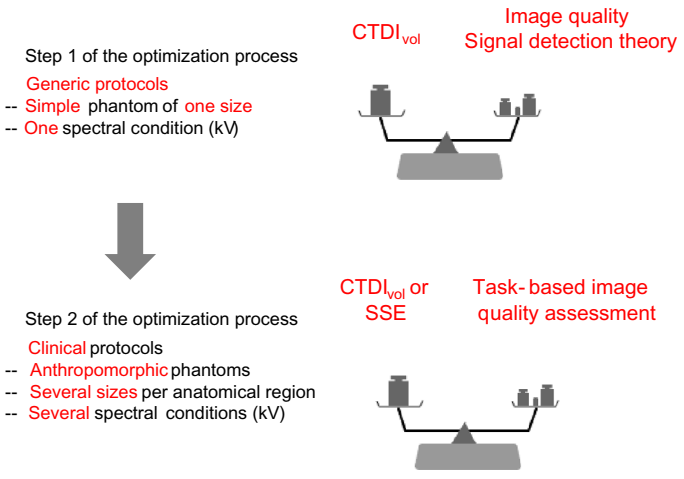

Figure 1. CT optimisation process in two steps: generic acquisition optimisation and clinical protocol optimisation.

To synthesise the information, and balance image quality with radiation doses, several figures of merit have been developed by combining image quality parameters such as the standard deviation in a region of interest (ROI) and the modulation transfer function (MTF). They were applied for specific clinical protocols to enable appropriate comparison of systems. This approach was quite useful during the development of CT technology, where performances between different units could vary drastically. These figures of merit can be based on simplified assumptions requiring caution in their interpretation. However it appears that the sensitivity of such methods is quite limited for newer systems, and, in addition, the effect of iterative reconstruction on the standard image quality parameters would mean that this approach would be difficult to implement.

Both clinical and phantom images can be assessed using the ROC paradigm or one of its derivatives (Localisation ROC, Free-response ROC). These methods give an accurate estimate of clinical image quality but, although carefully controlled measurements, they are still subjective because human observers are involved. These methods are time consuming and require large samples to obtain precise results. In spite of these limitations these methods can be used either by radiologists (when dealing with clinical images) or naïve observers when dealing with phantom images. To avoid the burden associated with ROC methods more simplified methods have been developed; for example, VGA (Visual Grading Analysis) in which image quality criteria can be used to give a relatively quick image quality assessment, without the explicit need for pathology or a task. Alternatively, phantom images can be assessed using the 2-AFC (two-alternative forced-choice) or M-AFC (multiple-alternative forced-choice) methods. Part 2 of this review discusses these methodologies, and these methods are used to validate the results produced by model observers presented in Part 3.

The introduction of iterative reconstruction in CT poses a new challenge in image quality assessment since most of the standard metrics presented in Part 1 cannot be used directly. In order to establish a bridge between radiologists and medical physicists, and therefore between clinical and physical image qualities, task related metrics can be used (even if the tasks are simplified versions of actual clinical tasks). Mathematical model observers are particularly suited to the routine image quality measurement of clinical protocols, with the results indicated to the user together with the standard dose report. Part 3 summarises the concepts behind these model observers, focusing on the anthropomorphic model observers that mimic human detection of simple targets in images, since the aim is to present tools for practical applications. The theory and description of the ideal observer can be found in the literature and a brief introduction to this model is done at the beginning of Part 3. Note that model observers can also be used when images are reconstructed with FBP. The inconvenience associated with the use of model observers is that they all lead to an overall outcome without the separation of the image quality parameters as with signal detection theory.

This paper is structured into three separate sections that provide an overview of the most common approaches taken when dealing with image quality in CT imaging. This structure is described in Fig. 2.

Traditional objective metrics

CT is a 3D imaging technique in which image quality assessment must be approached with some caution. Objective assessment of parameters that influence image quality is often made using physical metrics specified in either the spatial or spatial frequency domain. This duality is due to the fact that some features will produce overall responses which are independent of the location in the image, whereas other features will produce responses that are spatially correlated.

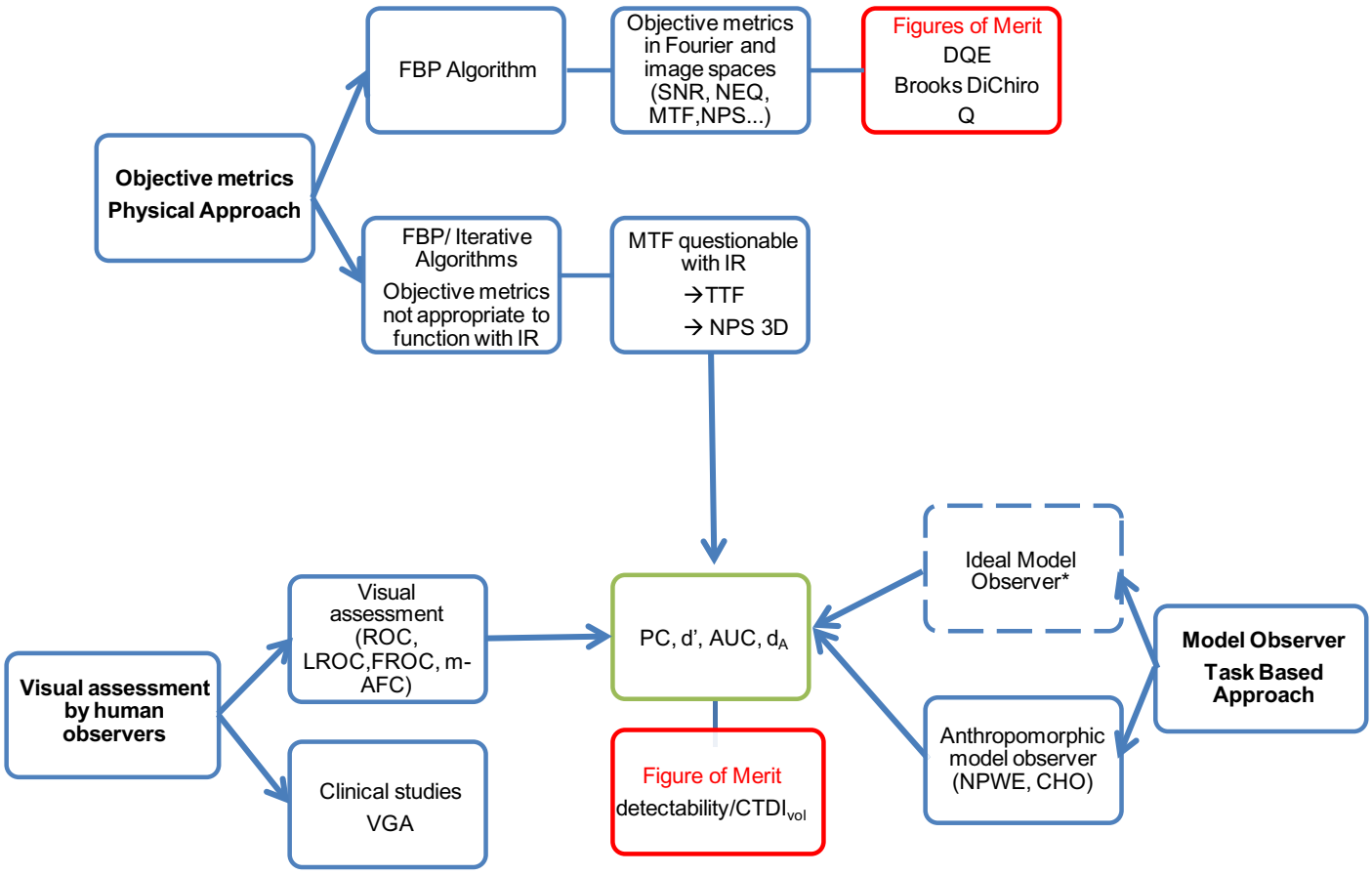

Figure 2. Summary of the content of the review (*this part will not be presented).

Objective metrics in spatial domain

Image signal and image noise are key parameters in image quality assessment. In the ideal and linear case, image signal (S) is directly linked to the detected number of photons N, whilst the noise (σ) may be seen as the pixel's stochastic fluctuation around their mean value. The photons are distributed according to Poisson's law, meaning that the quantity σ is equal to $\sqrt{N}$. The ratio of these two quantities yields the signal-to-noise ratio (SNR), expressed as:

$$SNR \propto \frac{S}{\sigma} = \frac{N}{\sqrt{N}} = \sqrt{N} \tag{1}$$

In an ideal device, each quantum could be counted by the detector and contributes towards the image. We could thus transpose Eq. (1) as:

$$SNR_{Ideal} \propto \frac{N_{Ideal}}{\sqrt{N_{Ideal}}} = \sqrt{N_{Ideal}} \tag{2}$$

However, due to the properties of the detector and its limited efficiency, a real measurement of the SNR would give the following result:

$$SNR_{Real} = \frac{N_{Real}}{\sqrt{N_{Real}}} = \sqrt{N_{Real}} < \sqrt{N_{Ideal}} \tag{3}$$

In Eq. (3), N_{Real} gives the number of quanta that contribute to the image for the real device and is also called noise-equivalent quanta (NEQ). Thus:

$$SNR^2_{Real} = N_{Real} = NEQ \tag{4}$$

Based on those parameters, we can eventually estimate the efficiency of a device by making the ratio between the number of photons actually used for the imaging and the incoming number of photons to the detector. This quantity is called detective quantum efficiency (DQE) and is defined as:

$$DQE = \frac{SNR^2_{Real}}{SNR^2_{Ideal}} = \frac{NEQ}{N_{Ideal}} \tag{5}$$

In Eq. (5), the NEQ can be measured in a straightforward manner, but some care must be taken when estimating quantity SNR^2_{Ideal}. Indeed, when considering a monochromatic beam, SNR^2_{Ideal} is simply the number of photons produced. However, for a polychromatic beam, SNR^2_{Ideal} should be the summed variance of the number of photons in each energy bin. In fact, some authors prefer to use an energy weighted variance because most detectors integrate energy [10] to form an image.

Another commonly used global image quality index is the signal difference-to-noise ratio ($SDNR$), defined for an object as the intensity difference from the background divided by the standard deviation:

$$SDNR = \frac{I_{object} - I_{Background}}{\sigma} \tag{6}$$

These metrics are extended to the spatial frequency domain in the following section.

Objective metrics in Fourier domain

Spatial resolution can be defined as the ability to distinguish two separate objects and is directly linked to the pixel size, the reconstruction kernel as well as the hardware properties of the imaging device. In order to derive an expression for image resolution, it is necessary to describe the imaging process generating a CT slice. Our analysis will be restricted to the axial plane. $I(x, y)$, which is the image slice of an input object denoted by $f(x, y)$, can be mathematically expressed as:

$$I(x,y)=\iint f(x-x',y-y')PSF(x',y')dx'dy' \tag{7}$$

with $PSF(x, y)$ being the point spread function in the axial plane and describing resolution properties of the device. It corresponds to the impulse response of a system, the response of the system to a Dirac input ($\delta(x, y)$).

Resolution can also be estimated through the line spread function (*LSF*), which is the response of the system to a straight line. Thus, the relationship between the *LSF* and the *PSF* can be derived from Eq. (7) in which the input function is replaced by the equation of a straight line in the axial plane (that is to say replacing $f(x, y)$ by $\delta(x)$ in Eq. (7)), yielding:

$$LSF(x)=\iint \delta(x-x')PSF(x',y')dx'dy'$$

leading to:

$$LSF(x)=\int_{-\infty}^{+\infty} PSF(x,y)dy \tag{8}$$

The point spread function needs to be similar at each location in the image (shift invariance) in order to ensure that the *LSF* will remain the same at every localisation. However, isotropy of the axial plane is a hypothesis which is not always true, especially when dealing with CT. In this case, the *LSF* will depend on the direction of the straight line in the axial plane. Assuming the straight line is positioned tilted with an angle θ the expression of the *LSF* will become:

$$LSF_\theta(x,y)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} PSF(x',y')\delta((x-x')\cos\theta+(y-y')\sin\theta)dx'dy' \tag{9}$$

Besides those two metrics, it is also possible to estimate the resolution using the edge spread function (*ESF*), that is to say the response of the device to an edge. An edge can be mathematically approached by the Heaviside function $H(x,y)=\begin{cases}1 & if\ x>0\\ 1/2 & if\ x=0\\ 0 & if\ x<0\end{cases}$. This function has the property: $\frac{dH(x)}{dx}=\delta(x)$.

Using this property, injecting $f(x, y)=H(x)$ in Eq. (7) and using Eq. (8) we obtain:

$$LSF(x)=\frac{\partial ESF(x)}{\partial x} \tag{10}$$

Hence, *PSF*, *LSF* and *ESF* are all related to each other and it is possible to use their representation in the frequency space thanks to the Fourier transform.

The Fourier representation of the *PSF* is the optical transfer function (*OTF*), which is defined as following:

$$OTF(u,v)\overset{\text{def}}{=}FT\{PSF(x,y)\} \tag{11}$$

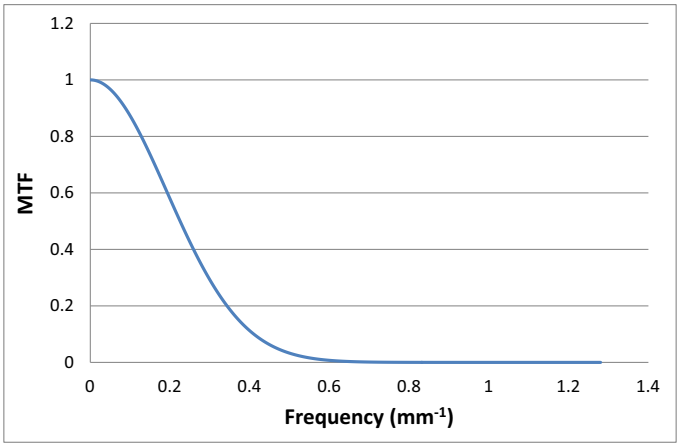

Figure 3. Example of a 1 dimension MTF curve of a GE VCT system with a 0.40 mm pixel size.

What is commonly used in order to estimate the resolution is the modulation transfer function (*MTF*), defined as the modulus of the *OTF normalised by its zero-frequency value*:

$$MTF(u,v)\overset{\text{def}}{=}\frac{|OTF(u,v)|}{|OTF(0,0)|} \tag{12}$$

Using Eqs. (8), (11) and (12) together with the Fourier slice theorem and assuming shift-invariance in the axial plane, we can state that a normalised radial *MTF* of the system is given by:

$$MTF_{1D}(f)=\left|\frac{FT\{LSF(x)\}}{\int_{-\infty}^{+\infty}LSF(x)dx}\right| \tag{13}$$

This metric describes how well frequencies are transferred through the system and is therefore used to make objective resolution estimation (Fig. 3).

Practically, the MTF can be computed from the image of a point (~PSF), a line (~LSF) or an edge (~ESF) [11–13]. In calculating MTF from the image of a point source (effectively from the PSF), a metal bead or taut wire fixed within a dedicated phantom is used to generate the signal [14]. Boone [12] used a tilted aluminium foil of thickness 50 μm to generate an oversampled *LSF*; the MTF is then computed using Eq. (13). Judy [13] was the first to describe calculation of MTF from an edge method in which the ESF was differentiated to give the LSF. This method has been developed over the years by various authors to include the use of spheres from which the oversampled ESF is built [15–17]. An older method was proposed by Droege and Morin, in which MTF is estimated from line pair test object images using the Coltman formula. Extensive details on the practical implementation of these techniques are given in ICRU Report 87 [18]. Several of these methods have been investigated by Miéville et al. in order to compare and contrast the advantages and drawbacks [19].

As with resolution, and of equal importance for SNR transfer, image noise can also be estimated in the frequency space. There are different sources of noise within the CT system, such as the electronic noise caused by the detector readout circuits (amplifiers) and the primary quantum noise which is inherent to the statistics of the limited quanta building the image. In a stationary system, the Wiener spectrum or noise power spectrum (NPS) gives a complete description of the noise by providing its amplitude over the entire frequency range of the image [20]. If the image noise is not stationary, the Wiener spectrum is not a complete description and the whole covariance matrix would be needed for complete description. However,

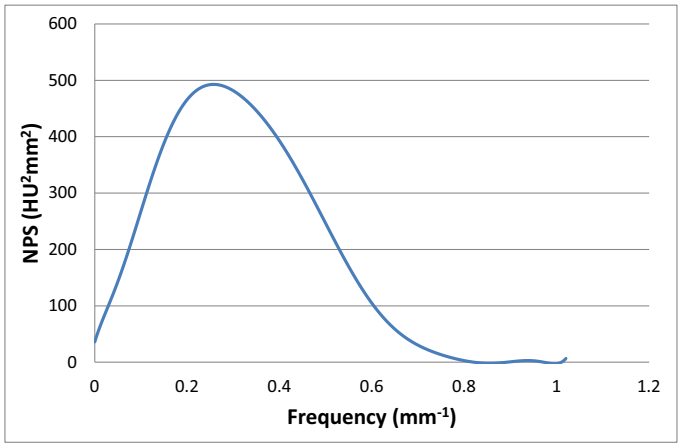

Figure 4. Example of a radially averaged NPS obtained on a GE revolution system with a standard convolution kernel.

if applied with care – for example working with small ROIs, extracted from a restricted region of the image – the NPS can be applied to both conventionally (i.e. FBP based) and iteratively reconstructed images. For NPS calculation, the assumption of 'small signal linearity' has to be made in order to apply Fourier analysis, which requires system linearity in order to be valid. This is the case for the logarithmic step applied to all reconstruction processes and also to the explicitly non-linear iterative methods.

In order to compute the *NPS* of an image, it is necessary to acquire homogeneous CT images and select region of interests (ROI) in this stack. The 2D *NPS* can then be computed as:

$$NPS_{2D}(f_x, f_y) = \frac{\Delta_x \Delta_y}{L_x L_y} \frac{1}{N_{ROI}} \sum_{i=1}^{N_{ROI}} \left| FT_{2D}\left\{ ROI_i(x,y) - \overline{ROI_i} \right\} \right|^2 \quad (14)$$

where Δ_x, Δ_y are the pixel sizes in the x and y dimension, $L_x L_y$ are the ROI's lengths (in pixel) for both dimensions, N_{ROI} is the number of ROIs used in the average operation and $\overline{ROI_i}$ is the mean pixel value of the ith ROI.

In practice, the *NPS* is largely affected by the detector dose, the hardware properties and the reconstruction kernel and algorithm. From each image of the stack a ROI is extracted and a custom computer program is generally used to compute the *NPS* according to Eq. (14). It is of common use to average the 2D *NPS* along a 1D radial frequency using the equation $f_r = \sqrt{f_x^2 + f_y^2}$ (Fig. 4). More details on the *NPS* computing can be found in ICRU Report 87 [18]. In the end, the NPS characterises the noise texture, thus giving a better and more complete description of noise than the simple pixel's standard deviation. Moreover, information about the pixel's standard deviation can still be retrieved with knowledge of the Wiener spectrum. Indeed, the Parseval theorem ensures that the total energy is obtained by summing the contribution of the different harmonics and that its value does not depend on the chosen space (image or frequency space). Since the *NPS* is a spectral decomposition of noise over frequencies, we have:

$$\sigma^2 = \iint NPS_{2D}(f_x, f_y) df_x df_y \quad (15)$$

As explained before, *MTF* shows how well the signal frequencies are transferred through an imaging system, that is to say it exhibits the signal response of a system at a given spatial frequency. As for the spatial domain, the ratio of signal (i.e. MTF) and noise (i.e. NPS) yields the output signal to noise ratio (the NEQ) and therefore the frequency dependent *NEQ* can be calculated as:

$$NEQ(f) = SNR^2_{Real}(f) = \frac{a^2 MTF^2_{1D}(f)}{NPS_{1D}(f)} \quad (16)$$

where a^2 is the mean pixel value squared.

The DQE in the frequency space can therefore be estimated by:

$$DQE(f) = \frac{SNR^2_{Real}(f)}{SNR^2_{Ideal}(f)} = \frac{a^2 MTF^2_{1D}(f)}{N_{Ideal} NPS_{1D}(f)} \quad (17)$$

Limitations of conventional and Fourier-based image quality metrics for the assessment of IR images

In order to compute an MTF that represents the spatial resolution of the entire image, the assumption of shift-invariance has to be made. That is to say that the device's response has to remain the same, whether measured at the image centre or periphery. If this assumption is not fulfilled it is necessary to make the measurements at the same location in different images to obtain an *MTF* that can be used to compare the resolution of different devices. Furthermore, the linearity hypothesis also needs to be fulfilled for the *MTF* to be reliable. That is to say, the output signal has to remain within the optimal range of response of the imaging system in terms of Hounsfield units (HU), usually in the range from –200 to +200 HU for clinical CT scanners [18]. Consequently, estimating the *MTF* with a high Z material can give a signal outside this range, yielding an incorrect assessment of resolution. In practice, estimating the *MTF* with high Z materials generally leads to a resolution overestimation because of the high SNR they generate [18].

Those two assumptions are approximately satisfied for CT images reconstructed with filtered back projection (FBP) algorithms and a standard reconstruction kernel, but the introduction of iterative reconstruction (IR) has changed the game [21]. Indeed, IR images exhibit stronger non-linear and non-stationary properties that force a change in the MTF measurement paradigm. Several authors have already highlighted the non-linearity problem of these algorithms, which manifests itself as contrast dependency of the resolution [21–23]. Also, investigations on how Fourier-based metrics are influenced by the characteristics of IR images have been described [24,25]. They showed, for example, that the shape of the NPS for some IR algorithms also depends on the dose level and that the resolution not only depends on the contrast but also on the radiation dose levels. These elements have highlighted the need to adapt the existing metrics to IR algorithms.

Adaption of Fourier metrics

These difficulties in estimating resolution can be overcome by using an adapted metric, such as the target transfer function (*TTF*), which makes it possible to characterise the resolution even in the presence of noise and contrast dependency [24,26]. *MTF* and *TTF* are similar but differ from one another in the sense that *MTF* only applies to a single given contrast level, whereas a *TTF* will exhibit three different curves at three different contrasts (corresponding to three different materials) for one measurement (Fig. 5). This enables a characterisation of the resolution when dealing with non linear algorithms for which contrast influences the resolution. As already demonstrated by several authors this will make full characterisation of the resolution possible when dealing with IR [24,27].

The technological evolution of CTs has also led to changes in the way *NPS* must be computed. The 2D axial *NPS* was well suited for the first generations of devices where only one CT image per axial scan could be acquired without noise correlation between slices. Now that the acquisitions are also made in helical mode and that the number of detectors along the z-axis is higher, a 3D NPS is required to fully characterise the noise (Fig. 6) [12,28]. 3D NPS can

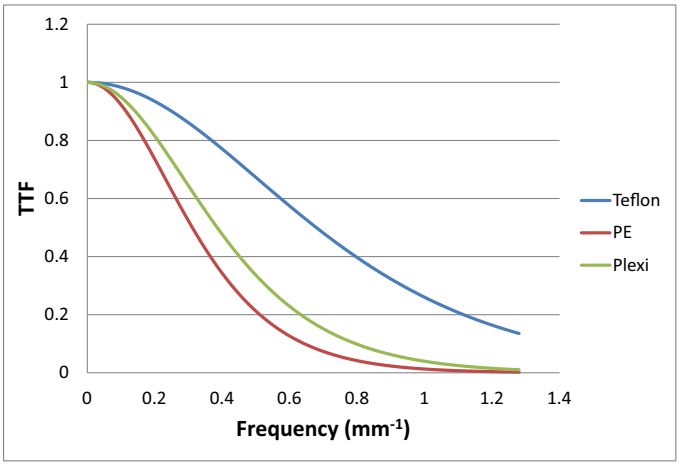

Figure 5. Resolution estimation through the *TTF* on a GE HD 750 system with a 0.4 mm pixel size and three different materials (Teflon polyethylene and plexiglass). Differences are observed on the resolution depending on the material. Such changes could not be observed when using the *MTF*.

be measured in a similar manner to the 2D NPS, but working with volumes of interests (VOI) instead of ROIs:

$$NPS_{3D}(f_x, f_y, f_z) = \frac{\Delta_x \Delta_y \Delta_z}{L_x L_y L_z} \frac{1}{N_{VOI}} \sum_{i=1}^{N_{VOI}} \left| FT_{3D}\left\{ VOI_i(x, y, z) - \overline{VOI_i} \right\} \right|^2 \quad (18)$$

For this case, the units of NPS are HU^2mm^3.

In this particular paradigm, Eq. (15) becomes:

$$\sigma^2 = \iiint NPS_{3D}(f_x, f_y, f_z) df_x df_y df_z \quad (19)$$

How to synthesise the information towards a figure of merit

Combining image quality and dose

In the clinical setting the focus for optimisation is balancing image quality and radiation dose in the context of the clinical question. Statistical noise, spatial resolution and imaged slice width are the fundamental parameters which describe the amount of object information retrievable from an image, and give rise to the perceived image quality. X-ray dose can be regarded as the cost of this information. It is meaningless to quote any of these image quality quantities without reference to the others, or to the radiation cost. The 'holy grail' is to try to find a way to combine the relevant parameters objectively and appropriately in a dose efficiency factor.

A dose efficiency factor, or figure of merit, can take a number of forms depending on how the various parameters are measured and quoted. Correctly developed and applied it can be used as a tool to compare scanner models, or simply different scan settings to optimise the balance of image quality and radiation dose.

How these parameters, resolution and noise in particular, are balanced is dependent on the clinical question and examination type. An important aspect that must be addressed is the influence of scan and protocol parameters that can be adjusted by the operator and how they affect image quality and radiation dose performance.

Clinical scanner settings – scan and protocol parameters

Any consideration of a theoretical approach to investigate a dose efficiency value needs to be in the scenario of the clinical question and the parameters used to create the image (Fig. 1).

Image quality and dose can be affected by the scanner design and also by the scan settings in the selected protocol (Table 1). The effect of the scan parameters, which form the examination protocol for the clinical question, can be seen in Table 2.

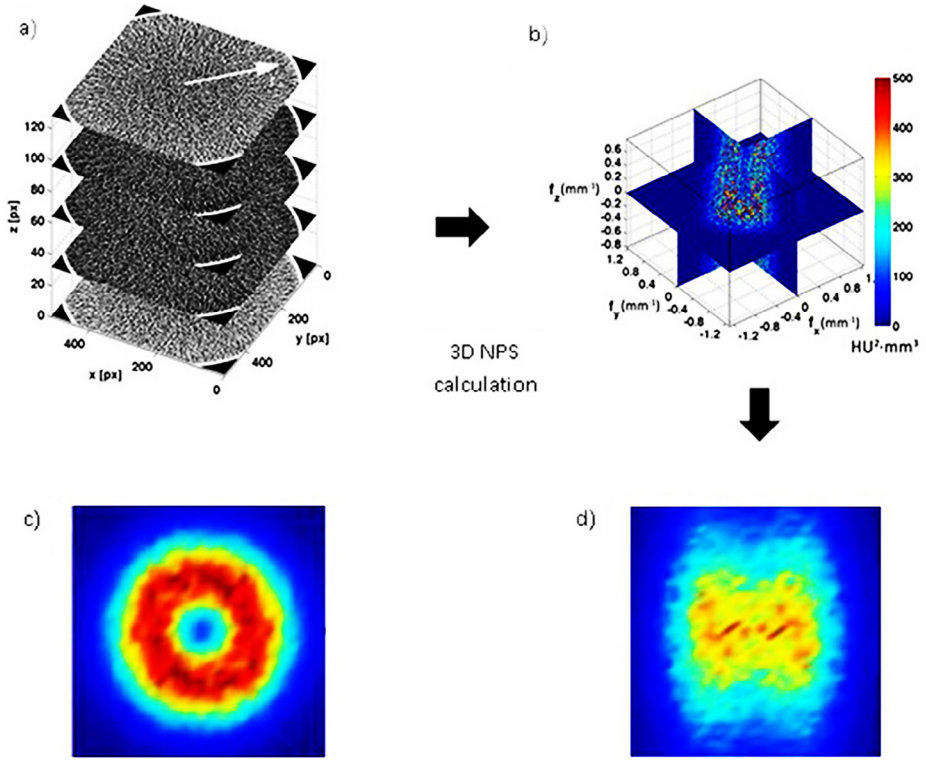

Figure 6. (a) The 3D homogeneous volume from which the 3D NPS is extracted. (b) The 3D NPS and the NPS sectioned in the (c) x–y (axial) and (d) the x–z (sagittal) planes. Figures extracted from Reference 22.

Table 1
Scanner design and scanner settings which can affect image quality and dose on scanner settings (courtesy of ImPACT [29]).

Scanner design factors	Scan protocol factors
Detectors material	Clinical application
Detector configuration	Tube current, tube voltage, focal spot size
Numbers of detectors, rows	Image reconstruction algorithms
Data acquisition rates	X ray Collimation width, detector acquisition width
Software corrections	
Filtration	Reconstructed image slice thickness
Focal spot size	Helical pitch
Geometry (i.e. focus–axis, focus–detector distances)	Interpolation algorithms

Combining image quality and dose metrics – theoretical background

The basic starting premise for a figure of merit for a dose efficiency parameter is that a dose efficient scanner will produce good resolution at minimum dose and noise.

There are a number of mathematical relationships that can be found in the scientific literature, both in terms of general imaging theory and for CT in particular [30–32]. The two of interest for CT are Brooks and Di Chiro [33] and Riederer et al. [34]. These were used in the development of the ImPACT Q value which became a useful, and relatively widely known, approach for comparing CT systems in the 1990s [35,36]. It was also explored by Fuchs and Kalender [37], more recently Kalender devoted a section to this subject in his book Computed Tomography: Fundamentals, System Technology, Image Quality, Applications [37,38]. However the fundamental relationship can also be found in standard textbooks on imaging with radiation [39,40]. The core of all these approaches is that the noise squared is inversely proportional to dose, and also inversely (in real or image space) proportional to the spatial resolution to the power 4. This encompasses spatial resolution in the x,y (to power 3) plane and also the z plane and quoted either as a size or frequency. In some equations the resolution is separated out into frequency for the x and y plane resolution, and the image thickness for the z-axis (z,x and z,y planes).

The relationship can be explored in more detail using the Brooks and Di Chiro equation [33]:

$$\sigma^2(\mu) = \frac{\pi^2 \beta \gamma(E) e^{\alpha} \mu_{en} E}{1200 \omega^3 h D} \tag{20}$$

Here σ^2 is the statistical error in the reconstructed image (i.e. the image noise); β is a beam spreading factor (non-parallel rays), $\gamma(E)$ is the average depth dose factor for photon energy (E), e^{α} is the logarithmic attenuation, μ_{en} is the energy absorption coefficient, E is the photon energy, ω is the detector aperture, h is the slice width, and D is the radiation dose.

For the purposes in this chapter, this can be simplified to:

$$\sigma^2 \propto \frac{1}{\omega^3 h D} \tag{21}$$

Similarly the Reiderer, Pelc and Chesler relationship is given as [33]:

$$\sigma^2 = \frac{\pi}{m N_p} \int_0^{2\pi} d\phi \int_0^{\infty} k dk \frac{|G(k)|^2}{k} \tag{22}$$

where m is the number of projections, N_p is the number of photons per projection, and $G(k)$ is the convolution function with frequency. The product $m\mathrm{N_p}$ could be regarded as a measure of radiation dose.

This essentially becomes $\sigma^2 \propto {k_c}^3/\mathrm{mN_p}$ (where k_c is the cut-off frequency, i.e. the limiting resolution). Or, indeed as the paper states; 'for all valid correction filters … σ^2 varies with the cube of the resolution'.

This is, in effect, the relationship of:

$$\sigma^2 \propto \frac{1}{N} \quad \text{i.e.} \quad \sigma^2 \propto \frac{1}{D} \tag{23}$$

where N is the number of photons and D is a measure of radiation dose for a fixed value of tube voltage. This can also be seen as a direct result from Eq. (1), assuming Poisson noise and without additive electronics noise.

Combining image quality and dose metrics – a practical approach

The discussion that follows is the approach taken by the UK CT scanner evaluation facility ImPACT [36]. It is a pragmatic solution to a complicated scenario of practical and computational effects on resultant image quality and dose for the operational CT scanner. This approach was reasonably successful for a number of years, and many scanner comparison reports were produced using this factor [32]. There is no other known work in this area covering a number of decades of scanner development. All measurements were undertaken according to a procedure with strict criteria, and in consultation with manufacturers as to the nature of their scan protocols, scanner features and reconstruction parameters. Measurements and analysis were carried out using typical clinical protocols, using the same image quality and dose assessment and calculation methods, and the same team of people. As scanners developed it became harder to apply such strict criteria, and with the development of adaptive filtration, and iterative reconstruction methods, it became very difficult to minimise the effects of other variables on ascertaining a dose efficiency value for a typical scan protocol.

$$Q \alpha \sqrt{\frac{f^3}{\sigma^2 z D}} \tag{24}$$

Table 2
Dependence of image quality and dose parameters on scanner settings (courtesy of ImPACT, adapted from Reference 29).

	Noise	Slice width	Scan plane resolution	Dose
kV				
Effective mA (mA/pitch)				
Focal spot selection				
Pitch				
X-ray beam collimation				
Detector configuration (e.g. 16 × 1.25 versus 32 × 0.62)				
Scan time (for a given mAs)				
Interpolation algorithm				
Convolution kernel				
Reconstructed slice thickness				
Use of iterative reconstruction				

The dark blue represents a major dependence of image quality and dose on scanner settings and the soft blue represents a minor dependence.

where σ is the image noise, f is a measure of the in-plane spatial resolution (in frequency space), z is a measure of the spatial resolution along the z-axis (in image space, and a measure of the z-sensitivity), and D, as indicated above, a is measure of the radiation dose. This is the approach used by the ImPACT CT scanner evaluation facility [32,36] and first proposed in 1978 by Atkinson [35]. Initially one form of the generic equation was used, and then altered some of the definitions of the parameters involved, to create what became known as Q2 [31,41] as shown in Eq. (25).

The Q-factor (Q_2 factor) is in part empirical, it was used with caution and with strict adherence to the calculation procedure, which included standardising certain scan and protocol variables. Since it is not an absolute figure, it cannot be applied to the overall scanner, only to the examination protocol. Each set of image quality and dose parameters was therefore focussed on a typical clinical type of examination; for example a standard brain or standard abdomen.

The first step in the process was to ascertain this scan protocol in conjunction with the manufacturer. Consideration of the effects of the scanner settings, as shown in Table 1, required some adjustment of the protocol. This was in order to minimise the effects of scan parameters whose effects confounded the aim of comparison of image quality and dose, in the context of dose efficiency of the system. The associated challenge was to maintain the integrity of the suggested protocol for that type of examination. The second step was to undertake the various image quality and dose measurements and calculations, and then finally to apply the Q2 relationship.

$$Q_2 = \sqrt{\frac{f_{av}^3}{\sigma^2 z_1 CTDI_{vol}}} \tag{25}$$

The specific parameters used in calculating this value were measured using standard techniques and quoted parameters, such as would be used for quality control or acceptance testing:

σ = the image noise, the standard deviation from the CT numbers of a specified sized region of interest (5 cm^2), expressed as a percentage (for water, standard deviation in HU divided by 10), measured at the centre of the field of view in a standard water phantom.
f_{av} = spatial resolution, given as $(MTF_{50} + MTF_{10})/2$, where MTF_{50} and MTF_{10} are the spatial frequencies corresponding to the 50% and 10% modulation transfer function values respectively (in line pairs per cm).
z_1 = the full width at half maximum (FWHM), (mm), of the imaged slice profile (z-sensitivity). This is measured using the inclined high contrast plates method (mm).
$CTDI_{vol}$ = volume weighted CT dose index (mGy).

To understand the dose efficiency relationship further in a practical manner, it can be helpful to consider the basic equation (Eq. 24) to be formed of three components:

$$\sigma^2 \propto \frac{1}{D}, \quad \sigma^2 \propto \frac{1}{z} \quad \text{and} \quad \sigma^2 \propto f^3 \tag{26}$$

which, in the Q_2 relationship, translate to:

$$\sigma^2 \propto \frac{1}{CTDI_{vol}}, \quad \sigma^2 \propto \frac{1}{z_1} \quad \text{and} \quad \sigma^2 \propto {f_{av}}^3 \tag{27}$$

Each of these relationships will be addressed more fully in the following sub-sections.

Dose value. The dose value in an earlier formulation of Q was the surface dose to a phantom, measured using thermoluminescent dosimeters. This was changed for Q_2 with the introduction of the standardised $CTDI_{vol}$ parameter. The cross-sectional averaging that contributes to the creation of the $CTDI_{vol}$ is more representative of the overall dose to the phantom and therefore a more appropriate value to be used.

The inverse relationship of dose with σ^2, $\left(\sigma^2 \propto \frac{1}{CTDI_{vol}}\right)$ has to be carefully considered with multi-slice CT beams. In CT it is generally acknowledged that the $CTDI_{vol}$ is a suitable dosimetry parameter; however the proportionality breaks down in MSCT since the penumbra contribution to the beam width is a constant value, and as such is a factor that affects the relative dose, and is not accounted for in the relationship. Therefore to accommodate this, the beam width needs to be kept as a constant when comparing one scanner to another, or to take it into account separately with a beam width correction factor.

Image slice width (z_1) – z-axis resolution. The effect on noise from the thickness of the slice (z_1) is from the imaged, as opposed to the nominal, slice width, with a dependence on the inverse proportionality of photons contributing to the image. For testing purposes the full width at half maximum (FWHM) of the imaged slice profile is a suitable parameter to use. However this does not fully describe the imaged slice profile, in terms of the photon distribution contributing to the reconstructed image. For ease of application the FWHM is used, even though a fuller description of this sensitivity profile would be better.

Spatial resolution (f_{av}). A similar approach is taken with the spatial resolution parameter. Rather than using a single value from the modulation transfer curve, a more complete description of the resolution takes into account the full function over all frequencies, and a resolution value based on an average of the 50% and the 10% values of the modulation transfer function is therefore used. These values, averaged, do not completely describe the spatial resolution function, however they are common values automatically extracted from MTF curves as part of a standard testing process, and together were deemed to provide a better indication of the compromise between high and low spatial resolutions, compared to only one of the parameters alone.

The derivation of the cubed relationship of noise with spatial resolution ($\sigma^2 \propto {f_{av}}^3$) relies on assumptions of the shape of the convolution filter used (for example in Brooks and Di Chiro [33], the convolution filter is a ramp filter). In this way comparisons between scanners are likely to be more reliable when comparing images reconstructed with similar convolution filters and, in particular, algorithms that best represent ramp filters. These are in general, the filtered back projection filters named for 'standard' applications, providing reasonably low spatial resolution in order to preserve the contrast detectability in an image. Filters that are slightly smoothed or slightly enhanced would be considered as close; however those with strong smoothing or strong edge enhancing would not be suitable. Reconstruction filters with 'standard' spatial resolution values were therefore chosen to minimise the dependency of Q_2 upon non-ramp like reconstruction filters. Fortunately, or appropriately, these were also the algorithms usually used in the standard clinical protocols under investigation. This aspect of the Q_2 equation is a pragmatic solution for the complexity of modern reconstruction algorithms. The reconstruction filter with MTF_{50} and MTF_{10} values as close as possible to 3.4 lp/cm and 6.0 lp/cm was used.

When investigating the empirical relationship with actual reconstruction filters, which range from ramp-like standard filters with conventional apodisation functions, to edge-enhancing high spatial resolution filters, it was found that the relationship was closer to a power of 4 or 5 [29,42].

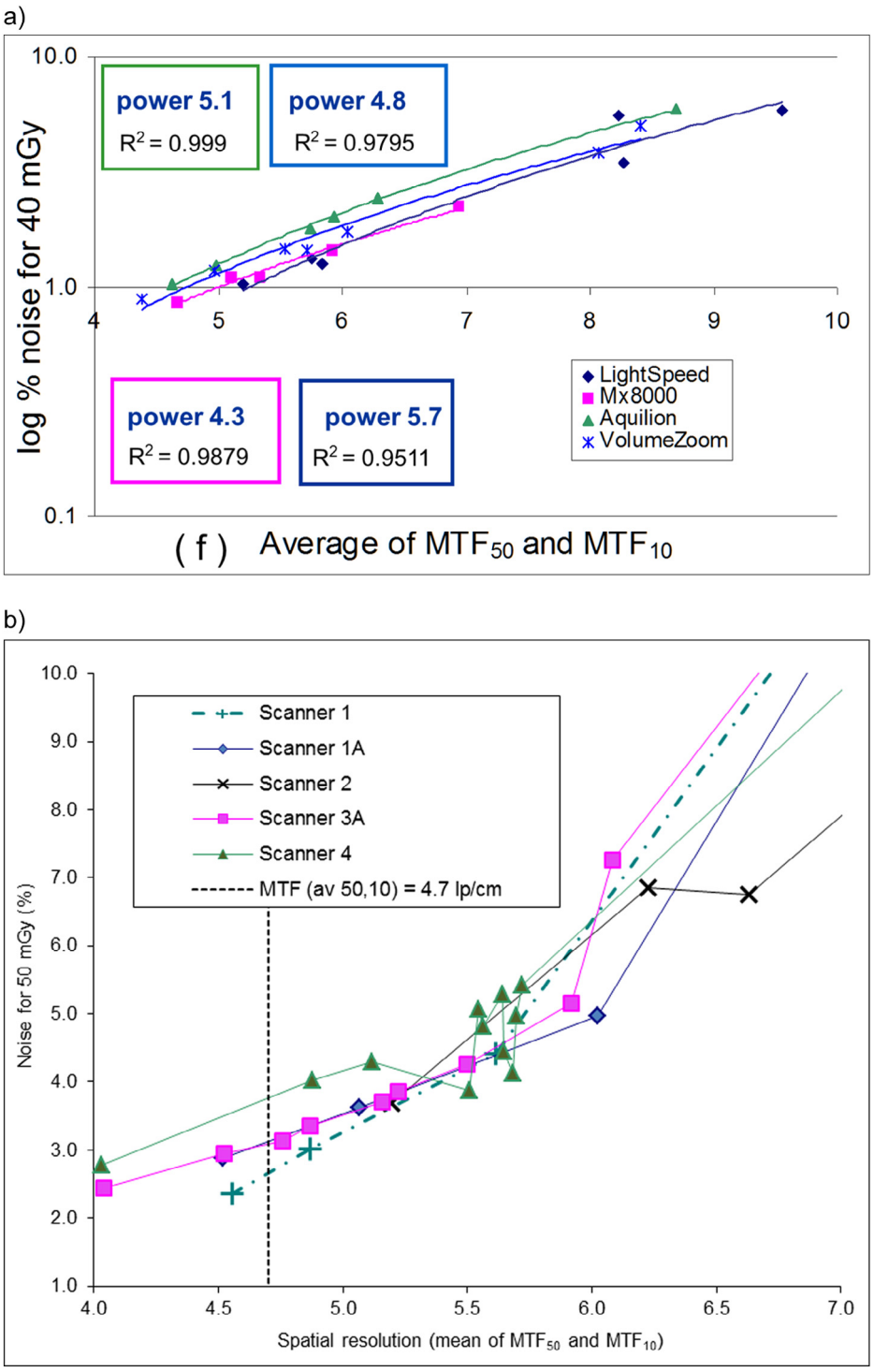

Figure 7. (a) Example for body algorithms, of logarithmic image noise against spatial resolution, with normalised dose (CTDI), demonstrating the deviation from the expected relationship. (The 'power' is the power to which f_{av} is raised against σ^2) (courtesy of ImPACT). (b) Head algorithms showing associated image noise against spatial resolution, with normalised dose (CTDI), demonstrating, particularly for scanner4, how small changes in spatial resolution give rise to large changes in measured noise [from data in Reference 41].

$$\sigma^2 \alpha \frac{f^{=4-5}}{zD} \tag{28}$$

This is illustrated in the following graph (Fig. 7a), for the body scans. The different points on the graph relate to different reconstruction algorithms. This reinforces the need to compare the 'Q' for scanners with image quality parameters measured using standard algorithms only, as the cubed power relationship is not valid across the whole range of spatial resolutions.

However, with modern scanners and reconstruction algorithms, even with a 'standard' algorithm there can be anomalies in the expected relationships. With adaptive filtration and special reconstruction techniques, even selecting the lower spatial resolution algorithms, inconsistencies in the 'straight line' relationship can

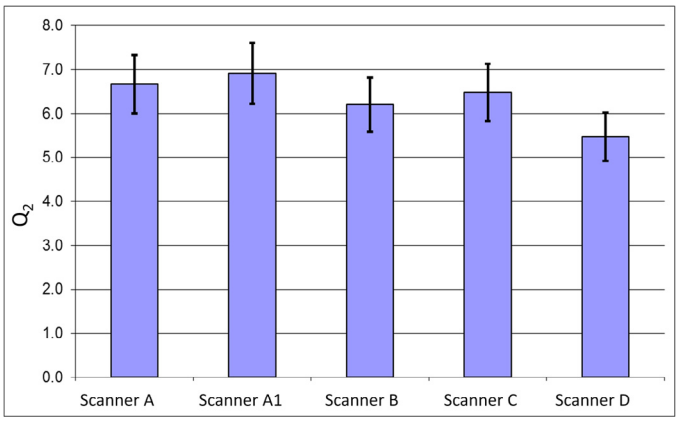

Figure 8. Q2 values for several 16-slice scanners for standard head scans (courtesy of ImPACT).

appear, where a small increase in spatial resolution may not bring the expected associated increase in image noise, as shown in Fig. 7b [41,43].

The uncertainty in the Q value was estimated to be about 15%, and therefore, even once the confounding variables are standardised, it cannot be used to look for fine differences in the image quality and dose relationship [36,41,43], as shown for a set of 16-slice scanners in Fig. 8 [41].

However, it can demonstrate larger differences – such as with the difference between the dose efficiency of xenon gas and solid state detectors. Figure 9 shows data from the original 'Q' value, where surface phantom dose measurements were used (giving the surface multiple scan average dose (MSAD)). By normalising for the spatial resolution both in the z-axis (the image slice thickness) and the scan plane, this can be shown graphically as a relative dose.

Alternative method for combining parameters. Another approach to define CT dose efficiency was suggested by Nagel [44]. This approach for image quality determination is based on a statistical method of determining low contrast detectability (LCD) as previously suggested by Chao et al. [45]. In this method, a uniform phantom is scanned with specified dose and parameter settings. An array of square regions of interest (ROIs) is defined on the uniform image that is covering approximately a third of the central image area. By measuring the distribution of mean CT numbers of the ROIs and assuming a normal distribution, a prediction can be made of

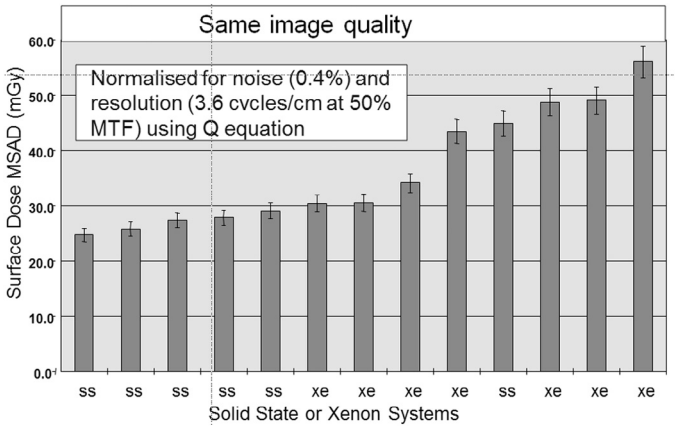

Figure 9. The use of a pervious version of Q to illustrate the relative dose, normalised by the other factors (courtesy of ImPACT) [slide 36 from Reference 29].

the CT number threshold of a low contrast detail having the same size as the ROIs in order to detect it at a 95% confidence level. This threshold contrast C is 3.29 times the standard deviation σ. This parameter is obtained by measuring the mean CT numbers of the ROIs before calculating their standard deviations. There is a 95% probability that a low contrast object of the same size as the ROIs is missed if the contrast is within the normal variation in the ROI means, i.e. if $C < 3.29\,\sigma$. Similarly, with a probability of 5%, a randomly high fluctuation of some ROI numbers could be mistaken for an actual low contrast object if the contrast of interest is sufficiently small. According to the Nyquist theorem, the ROI size limits the noise power spectrum (NPS) at a relatively low spatial frequency (here, approximately 1 lp/cm). Therefore, a measure of the detectability of low contrast objects having the same size as the ROIs suppresses spectral noise components at high spatial frequencies that are strongly affected by the detector and reconstruction algorithm.

The CT dose efficiency value (CTDEV) puts all parameters that are relevant for the specification of LCD into a single number that is based on the fundamental theory of Rose [46]:

$$CTDEV = 10^5 \frac{e^{0.207(D_{eq}-16)}}{d^2 C^2 h_{rec} CTDI_{Vol,H}} \tag{29}$$

with the diameter d of the low contrast detail (here, $d = 5$ mm), the slice thickness h_{rec} (in mm), the volume CT dose index $CTDI_{vol,H}$ for the 16 cm head phantom, the PMMA-equivalent phantom diameter (in cm) D_{eq}, and the detail contrast (in %, with 1% = 10 HU) $C = 3.29\,\sigma$.

The method of Chao et al. can be easily implemented by applying customary CT phantoms and reduces the variability in LCD visually specified by human observers in conventional image quality assessments [45]. Chao's method has been applied by two CT manufacturers for the assessment of low contrast specifications [47]. The result of the method, however, depends on the size of the predefined ROI, the location of the CT image slice within the cone beam, and the filter used for image reconstruction [48]. As with other figures of merit, such as the Q_2 value, to apply the CTDEV for CT benchmarking, certain features must be standardised in detail. These are the protocol parameter set, reconstruction filter, phantom and method used.

Measures of diagnostic performance

Visual grading analysis (VGA)

Complementing the physical measurements of image quality, the assessment by observers is a subjective way to evaluate the image quality. Several general principles apply to all subjective observer studies: patients should be selected to have a wide range of body habitus, they should involve as many observers as possible, and they should cover the range of expected competencies in the field [49]. When these assumptions are verified, the visual grading analysis (VGA) based on observer scorings can be used to assess image quality.

VGA provides two types of information [50]:

Firstly, this subjective analysis provides information on the acceptability of the appearance (i.e. image noise level) of the clinical images and how the anatomical structures are visualised. For example the VGA grades the visibility of important structures for different noise levels, because the detectability of low contrast structures is affected by noise, decreasing as the noise level increases.

Secondly, the subjective evaluation provides a context to interpret the physical metrics (i.e. *MTF*, *NPS*). Human observer evaluation is subject to change depending on context

(i.e. brightness, tiredness), so the variability is not negligible and it is important to have a sufficient number of observers. For instance if a CT has 40% better *MTF* at high frequencies than another, but both CTs are rated by a single observer the difference between both systems will not become significant.

The VGA paradigm is split into two categories: relative grading and absolute grading.

Relative grading: The observer grades the image quality compared to a reference image or to the other images. The images should be displayed in random order to avoid any bias (i.e. first image read bias) and the viewing conditions should reproduce the darkened environment of the reading diagnosis room [51]. The parameter studied should be as specific as possible, but it is possible to ask more than one question in order to evaluate several specifications. The rating scale used in relative grading can have 3, 5, or more steps/ranks. The scale with 3 steps is not ideal because it is impossible to differentiate sufficiently. But when the degree of difference is small, a two step scale can be a possibility. The quality of the test is dependent on the reference image.

For instance, a scale with 5 steps can be represented by:

–2: A is much better than B
–1: A is slightly better than B
0: A and B are equal
+1: B is slightly better than A
+2: B is much better than A

Absolute grading: The observers do not have any references and the images are displayed one by one. The evaluation is performed for one image at a time unlike the relative grading. To avoid bias from observer learning, the reading sessions must be separated in time. The grading scale should be numerical (i.e. from 1 to 10) or adjectival. With the adjectival scale, the descriptor should be expressive in order to create a difference between the worst and best cases. For instance, the Likert scale is a non-comparative ordinal scale used especially in psychometric studies where the participants express their level of agreement with a given statement. Note that reproducibility is low with this type of study [52–54].

The results of a VGA study can be summarised with the VGA Score (VGAS):

$$VGAS = \frac{\sum_{o,i} S_c}{N_i N_o} \tag{30}$$

where Sc = the given individual scores for observer (o) and image (i), Ni = total number of images, and No = total number of observers. In a VGA study to analyse the statistical difference, the analysis of variance (ANOVA) is calculated, associated with procedures for multiple comparisons.

For VGA, clinical images are required, which increases the implementation difficulties and also forces the avoidance of naïve observers. Indeed, to assess image quality in the VGA paradigm, the observer experience is very important if we want the obtained results to be as little distorted as possible. Nonetheless, VGA results are subjective and the analysis may be influenced by the experience of the radiologist, for instance in visualising different noise textures.

Decision theory: the statistical approach

It is common practice to specify the performance of diagnostic systems in physical terms as described in Part 1. However, it is complicated to translate these results to clinical performance. For instance, in detection tasks, certainty is rarely present. When an observer is asked to detect a signal on a medical image **g**, the result is a degree of belief that the signal is present. This degree of belief

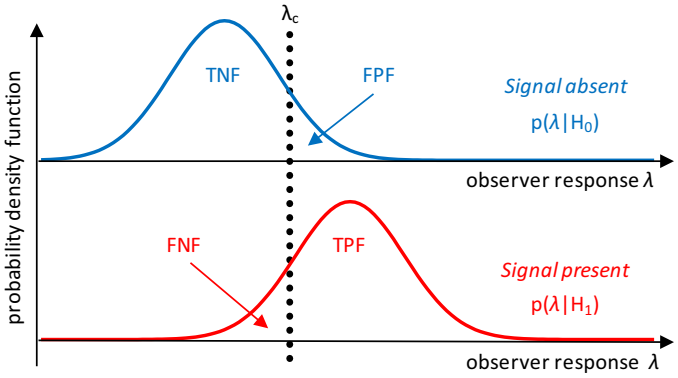

Figure 10. Probability density function of the observer response λ when presented with signal-absent images (top) or signal-present images (bottom). The vertical line λ_c indicates the threshold response above which the observer gives a positive response. TNF: true negative fraction; FPF: false positive fraction; FNF: false negative fraction; TPF: true positive fraction.

is commonly called the response $\lambda(\boldsymbol{g})$ of the observer: a low value denotes a confidence that the signal is absent, whereas a high value corresponds to the conviction that the signal is present. As shown in Fig. 10, the probability of obtaining a response can be plotted over all possible responses for two categories of images: those that do not contain a signal (top) and those that do contain a signal (bottom). These two curves are called probability density functions (pdf): respectively $P(\lambda|H_0)$ and $P(\lambda|H_1)$, where H_0 is the null-hypothesis corresponding to signal absent and H_1 is the alternative hypothesis corresponding to signal present. In radiology, the observer is forced to make a decision. In the present framework, this means that the observer chooses a threshold λ_c above which a positive decision is made. Below λ_c the observer makes a negative decision.

The integral of the distribution $P(\lambda|H_0)$ that is below the threshold is called the true negative fraction (TNF), or specificity. On the other hand, the integral of the distribution $P(\lambda|H_1)$ that is above the threshold is called the true positive fraction (TPF), or sensitivity. If the detection strategy is good, one expects both specificity and sensitivity to be as high as possible. However, Fig. 10 shows that changing the threshold changes the balance between specificity and sensitivity: increasing one parameter leads to a decrease of the other.

There are mainly two ways to quantify the effectiveness of the strategy. The first is the signal to noise ratio defined as follows:

$$SNR_\lambda = \frac{|\mu_1 - \mu_0|}{\sqrt{\frac{1}{2}\left(\sigma_0^2 + \sigma_1^2\right)}} \tag{31}$$

where μ_0 and μ_1 are the means of $P(\lambda|H_0)$ and $P(\lambda|H_1)$, respectively, and σ_0 and σ_1 are the corresponding standard deviations. SNR_λ is a global figure of merit that broadly describes how two distributions are separated. This equation is similar to Eq. (6) about SDNR and its purpose is to compare two situations (with and without noise). However, Eq. (31) characterises the response of an observer and not a signal or a noise directly measurable on an image.

$SNR_\lambda = 0$ corresponds to the situation where the two pdfs have the same mean. If their shapes are the same, the decision based on such a strategy will be just guessing, and therefore the image does not transfer any information about the presence of the signal. A large SNR_t corresponds to well-separated pdfs. If the threshold is chosen between the distributions, then a large number of correct responses are expected.

A second way to quantify the effectiveness of the strategy is the receiver operating characteristics (ROC) curve, which displays all the possible combinations of sensitivity and specificity obtainable whilst

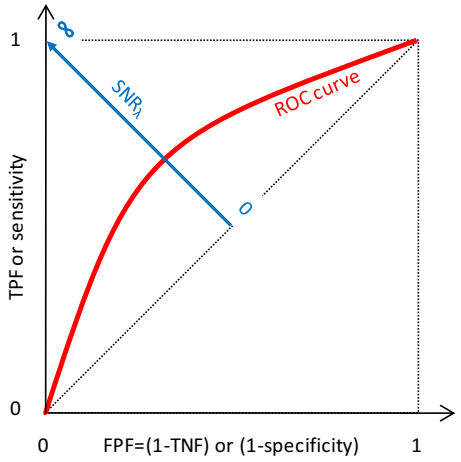

Figure 11. The ROC curve displays the true positive fraction versus the false positive fraction. If both response distributions are Gaussian with the same variances, then the intercept between the ROC curve and the second diagonal corresponds to SNR_λ.

we vary the threshold from the lowest to the highest possible values [55].

For historical reasons, the ROC curve displays the TPF versus the FPF, which is the sensitivity versus the (1 – specificity). If the pdfs are superimposed, the ROC curve is the straight line TPF = FPF. If the pdfs are well separated then the ROC curve has a square shape that passes close to the perfect point defined by sensitivity = 1 and specificity = 1. If pdfs are Gaussian with equal variances (this is often assumed in practice), the ROC curve is symmetrical and its intercept with the secondary diagonal corresponds to SNR_λ (Fig. 11). The value computed from the intercept between the ROC curve and SNR_λ is called the detectability index and usually represented with the symbol *d′*.

In practice, the observer (e.g. the radiologist) chooses a given threshold that corresponds to an operating point on the ROC curve. An objective way to define an optimal combination of sensitivity and specificity consists of computing the mean cost associated with all possible combinations of decision (negative or positive) and reality (signal absent or present):

$$\begin{aligned}\bar{C} = {} & C_{00}P(D_0|H_0)P(H_0) + C_{01}P(D_0|H_1)P(H_1) \\ & + C_{10}P(D_1|H_0)P(H_0) + C_{11}P(D_1|H_1)P(H_1)\end{aligned} \tag{32}$$

where C_{ij} is the cost associated with decision D_i and reality H_j, $P(D_i|H_j)$ is the pdf to make a decision D_i when the reality is H_j, and $P(H_1)$ is the probability to have a signal present. The latter is called prevalence in the case of the disease present in a population. By taking into account the basic properties of probabilities (e.g. $P(H_1) = 1 - P(H_0)$), Eq. (31) can be easily rewritten in terms of the four costs, sensitivity, specificity and prevalence.

All measures of clinical image quality using the decision theory are based on the truth. This truth can either be the ground truth (the truth is known exactly) or a gold standard (based on for instance the pathology outcome or experts opinion). Human observer studies are valuable as they are able to directly measure clinical image quality. Unfortunately, these methods are time consuming, expensive, and the inter- and intra-observer variability is often large. As a result assessment of clinical image quality is only applied incidentally. These limitations, together with the growing awareness of the importance of the evaluation of clinical image quality, make it more relevant to investigate whether model observers can be used as an objective alternative to human observers. This section is however limited to the discussion of rating scale experiments and m-AFC experiments using human observers. Part 3 provides an in-depth discussion about the use of model observers for this purpose. To gain insight into the decision making process rating scale experiments where observers are asked about their decision confidence can be performed. By varying variation in the decision threshold ROC curves can be drawn. The section "Rating Scale Experiments" provides more in-depth background of rating scale experiments. Another way to deal with observer decision criteria is by using multiple-alternative forced choice (m-AFC) experiments. In m-AFC experiments multiple alternatives are shown to the observer who is asked (forced) to choose the m-alternative which is most likely to contain the signal. This type of experiment will be discussed in detail in the section "Alternative Forced Choice Experiments."

Rating scale experiments

ROC analysis is a quantitative method applicable to a binary decision task. The method results in a graphical plot, the so-called ROC curve (Fig. 11), that illustrates the performance of observers (either humans or computer models) in the detection or classification tasks [50,56–58]. In this chapter we focus on the use of ROC analysis with respect to diagnostic imaging. In diagnostic imaging ROC studies, observers are asked to evaluate different cases and give a confidence about the presence or absence of an abnormality in each case. The TPF and the FPF depend on the choice of the confidence level which results in a positive decision (threshold). Generally, the ROC curve will be determined from the continuous confidence scale by varying the discrimination threshold. However, discrete binary confidence intervals can also be used in ROC analysis. An example of a continuous data experiment could be the assessment of the average CT number of pulmonary nodules from CT images to classify benign from malignant nodules (nodules with higher CT numbers are more likely to be calcified which is a sign of benignity; the average CT number will generate the continuous data). Discrete data could be obtained, for example, in a study with radiologists providing a five-point discrete confidence rating of abnormality concerning a set of normal and abnormal diagnostic images. For examples of ROC analysis used in computed tomography see References 59–61.

Theoretically, ROC curves are continuous and smooth. Unfortunately, the empirically derived ROC curves are most often jagged. Fitting algorithms can aim to create the smoothest curve according to the available data points. A wide range of algorithms is available for this purpose [56]. Often the area under the ROC curve (AUC or Az) is determined as figure of merit for ROC studies. This AUC provides a summary measure of the accuracy of the diagnostic test that is independent of class prevalence (in contrast to accuracy measures mentioned earlier). The AUC would be 1.0 for a perfectly performed test. A test performance that is equal to chance results in an AUC value of 0.5. Sometimes it can be more useful to look at a specific region of the ROC curve rather than at the whole curve. In these scenarios, it is possible to compute partial AUC. For example, one could focus on the region of the curve with a low false positive rate, which could be relevant for population screening tests [56]. The detectability, d_A, related to a rating scale experiment can be derived from the AUC:

$$d_A = \sqrt{2}\Phi^{-1}(AUC) \tag{33}$$

where, $\Phi = \int_{-\infty}^{x} \phi(y)dy$ is the cumulative Gaussian function and $\phi = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$ is a Gaussian function.

If the decision variable distribution is Gaussian under both hypotheses (signal present and signal absent), and their variances are equal, then d_A is equivalent to *d′*.

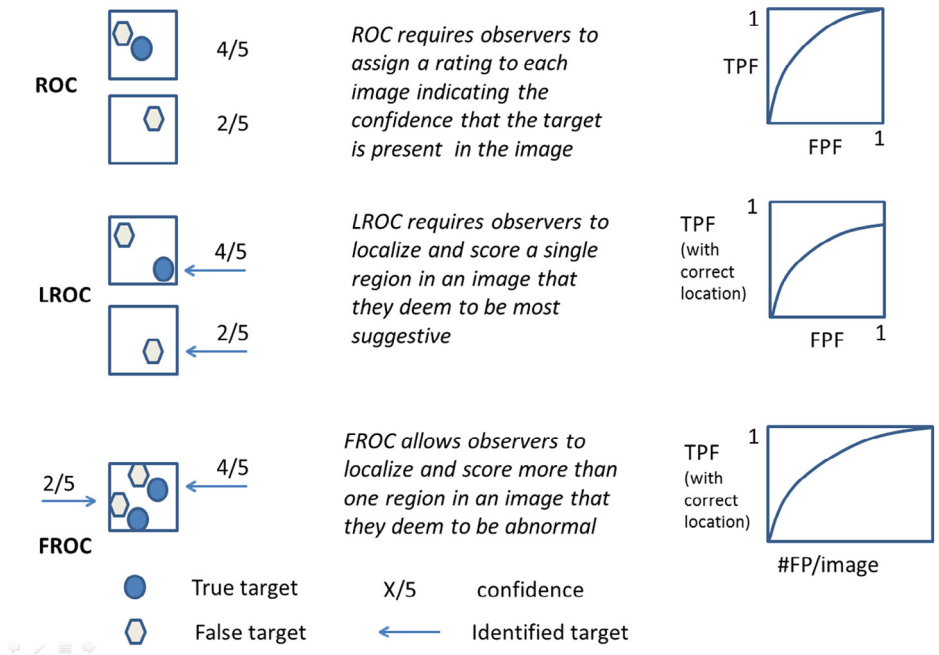

Figure 12. Related methodologies: ROC, LROC, FROC. The task in each of the methods is to give a confidence level concerning the presence of a true target (ROC) eventually in combination with the perceived location (LROC/FROC). In these examples the confidence level scale runs from 1 to 5. A rating of 4 on this scale is given as 4/5 (4 out of 5). Arrows indicate the perceived location.

Several advantages of ROC analysis can be considered. Among these is for instance the fact that the ROC approach provides a simple graphical plot that facilitates visual interpretation of data. Furthermore, depending on the implications of false positive and false negative results, and the prevalence of the condition, one can choose the optimal cut-off for a test from this graph, as the method provides a description of diagnostic accuracy for the full range of sensitivity and specificity. Moreover, two or more tests (for instance radiologists and a Computer Aided Diagnosis (CAD) system) can be compared, for example, analysing the area under each curve (where the better test has the largest AUC) [62]. Shortcomings of ROC analysis are related to its need for specialised computer software (regarding the curve fitting, AUC value calculation and confidence analysis on the ROC curve). Also, large sample sizes may be needed to generate reliable ROC curves. Finally, the ROC methodology does not optimally take the localisation task or the option of multiple abnormalities into account. For this purpose the so-called localisation ROC (LROC) and free response ROC (FROC) have been introduced. Figure 12 gives a graphical impression of the different methods and their concepts. Figure 13 gives a decision tree that illustrates the application of the different methods.

In LROC studies the observers' task is to mark a single location of a suspicious region in each case with a confidence level regarding the observed suspiciousness [56,57,63]. If the marked region is "close enough" to the true abnormal location, the observers' mark is considered a correct localisation. The definition of closeness is not uniformly defined and changes from study to study. Images with no targets (controls, benign, or negative cases) are also scored by marking a "most suspicious" area in the image and by giving this suspicious area a rating (forced localisation choice). To create an LROC curve, the TPF of decisions with correct localisation versus the FPF are plotted. It should be noted that the LROC curve does not necessarily pass the point (1, 1). Unlike the ROC methodology, in LROC the TPF of decisions with correct localisation may well be less than 1.0 at FPF = 1.0 because of incorrect localisations. Similar to the ROC methodology, the area under the LROC curve is considered to be a figure of merit for task performance.

To account for both the localisation and detection of abnormalities in images containing an arbitrary number of them, the free-response ROC (FROC) methodology can be used [56,57,63]. If the localisation mark is within a tolerance range around the true location and the rating of this mark is above a threshold, then a TP is realised. Otherwise a FP decision occurs. The free-response ROC curves are plotted by plotting the TPF (y-axis) versus the number of FP detections per image (x-axis) [64,65].

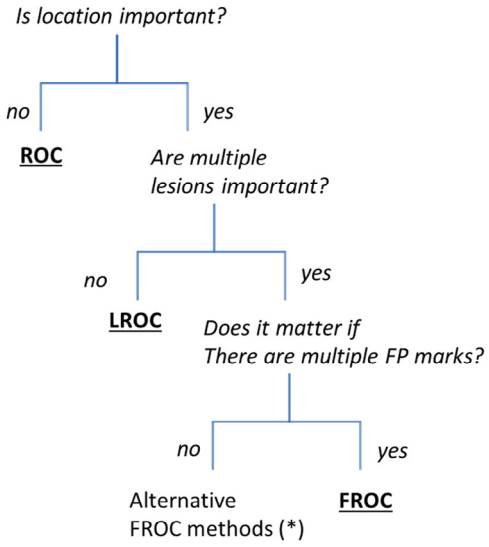

Figure 13. Decision tree illustrating the application of the different methods. The figure is a simplification of a figure provided by Wunderlich and Abbey [63]. Alternative methods (*) concern so-called Alternative FROC (AFROC) methods [54].

Alternative forced choice experiments

In forced choice experiments the observer has to make the decision 'signal present' between alternatives which are offered, even if this means that he has to guess. Compared to ROC studies, m-AFC experiments are faster and easier to perform [66]. However, m-AFC experiments do not provide insight into the underlying distribution functions and the trade-off between sensitivity and specificity [56]. Therefore, m-AFC is sometimes referred to as a poor measure of sensitivity [67].

The natural outcome of m-AFC experiments is a proportion of correct (PC) response. In m-AFC experiments and under assumption of Gaussian distribution of the decision variables (λ), d′ and PC_m of a m-AFC task are related by:

$$PC_m = \int_{-\infty}^{\infty} \Phi^{m-1}(d')\phi(d') \tag{34}$$

where $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$ and $\Phi(x) = \int_{-\infty}^{x} \phi(y)dy$ are respectively Gaussian and cumulative Gaussian functions [68].

This formula can be solved using tabulated values or numerical analysis (standard root finding methods) [69–72]. In the 2AFC experiment, this can be rewritten to:

$$d' = \sqrt{2}\Phi^{-1}(PC_2) \tag{35}$$

For 2-AFC experiments, the PC is equal to the AUC but with human observers, the detectability obtained with the alternative forced-choice paradigm is larger than the detectability obtained with the ROC paradigm [50].

An example of setting for 2-AFC Signal Known Exactly/Background Known Exactly detection experiments is depicted in Fig. 14, where samples with signal present or absent are displayed together with a template of the target.

A detailed comparison and discussion about the use of ROC and AFC experiments as well as the optimum selection of m has been presented by Burgess [66]. This paper concludes that depending on the research question, a deliberate choice between ROC – m-AFC

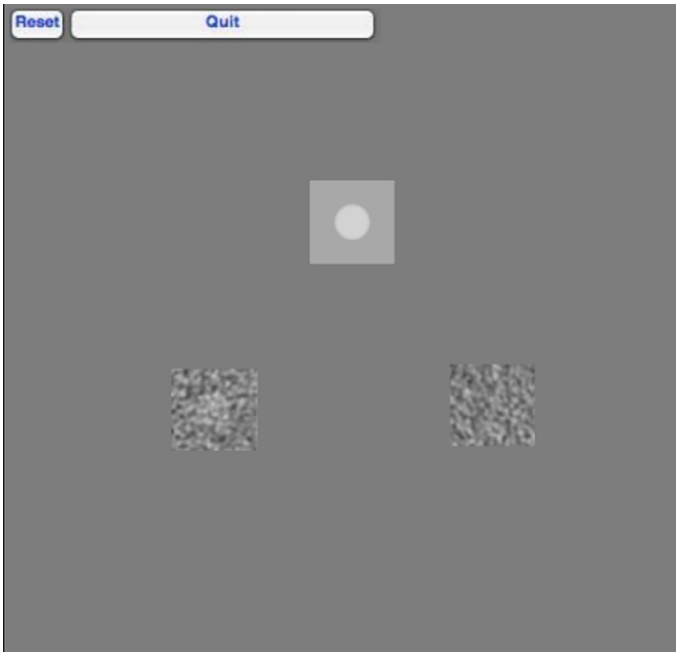

Figure 14. Interface of a 2-AFC human observer SKE/BKE detection experiment.

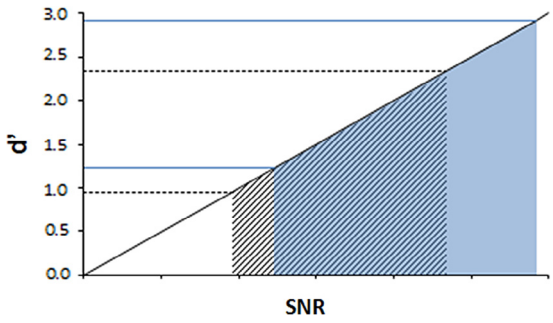

Figure 15. Selecting SNR range for a 2-AFC experiment (dotted black line) and a 4-AFC experiment (solid grey line).

experiments and the value of m is possible. In general m-AFC experiments are chosen if the study goal is to determine how well a certain task can be performed and when there is full control over both the ground truth and the SNR associated with the task. Most commonly m has a value of 2 or 4 but any scalar number larger than two is possible [73]. Burgess has demonstrated that a higher value of m will result in a smaller coefficient of variance. Besides this, he has shown that if d′, for experiments with different values of m, is plotted against the signal to noise ratio (SNR) of the task they will fall on a the same line, independent of m [74]. From this it can be concluded that the choice of m depends essentially on the SNR range for the experiments and the accuracy needed. The SNR range which can be used for an experiment is dictated by the SNR related to the lower threshold (halfway between chance and 1) and 0.95–0.98. This upper level is advised to avoid issues due to observer inattention and their impact on d′ [66,75]. This means that in a 2-AFC experiment, the SNR range should be chosen to result in d′ values between 0.95 and 2.33, whilst this should be between 1.23 and 2.92 for 4-AFC (Fig. 15).

m-AFC experiments can be designed with m independent image combinations or single images which are divided into m areas in which the task can either be signal detection (present-absent) or classification (benign–malignant) [76,77]. Sample sizes for m-AFC experiments are based on the comparison of the expected difference between the PCs of the settings under evaluation for which standard statistical approaches can be followed. m-AFC experiments are based on the signal-known-exactly (SKE) paradigm, which implies that clues should be provided regarding the signal and its position. Therefore, a template of the signal should be visualised together with the m alternatives and an indication of the possible position of the lesion should be indicated. Failure to provide clues on the signal position will result in non-linearity between SNR and observer d′ [66]. Finally, when designing m-AFC experiments care should be taken to avoid bias. For this purpose, the signal should be randomly assigned to one of the m alternatives and the observer PCs should be investigated for the tendencies to favour certain alternatives (e.g. the observer tends to choose left when he is unsure) [66].

Simulated and phantom images are generally well suited to conduct m-AFC experiments because of the full control of ground truth and SNR related to the task [66]. Phantom studies with the m-AFC paradigm are used to evaluate image quality of CT with both human and model observers [77–79]. But also for other modalities m-AFC methodologies are adapted into phantoms for quality control procedures like the CDMAM test object in mammography [80] or the CDRAD for general radiology [81,82].

Yes–no detectability experiments

In yes–no experiments observers only need to decide about the presence of an abnormality. Since yes–no experiments do not provide

insight into the decision-making process of the observer they are not often used for measuring clinical performance very often. In the yes–no experiment the observer inspects one displayed image at a time and must indicate if the signal is present or absent. For a model observer, the yes–no performance is computed by comparing the decision variable to a threshold [50]. If the decision variable is higher than a threshold, the decision is: the signal is present. If the decision variable is less than the same threshold, the decision is: the signal is absent. In this test we assume that the case where the decision variable is equal to a threshold is negligible. With this performance it is possible to obtain four outcomes: true positive (the signal is present and the observer outcome is present), false positive (the signal is absent and the observer outcome is present), true negative (the signal is absent and the observer outcome is absent) and finally false negative (the signal is present and the observer outcome is absent). In the yes–no experiment the detectability index is given by:

$$d_{YN} = \Phi^{-1}(TPF) - \Phi^{-1}(FPF) \tag{36}$$

The TPF represents the True Positive Fraction, and it means the probability given that the signal is truly present in the image. The False Positive Fraction represents the probability that when the signal is absent the observer indicates that the signal is present.

Model observers

ICRU Report 54 suggests that methodologies based on statistical decision theory should be used in medical imaging [58]. Under this framework it is understood that the imaging performance depends on various factors: (1) measures describing the image contrast, image sharpness and the quantity and character of noise; (2) the detailed nature of the diagnostic task, including the clinically important details and the figure of the patient, and the complexities arising from variability between patients; and (3) the degree to which information provided in the image is perceived by the clinician. Points (1) and (2) above are related to the information that is being recorded in the image data, but the ability of the human observer to extract the image information (Point 3 above) may also be an important or even the single limiting factor affecting diagnostic outcome.

Related to this, to simplify image quality assessment, the imaging process is often divided into two separate stages: the first stage consists of the image data acquisition and image formation stage; the second stage consists of the further processing of these data and their actual display to the human observer [58,83]. The first stage can be analysed rigorously by using the concept of the ideal observer, at least in principle and also in practice in simple cases. The ideal observer uses all available information in an optimal way for its decision; the performance of the ideal observer in a given imaging task can then be taken as a measure of the image information related to this task. The ability of the human observer to extract this image information can be measured separately; if the human observer is not able to use the recorded image information this implies leeway –and a need– to improve the image processing or display stage to be better suited to the human observer. This chapter will mainly concentrate on the imaging stage and leave the display stage largely outside the scope; the main aim of this paper is to review methods for evaluating CT scanners and their performance and not the quality of display equipment and display conditions. However, some methods which try to include features of human observers are shortly presented.

The performance of the ideal observer can usually be evaluated only for simplified classification tasks, such as the signal-known-exactly/background-known-exactly case, denoted as SKE/BKE. In this case the ideal observer has all a-priori information of the task, and its performance for classifying images to signal-present and signal-absent cases depends only on the amount of information in the image [58]. The performance of the ideal observer can therefore be taken as a measure of the task-related image information. Other tasks, involving uncertainty of the signal and the background, would be better related to clinical image quality assessment than the SKE/BKE. In such tasks the performance of the observer is not just dependent on the information in the image. The amount of a-priori information about the task that the observer has needs to be taken into account and will affect the performance. It may then sometimes be difficult to quantify the actual effect that this a-priori image information has in the task performance.

Relying on stylised imaging tasks based on the SKE/BKE paradigm may not always be reasonable; see, e.g., Myers et al., where the problem of aperture-size optimisation in emission imaging was considered and it was shown that the optimal aperture would be highly different for the detection of a simple signal in a known background and in a lumpy background [84]. Often, however, it may be considered plausible that the performance of an imaging system in tasks involving incomplete a-priori information could be monotonically related to the outcome in similar detection tasks in the case of full a-priori information (SKE/BKE) [85–88]. This appears to be the case in the paper of Brown et al., where the ideal observer's performance was studied for the signal position unknown case [89]. However, we are still far from completely understanding how a-priori information and the actual image information interact in medical imaging.

In phantom measurements the variability and non-uniformity of real patient images are not usually present. In the SKE/BKE paradigm any background structure is treated as being a deterministic known structure, which does not impair detail detectability. This may not always be realistic for a human observer, whose detection performance may in some cases be more impaired because of background variability than because of actual stochastic noise [90–94], but is certainly applicable to the ideal observer. Human observers seem to operate somewhere between two interpretations: background variability appears to function as a mixture of noise and deterministic masking components. For a more detailed discussion on this matter, see, e.g. Burgess and the references therein [91].

For a thorough presentation of modern image science, see the book by Barrett and Myers [57] and by Samei and Krupinski [56]. Another useful handbook on imaging systems, image quality and measurements has been published by the International Society for Optical Engineering [50]. Also, a discussion and review of task-based methods for assessing the quality of iteratively reconstructed CT images have been published recently [25]. They conclude that Fourier-based metrics of image quality are convenient and useful in many contexts, e.g., in quality assurance, but the assessment of iteratively reconstructed CT images requires more sophisticated methods which do not rely on assumptions of system linearity and noise stationarity; these assumptions are prerequisites in the Fourier-based methods [95–97].

Linear observers

Mathematical theory

A linear observer can be described with a decision statistic $\lambda(g)$ which is a linear function of the image data, instead of being a more general function. In the vector notation of images this can be written as an inner product of a template $\mathbf{w}$ and the image $\mathbf{g}$

$$\lambda(\mathbf{g}) = \mathbf{w}^T\mathbf{g} \tag{37}$$

The non-zero elements of the template correspond to image locations where the pixel value needs to be taken into account, and by what weight. The weight can be either positive or negative. Pixels

with the value zero in the template do not influence the decision statistic at all, and the observer considers the data in those pixels to be irrelevant for the decision.

The importance and frequent use of linear observers stems mainly from their manageability and ease of use. Further, as was seen in the preceding chapter, the ideal observer of many cases may be obtained in a linear form. This is not the case for all detectability tasks, however. For example, the ideal detection in the case involving uncertainty of the signal position will result in a non-linear test statistic (see, e.g., Brown et al. [89]). A linear observer for this task would consist just of a template which is obtained as the convolution of the pdf of the signal position and shape. Therefore, essentially, this observer would measure only the mean brightness of the image and it seems clear that it would be much less efficient than a human observer, for example.

In order to compute the SNR of a linear observer, we first need to express the mean response under hypothesis H_j as well as its associated variance:

$$\begin{aligned}\bar{\lambda}_j &= \langle \lambda(\mathbf{g})|H_j\rangle = \mathbf{w}^T\langle \mathbf{g}|H_j\rangle \\ \sigma_j^2 &= \left\langle \left(\mathbf{w}^T\mathbf{g} - \langle \mathbf{w}^T\mathbf{g}|\mathbf{H_j}\rangle\right)^2 \middle| H_j \right\rangle = \mathbf{w}^T\mathbf{K}_j\mathbf{w}\end{aligned} \tag{38}$$

This allows us to easily express the signal to noise ratio of a linear observer by injecting Eq. (38) into Eq. (31):

$$SNR_\lambda^2 = \frac{\left(\mathbf{w}^T(\langle \mathbf{g}|H_1\rangle - \langle \mathbf{g}|H_0\rangle)\right)^2}{\mathbf{w}^T\frac{1}{2}(\mathbf{K}_0 + \mathbf{K}_1)\mathbf{w}} \tag{39}$$

Here, it is important to recall the assumptions required for Eq. (39) to be meaningful. First, this requires that the conditional distributions of λ are normal. This is the case at least when the noise in the images is multivariate normal. Secondly, if the covariance matrices for the signal and background cases are different, the SNR does not define the entire ROC curve, but the area under the ROC curve and the percentage of correct answers in a two-alternative forced-choice test using the same images are still specified by the SNR. An inequality of covariance matrices $\mathbf{K}_0$ and $\mathbf{K}_1$ would also infer that a linear observer is not ideal, and may fall far beyond the true ideal observer [98]; however, if measured covariance data are used, it is useful to improve the precision of the **K**-estimate by including both measured covariance, $\mathbf{K_0}$ and $\mathbf{K_1}$.

By inserting the w-templates of the PWMF and the NPWMF to Eq. (39) we obtain the well-known expressions for their SNR

$$SNR^2{}_{PWMF} = \mathbf{s}^T\mathbf{K}^{-1}\mathbf{s} = \mathbf{S}^T\mathbf{W}^{-1}\mathbf{S} \tag{40}$$

and

$$SNR^2{}_{NPWMF} = \left(\mathbf{s}^T\mathbf{s}\right)^2 / \mathbf{s}^T\mathbf{K}\mathbf{s} = = \left(\mathbf{S}^T\mathbf{S}\right)^2 / \mathbf{S}^T\mathbf{W}\mathbf{S} \tag{41}$$

where we have denoted the Fourier transform of **s** by **S** and that of matrix **K** by **W**. If the noise is stationary, **W** is a diagonal matrix and its diagonal values represent the NPS. Then, decomposing the SNR^2 to components: each frequency k contributes by amount

$$SNR^2{}_{PWMF,k} = I|S_k|^2 / W_k \tag{42}$$

to the total $SNR^2{}_{PWMF}$. This simplicity is lost if W is not diagonal.

The best possible linear observer is called the Hotelling observer. The Hotelling observer is equal to the PWMF in the case of signal-independent (additive), normally distributed noise and both of these reduce to the NPWMF, when the noise is white. As discussed above, the Hotelling observer may also fall far below ideal performance, for example, in the signal position unknown case, where the ideal decision statistic is not a linear function of image data [89].

The strategy of the ideal observer may be complicated by **K** not being diagonal. However, in the case of uncorrelated image noise the strategy is self-evident: the ideal observer then just looks more keenly to image pixels where the presence of the signal is known to have a strong effect and where the uncertainty of the measurement (noise) is small. Image areas that are not affected by signal presence need not be observed at all. This same interpretation applies to the case of coloured, stationary noise as well; then the Fourier transformed data will have a diagonal covariance matrix, where the diagonal elements constitute the noise power spectrum. In this case the ideal observer puts more emphasis on spatial frequencies where the signal presence makes a large contribution and less emphasis on frequencies which contain more noise.

If the image noise is not white, the NPWMF observer is suboptimal because it does not take into account the noise correlations between pixels, or equivalently, the different noise power at various spatial frequencies. Therefore, in this case, the observer is not tuned against the noise similarly as the ideal observer and it shows a penalty of this in its performance. However, if the frequency spectrum of the signal is concentrated on a relatively narrow band of frequencies where the frequency dependence of the NPS is modest, one can expect the NPWMF observer to perform nearly as well as the ideal PWMF does. This may happen, for example, when the signal to be detected does not have sharp details and is of a relatively large size.

However, note that by definition, the NPWMF believes that the background level is equal in all images and therefore needs not be observed. The NPWMF measures the image intensity only in the pixels that belong to the expected signal position and discards the data in all other pixels. For a disk signal this would be equivalent to observing just the total image intensity of the signal disk area and masking away all other image areas: no reference to the contrast between the signal and the background is made. If in fact, there is any – even small – variation in the background level from image to image, or if there is any low-frequency background variability (e.g., variable lumpy background structures) which in reality can have an effect on the image intensity in the signal detail area, the NPWMF can be considered as being a misled observer, which will perform extremely poorly and often performs worse than human observers. This was the case, for example, in a paper that considered signal detection in added low-pass correlated noise and found that the NPWMF observer was very inefficient and even humans significantly outperformed it [99]. This and other similar results greatly diminished the interest in the NPWMF observer.

To improve this situation, Tapiovaara and Wagner [98] introduced the DC-suppressing observer, which leaves the average brightness of the image (or the zero-frequency channel) outside of the decision.[1] This observer is achieved by subtracting the mean pixel value of the NPWMF-template from every pixel of the template

$$\lambda_{DCS} = \left[\mathbf{s} - \left(N^{-1}\Sigma s_k\right)\mathbf{1}\right]^T\mathbf{g} \tag{43}$$

Here, N is the number of pixels in the analysed image area and **1** denotes a vector with all elements equal to unity. In the Fourier domain this observer is:

$$\lambda_{DCS} = \left[\mathbf{S} - S_0\mathbf{e}^0\right]^T\mathbf{G} = \sum_{k=1}^{N-1} S_k^* G_k \tag{44}$$

[1] In practical imaging measurements one often does not analyse the whole image area, but considers only a relatively small sub-area containing the signal and a reasonable surround of it. Then the image vector **g** corresponds to this sub-area, and the zero-frequency of this image data includes contributions from very low-frequencies in addition to the strict zero-frequency of the whole image data.

This modification of the NPWMF-observer turned out to be crucial for the performance of the observer in measurements of fluoroscopic imaging, where excess noise in the mean image brightness strongly and variably impaired the performance of the NPWMF [100]. This zero-frequency variability can be assumed to be common in other fields of radiology as well: the exact mean image brightness is not probably an important diagnostic feature in any imaging modality, and, on the other hand, if there is excess variability in the brightness, including it – as the NPWMF does – will result in a notable performance penalty. Such a variability in average brightness can be seen as a delta spike at the origin of the NPS and can be properly weighted by the PWMF, of course. However, in many recipes for measuring the NPS, the DC-component is normalised out and therefore equals zero in the NPS results (e.g. Boedeker et al. [101]). Whilst noiseless data in real systems are not realistic, it is then important not to include the zero frequency signal component in the SNR calculation either.

Non-prewhitening with eye filter

Another modification of the NPWMF includes filtering of the image with an eye-filter, intended to obtain a better agreement of the performance of this model observer and human observers. The observer is often denoted as NPWE [102] (a similar observer model has been presented earlier by Loo et al. [103]). This observer is usually expressed in the spatial frequency domain and the eye filter **E** mimics the visual spatial frequency response function (or the contrast sensitivity function) of the human eye. The application of **E** requires knowing the dimension of the image and the viewing distance. The decision function of this observer is then:

$$\lambda_{NPWE} = [\mathbf{ES}]^T \mathbf{EG} = \mathbf{S}^T \mathbf{E}^T \mathbf{EG} \tag{45}$$

It is noted here that the eye filter also suppresses the zero-frequency, like the DCS-observer above, but the NPWE observer also takes very low frequencies into account with a low weighting. This is the main factor for the NPWE observer performing much better than the NPWMF in studies involving excess noise in very low frequencies [25,102]. This means that the usefulness of this observer model may actually be more related to its suppressing of low-frequency noise than in its attempt to mimic human vision.

As an example of NPWE performance, Fig. 16 shows the detectability index (d′) or SNR as a function of object diameter for the 0.5% contrast group of the Catphan and three mAs levels acquired in a Toshiba Aquilion ONE 320 detector-row CT scanner. The NPWE detectability improved with increasing mAs, as the noise level of the images decreased, for all the objects [50].

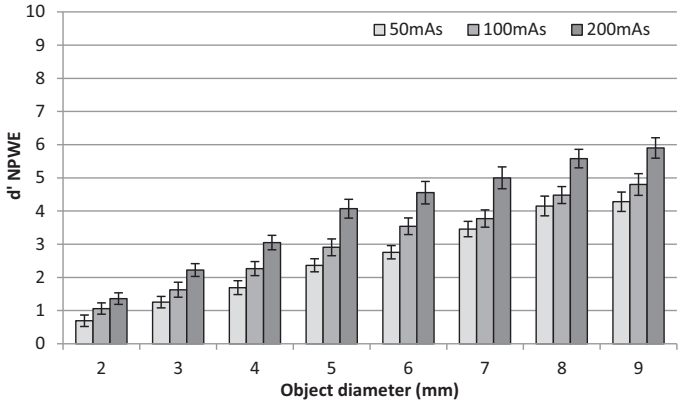

Figure 16. Detectability index (d′) as a function of object diameter for the different levels of mAs for the 0.5% contrast group (2-9 mm) in the Catphan 600 Phantom (Phantom Laboratories, New York).

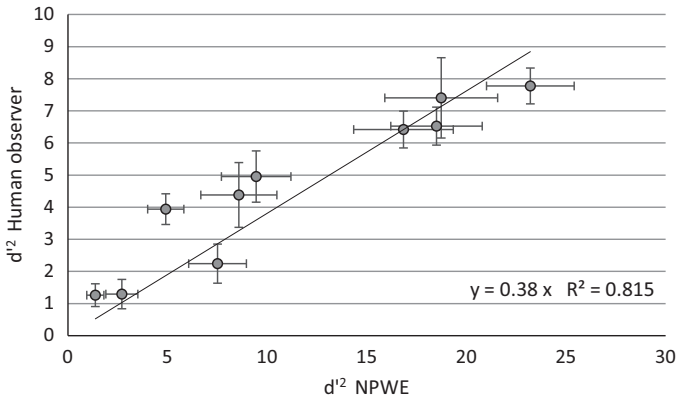

Figure 17. Detectability index (d′) of the average human observer as a function of the NPWE model observer d′, both squared, for 1% contrast objects and all dose values. The efficiency, η, tallies the slope of the linear fit.

The detectability index is given when two assumptions are verified [73]. Firstly the template responses must be Gaussian and secondly the template responses are statistically independent [90]. This performance is given in terms of distance in standard deviation units between the signal distribution and the noise distribution.

$$d' = \frac{\langle \lambda_s \rangle - \langle \lambda_n \rangle}{\sigma_\lambda} \tag{46}$$

where λ_s is the mean model response to the signal, and λ_n is the mean model response to the background. σ_λ is the standard deviation of the model response.

The advantage of this metric is that it computes directly from the image statistic.

Model observers can also be otherwise modified in order to mimic human performance better, for example, by including internal noise [104,105]. Internal noise degrades the model's performance, and takes into account the fact that human observers have "noise" by not giving necessarily the same answer when a certain image is presented twice or more to be scored [106]. Many approaches can be used to decrease the model's performance, and each has pros and cons [105,107]. Such models are of interest in efforts to reproduce the efficiency of the visual detection performance of humans, but are not explained in this review. In Fig. 17 the PC values were translated into d′ and an efficiency (η) was calculated to normalise the model observer results, fitting d′ human as a function of d′NPWE, both squared. As the curve representing d′ as a function of PC saturates above 3 (PC ≈ 0.98) for 2-AFC experiments, only the values below this threshold were taken into account [108].

Channelised Hotelling Observer

Another type of linear observer models is the Channelised Hotelling Observer [109] (CHO) either with or without internal noise; only the latter model is considered here. A thorough treatment of both can be found in Abbey and Barrett [110]. The motivation for this observer results both from its effect in reducing the image data from a large number of pixels to a much lower number of scalars, called channel responses and by the ability of such models to mimic the detectability results of human observers. If the channels in the model are selected such that they help in the tuning against the noisy frequencies without losing too much of the signal energy they may also provide an improvement over the non-prewhitening observer types and a useful approximation for the ideal observer. The reduction of dimensionality especially simplifies computing and inversing the covariance matrix.

The CHO does not have access directly to the pixel values (or the Fourier transform) of the image. Instead, first the image data (**g**) are linearly combined to a small number of channelised data (**u**) by multiplication with a matrix **T**:

$$\mathbf{u} = \mathbf{T}^{\mathrm{T}}\mathrm{g} \tag{47}$$

Here the column vectors of **T** represent the spatial profiles of the channels. These channelised data are then combined with a weighting template **v** to a linear decision function:

$$\lambda = \mathbf{v}^{\mathrm{T}}\mathbf{u} \tag{48}$$

If the noise in the image data **g** is Gaussian, it is also Gaussian in the channel **u**, and we already know that the ideal observer (which, however, has access only to the channelised data) is obtained with weighting $\mathbf{v}^{\mathrm{T}} = (\mathbf{u}_1 - \mathbf{u}_0)^{\mathrm{T}}\mathbf{K}_{\mathbf{u}}^{-1}$, and the decision function of this observer is:

$$\lambda_{\mathrm{CHO,T}} = (\mathbf{u}_1 - \mathbf{u}_0)^{\mathrm{T}}\,\mathbf{K}_{\mathrm{u}}^{-1}\mathbf{u} = (\mathbf{u}_1 - \mathbf{u}_0)^{\mathrm{T}}\,\mathbf{K}_{\mathrm{u}}^{-1}\mathbf{T}^{\mathrm{T}}\mathbf{g} \tag{49}$$

Above, the channels were presented in the image domain. Usually, however, the channels are specified in the frequency domain, and may be either non-overlapping frequency intervals or overlapping functions of various forms, such as sparse or dense difference-of Gaussians, Laguerre–Gauss polynomials or other functions [109,111].

Note that in the case of stationary image noise the non-overlapping channel models result also to a diagonal covariance matrix, because the frequency channels remain independent, whereas the overlapping channels cause correlations in the noise. If one prefers working in the image domain, one can obtain the spatial representations of the frequency selective channels by taking the inverse Fourier transforms of the latter.

In image quality assessment when using these channelised models it is important to note that the channelised Hotelling observer can adapt to the signal and the image covariance only after they have passed through **T**. Then, for example, the observer is sensitive only to signals that cause a change in the channelised signal $\mathbf{T}^{\mathrm{T}}\mathbf{s}$ (or, equivalently, in the frequency domain representation). For sparse channel models with just a few channels, a significant loss of information may occur in the formation of the channel responses [110].

Also, these observers are typically zero-frequency suppressing, although, being tuned against the noise in the different channels, they could also otherwise handle variability in the average image brightness better than the NPWMF. This would require, however, that if zero-frequency is included in the lowest frequency channel, not much of the important signal energy shall be included in this channel.

Usually, in applications related to medical imaging, the channels are defined to be cylindrically symmetric and are specified in terms of the radial frequency. The use of such models is usually restricted to image signals that are also cylinder-symmetric. Channelised Hotelling observers have been used with good success to predict the performance of human observers in detection tests.

As an example, Fig. 18 shows the CHO performance (detectability index (d′)) with dense of difference of Gaussian for an 8 mm sphere at 20 HU of the QRM 401 phantom and three $\mathrm{CTDI_{vol}}$ levels acquired in GE HD 750 CT scanner.

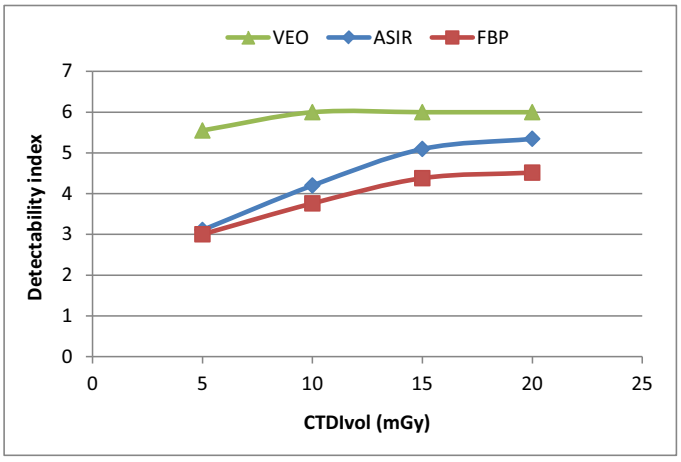

Figure 18. Detectability index (d_A) as a function of $\mathrm{CTDI_{vol}}$ for the different algorithms for a sphere of Ø 8 mm and a contrast with background of 20 HU in the QRM 401 Abdomen Phantom (QRM, Moehrendorf, Germany).

Agreement between observers

The first step to compare model observers and model/human observers is to have the same metrics to measure their performance. For a specific task, background, signal and model the investigator must choose between the area under the curve (AUC), sensitivity/specificity pairs, the percent correct (PC), the signal to noise ratio (SNR) or the detectability index (d′), then a comparison is possible.

Kappa test

To measure the agreement between observers it is common to use the Kappa coefficient. When observers are two or more the inter-observer variation can be computed. The Kappa test is based on the difference between the observer agreement (percentage where observers agree among themselves) and the expected agreement (agreement obtain just by chance). The formula for the Kappa test is then as follows:

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \tag{50}$$

where p_0 is the relative observed agreement among reviewers, and p_e is the probability of chance agreement.

The Kappa scale ranges from –1 to 1. 1 represents a perfect agreement; 0, the agreement is obtained just by chance; and –1 represents a systematic disagreement. A generic scale proposed by Landis and Koch is used to help the investigator to interpret the Kappa coefficient (Table 3) [112].

The Kappa coefficient estimated itself could be obtained just by chance, so a P value can be calculated to interpret the result of the Kappa test. The P value is sensitive to sample size, so another Kappa test can be used to interpret the result, the weighted Kappa assigns weighting more or less important to different categories, to focus on categories where the difference is significant. But the weighting is defined by the investigator, and the expert can disagree on the tuning of the weighted Kappa. The Kappa test is used to interpret the agreement, but this test is affected by the prevalence of the disease [50] (Fig. 19); in rare cases a low Kappa test does not reflect a low agreement. Moreover, the Kappa test can give strange results when the observers have a high degree of agreement and when they are close to PC = 1.

Table 3
Genetic scale investigator to interpret the Kappa coefficient.

0.01–0.20	Slight agreement
0.21–0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–0.99	Almost perfect agreement

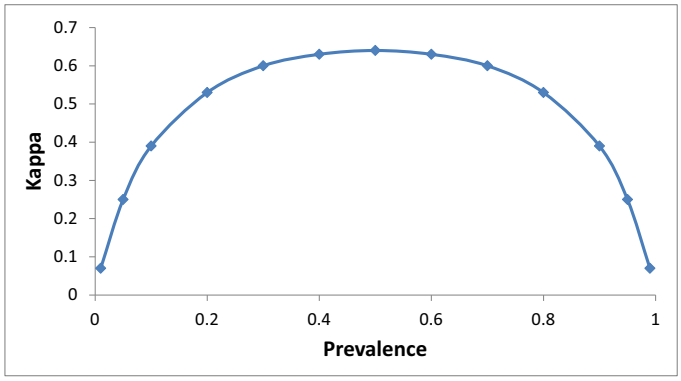

Figure 19. Kappa coefficient in terms of prevalence.

Bland–Altman test

A Bland–Altman plot is often used to compare results between model observers and human observers [113]. When both observers measure the same parameter (i.e. d′ or PC) with the same images, most of the time the correlation is good [57,108]. A good correlation for two observers that measure the same parameter does not imply a good agreement between the two observers.

A Bland–Altman plot shows the mean of the two observers in the abscissa, and the difference between the two observers in the ordinate. The limits of agreement are defined by the mean of the difference and the standard deviation of the difference. If a method is the gold standard then d represents the bias, whereas if any methods are standard, d represents only systematic differences. Figure 20 shows an example comparing the performance of the NPWE model and human observers for a given detection task.

Conclusion and perspectives

Since the introduction of CT many efforts have been made to balance image quality with patient exposure. Image quality was first assessed using signal detection theory, and basic parameters such as image noise and spatial resolution, which made it possible to evaluate the strengths and weaknesses of acquisition protocols. With the technological developments of CT it became necessary to assess units in order to objectively enhance the benefit of new technological solutions. Global figures of merit of image quality were derived, still using signal theory functions, normalising the result by a standardised dose indicator: the $CTDI_{vol}$. If this approach seems enticing one has to remember that the use of one number to judge image quality is a simplified solution that can lead to false conclusions. Moreover, image quality assessment methods based on signal theory only do not include a clinically relevant task. With this kind of approach one could optimise aiming at getting the best theoretical image quality, rather than ensuring that images convey the relevant clinical information to make a correct diagnosis. In such a context, image quality assessment in the field of medical imaging should be task oriented and clinically relevant.

The use of mathematical model observers may be an appropriate solution, opening a way forward, even if the tasks investigated remain very simple and far from clinical reality. As shown in the review, there are several types of model observers, and the choice of a single solution might not be optimal. The disadvantage of model observers is that they are defined for simple situations, like the detection of a representative signal in a given phantom, and surely do not cover the whole range of characteristics that define image quality at the clinical level. This drawback can nonetheless become an advantage because their calculation can be kept relatively simple; they are objective and compatible with new image reconstruction techniques such as iterative reconstruction. They also lead to reproducible results which can be representative of human perception whilst avoiding the burden of actual studies with human observers. They could be used to compare clinical protocols in terms of image quality and dose levels to initiate an optimisation process. Nevertheless, more studies should be performed in the future on correlations between model observer outcomes and human diagnostic accuracy.

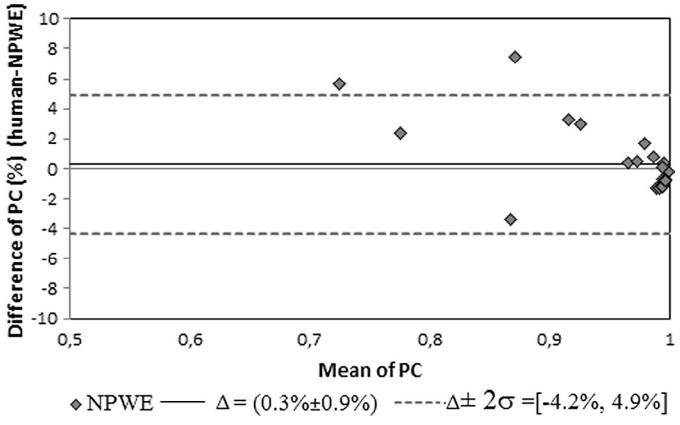

Figure 20. Bland–Altman plot of proportion correct (PC) difference between human and NPWE for 1% contrast and all mAs. The straight line represents the average difference (Δ) and the dash lines, the range of the differences [Δ ± 2σ], where σ is the standard deviation of the differences. The NPWE model was corrected by an efficiency of 0.38.

Acknowledgements

This work was supported by the German Radiation Protection Agency (BfS – UFO-Plan Vorhabens 3613S20007) and the Swiss National Science Foundation for research (FNS 320030_140995/1). The authors are thankful to R. van Engen from LRCB, Nijmegen and Pr H. Bosmans from Katholieke Universiteit, Leuven, and P. Monnin from CHUV, Lausanne, for their help for improving the structure of the manuscript.

References

[1] Thurston J. NCRP report no. 160: ionizing radiation exposure of the population of the United States. Phys Med Biol 2010;55:6327. doi:10.1088/0031-9155/55/20/6327.
[2] Aubert B, Sinno-Tellier S, Etard C. Exposition de la population française aux rayonnements ionisants liée aux actes de diagnostic médical en 2007. Institut de Radioprotection et de Sûreté Nucléaire et the Institut de Veille Sanitaire: 2010.
[3] Samara ET, Aroua A, Bochud FO, Ott B, Theiler T, Treier R, et al. Exposure of the Swiss population by medical x-rays: 2008 review. Health Phys 2012;102:263–70.
[4] Federal Ministry for the Environment, Nature Conservation, Building and Nuclear Safety. Bernhard-Ströl C, Hachenberger C, Trugenberger-Schnabel A, Loebke-Reinl A, Peter J, editors. Umweltradioaktivitaet und strahlenbelastung. Annual report. 2012.
[5] IRSN. Exposition de la population française aux rayonnements ionisants liés aux actes de diagnostic médical en 2012. Rapport PRP-HOM N°2014-6. 2012. <http://www.irsn.fr/FR/Actualites_presse/Communiques_et_dossiers_de_presse/Pages/20141013_Rapport-Expri-Exposition-rayonnements-diagnostic-medical.aspx#.VOHGuuK1Z8E> [accessed 16.02.15].
[6] Boone J, Strauss K, Cody D, McCollough C, McNitt-Gray M, Toth T. Size-specific dose estimates (SSDE) in pediatric and adult body CT exams. Report of AAPM Task Group 204. 2011.
[7] Nguyen TTA, Le HND, Vo M, Wang Z, Luu L, Ramella-Roman JC. Three-dimensional phantoms for curvature correction in spatial frequency domain imaging. Biomed Opt Express 2012;3:1200–14. doi:10.1364/BOE.3.001200.
[8] Mironov V, Boland T, Trusk T, Forgacs G, Markwald RR. Organ printing: computer-aided jet-based 3D tissue engineering. Trends Biotechnol 2003;21:157–61. doi:10.1016/S0167-7799(03)00033-7.

[9] Solomon J, Samei E. Quantum noise properties of CT images with anatomical textured backgrounds across reconstruction algorithms: FBP and SAFIRE. Med Phys 2014;41:091908. doi:10.1118/1.4893497.
[10] Tapiovaara MJ, Wagner R. SNR and DQE analysis of broad spectrum X-ray imaging. Phys Med Biol 1985;30:519. doi:10.1088/0031-9155/30/6/002.
[11] Nickoloff EL, Riley R. A simplified approach for modulation transfer function determinations in computed tomography. Med Phys 1985;12:437–42.
[12] Boone JM. Determination of the presampled MTF in computed tomography. Med Phys 2001;28:356–60.
[13] Judy PF. The line spread function and modulation transfer function of a computed tomographic scanner. Med Phys 1976;3:233–6.
[14] Nakaya Y, Kawata Y, Niki N, Umetatni K, Ohmatsu H, Moriyama N. A method for determining the modulation transfer function from thick microwire profiles measured with x-ray microcomputed tomography. Med Phys 2012;39:4347–64. doi:10.1118/1.4729711.
[15] Thornton MM, Flynn MJ. Measurement of the spatial resolution of a clinical volumetric computed tomography scanner using a sphere phantom, vol. 6142. 2006. p. 61421Z doi:10.1117/12.654969.
[16] Grimmer R, Krause J, Karolczak M, Lapp R, Kachelriess M. Assessment of spatial resolution in CT. IEEE Nucl Sci Symp Conf Rec 2008;5562–6. doi:10.1109/NSSMIC.2008.4774508. 2008 NSS 08.
[17] Friedman SN, Fung GSK, Siewerdsen JH, Tsui BMW. A simple approach to measure computed tomography (CT) modulation transfer function (MTF) and noise-power spectrum (NPS) using the American College of Radiology (ACR) accreditation phantom. Med Phys 2013;40:051907. doi:10.1118/1.4800795.
[18] International Commission on Radiation Units and Measurements. ICRU Report No. 87: radiation dose and image-quality assessment in computed tomography. J ICRU 2012;12:1–149. doi:10.1093/jicru/ndt007.
[19] Miéville F, Beaumont S, Torfeh T, Gudinchet F, Verdun FR. Computed tomography commissioning programmes: how to obtain a reliable Mtf with an automatic approach? Radiat Prot Dosimetry 2010;ncq050. doi:10.1093/rpd/ncq050.
[20] Dainty JC, Shaw R. Image science: principles, analysis and evaluation of photographic-type imaging processes. Academic Press; 1974.
[21] Thibault J-B, Sauer KD, Bouman CA, Hsieh J. A three-dimensional statistical approach to improved image quality for multislice helical CT. Med Phys 2007;34:4526–44. doi:10.1118/1.2789499.
[22] Hsieh J, Nett B, Yu Z, Sauer K, Thibault J-B, Bouman CA. Recent advances in CT image reconstruction. Curr Radiol Rep 2013;1:39–51. doi:10.1007/s40134-012-0003-7.
[23] Richard S, Li X, Yadava G, Samei E. Predictive models for observer performance in CT: applications in protocol optimization, vol. 7961. 2011. p. 79610H doi:10.1117/12.877069.
[24] Ott JG, Becce F, Monnin P, Schmidt S, Bochud FO, Verdun FR. Update on the non-prewhitening model observer in computed tomography for the assessment of the adaptive statistical and model-based iterative reconstruction algorithms. Phys Med Biol 2014;59:4047–64. doi:10.1088/0031-9155/59/4/4047.
[25] Vaishnav JY, Jung WC, Popescu LM, Zeng R, Myers KJ. Objective assessment of image quality and dose reduction in CT iterative reconstruction. Med Phys 2014;41:071904. doi:10.1118/1.4881148.
[26] Richard S, Husarik DB, Yadava G, Murphy SN, Samei E. Towards task-based assessment of CT performance: system and object MTF across different reconstruction algorithms. Med Phys 2012;39:4115–22. doi:10.1118/1.4725171.
[27] Brunner CC, Abboud SF, Hoeschen C, Kyprianou IS. Signal detection and location-dependent noise in cone-beam computed tomography using the spatial definition of the Hotelling SNR. Med Phys 2012;39:3214–28. doi:10.1118/1.4718572.
[28] Miéville FA, Bolard G, Bulling S, Gudinchet F, Bochud FO, Verdun FR. Effects of computing parameters and measurement locations on the estimation of 3D NPS in non-stationary MDCT images. Phys Med 2013;29:684–94. doi:10.1016/j.ejmp.2012.07.001.
[29] Edyvean S. Understanding image quality and dose. ImPACT Feb. 2007. <http://www.impactscan.org/slides/course07/lect9/frame.htm> [accessed 29.04.15].
[30] Seeram E. Computed tomography: physical principles, clinical applications, and quality control, 3e. 2nd ed. Saunders; 2000.
[31] Edyvean S. The relationship between image noise and spatial resolution of CT scanners, <http://www.ctug.org.uk/meet02/noiseandspatialresct.pdf>; 2002 [accessed 11.03.15].
[32] Edyvean S, Keat N. Comparison of CT scanner image noise, image width, dose and spatial resolution using standard test methods. <http://www.aapm.org/meetings/04AM/pdf/14-2350-75226.pdf>; 2004 [accessed 11.03.15].
[33] Brooks RA, Di Chiro GD. Statistical limitations in x-ray reconstructive tomography. Med Phys 1976;3:237–40. doi:10.1118/1.594240.
[34] Riederer SJ, Pelc NJ, Chesler DA. The noise power spectrum in computed X-ray tomography. Phys Med Biol 1978;23:446–54.
[35] Atkinson JK. The quantitative assessment of CT scanners. University of London; 1980.
[36] Edyvean S. ImPACT MDA Report type testing of CT scanners – methods and MEthodology, <http://www.impactscan.org/reports/MDA9825.htm>; 1998 [accessed 11.03.15].
[37] Fuchs T, Kalender WA. On the correlation of pixel noise, spatial resolution and dose in computed tomography: theoretical prediction and verification by simulation and measurement. Phys Med 2003;XIX(2):153–64.
[38] Kalender WA. Computed tomography: fundamentals, system technology, image quality, applications. 3rd ed. Erlangen: Publicis; 2011.
[39] Bassano DA. Specification and quality assurance for CT scanners. AAPM Summer Sch 1980.
[40] Allisy-Roberts P, Williams JR. Farr's physics for medical imaging. Elsevier Health Sciences; 2007.
[41] Edyvean S. ImPACT NHS PASA 16 slice CT scanner comparison report version 14. <http://www.impactscan.org/reports/Report06012.htm>; 2006 [accessed 11.03.15].
[42] Edyvean MS. A methodical approach for comparison of CT scanner image quality relative to dose. Radiological Society of North America 2003 Scientific Assembly and Annual Meeting, November 30–December 5, 2003, Chicago, IL. <http://archive.rsna.org/2003/3107396.html> [accessed 29.04.15].
[43] Platten D, Keat N, Lewis M, Barret J, Edyvean S. ImPACT MHRA 04045 Toshiba 16 Report, page 26, <http://www.impactscan.org/reports/MHRA04045.htm>; 2003 [accessed 30.03.15].
[44] Nagel H-D. CT dose efficiency parameters. European medical ALARA network. WG 1: optimisation of patient exposure in CT procedures, synthesis document 2012. <https://www.yumpu.com/en/document/view/42382097/wg1-synthesis-report-pdf-a-11-mb-european-medical-alara-/5>; 2004 [accessed 17.02.15].
[45] Chao EH, Toth TL, Williams EC, Fox SH, Carleton CA, Bromberg NB. A statistical method of defining low contrast detectability, poster presented at RSNA Meeting; 2000.
[46] Rose A. Vision: human and electronic. New York: Plenum Press; 1973.
[47] Torgersen GR, Hol C, Møystad A, Hellén-Halme K, Nilsson M. A phantom for simplified image quality control of dental cone beam computed tomography units. Oral Surg Oral Med Oral Pathol Oral Radiol 2014;118:603–11. doi:10.1016/j.oooo.2014.08.003.
[48] Nagel H-D. Methoden zur bestimmung der dosiseffizienz von CT-scannern, presented at APT meeting; 2008.
[49] Månsson LG. Methods for the evaluation of image quality: a review. Radiat Prot Dosimetry 2000;90:89–99.
[50] Beutel J, Kundel HL, Van Metter RL. Handbook of medical imaging: physics and psychophysics. SPIE Press; 2000.
[51] Samei E, Badano A, Chakraborty D, Compton K, Cornelius C, Corrigan K, et al. Assessment of display performance for medical imaging systems: executive summary of AAPM TG18 report. Med Phys 2005;32:1205–25. doi:10.1118/1.1861159.
[52] Likert R. A technique for the measurement of attitudes. Arch Psychol 1932;22(140):55.
[53] Jamieson S. Likert scales: how to (ab)use them. Med Educ 2004;38:1217–18. doi:10.1111/j.1365-2929.2004.02012.x.
[54] Norman G. Likert scales, levels of measurement and the "laws" of statistics. Adv Health Sci Educ 2010;15:625–32. doi:10.1007/s10459-010-9222-y.
[55] International Commission on Radiation Units and Measurements. Receiver operating characteristic analysis in medical imaging, vol. 79. ICRU Rep. N°79. Bethesda (MD): International Commission on Radiation Units and Measurements; 2008.
[56] Samei E, Krupinski E. The handbook of medical image perception and techniques. Cambridge UK: Cambridge University Press; 2014.
[57] Barrett HH, Myers KJ. Foundations of image science. Hoboken (NJ): Wiley-Interscience; 2004.
[58] Vennart W. ICRU Report 54: medical imaging – the assessment of image quality. Radiography 1996;3:243–4. doi:10.1016/S1078-8174(97)90038-9.
[59] Way T, Chan H-P, Hadjiiski L, Sahiner B, Chughtai A, Song TK, et al. Computer-aided diagnosis of lung nodules on CT scans: ROC study of its effect on radiologists' performance. Acad Radiol 2010;17:323–32. doi:10.1016/j.acra.2009.10.016.
[60] Li F, Aoyama M, Shiraishi J, Abe H, Li G, Suzuki K, et al. Radiologists' performance for differentiating benign from malignant lung nodules on high-resolution CT using computer-estimated likelihood of malignancy. Am J Roentgenol 2004;183:1209–15.
[61] Okumura M, Ota T, Kainuma K, Sayre JW, McNitt-Gray M, Katada K. Effect of edge-preserving adaptive image filter on low-contrast detectability in CT systems: application of ROC analysis. Int J Biomed Imaging 2008;2008:379486. doi:10.1155/2008/379486.
[62] Bushberg JT, Seibert JA, Leidholdt EM, Boone JM. The essential physics of medical imaging. Philadelphia (PA): Wolters Kluwer Health; 2011.
[63] Wunderlich A, Abbey CK. Utility as a rationale for choosing observer performance assessment paradigms for detection tasks in medical imaging. Med Phys 2013;40:111903. doi:10.1118/1.4823755.
[64] Chakraborty DP. A brief history of FROC paradigm data analysis. Acad Radiol 2013;20:915–19. doi:10.1016/j.acra.2013.03.001.
[65] Popescu LM. Nonparametric signal detectability evaluation using an exponential transformation of the FROC curve. Med Phys 2011;38:5690–702. doi:10.1118/1.3633938.
[66] Burgess A. Comparison of receiver operating characteristic and forced-choice observer performance-measurement methods. Med Phys 1995;22:643–55. doi:10.1118/1.597576.
[67] Macmillan NA, Creelman CD. Detection theory: a user's guide. Mahwah (NJ): Lawrence Erlbaum Associates; 2005.
[68] Bochud FO, Abbey CK, Eckstein MP. Visual signal detection in structured backgrounds. III. Calculation of figures of merit for model observers in statistically nonstationary backgrounds. J Opt Soc Am A Opt Image Sci Vis 2000;17:193–205. doi:10.1364/JOSAA.17.000193.

[69] Swets JA. Signal detection and recognition by human observers: contemporary readings. New York: Wiley; 1964.
[70] Craven BJ. A table of d′ for M-alternative odd-man-out forced-choice procedures. Percept Psychophys 1992;51:379–85.
[71] Hacker MJ, Ratcliff R. A revised table of d′ for M-alternative forced choice. Percept Psychophys 1979;26:168–70. doi:10.3758/BF03208311.
[72] Dahlquist G, Björck Å. Numerical methods. New York: Courier Corporation; 2012.
[73] Green DM, Swets JA. Signal detection theory and psychophysics. New York: John Wiley and Sons; 1966.
[74] Burgess AE. Visual perception studies and observer models in medical imaging. Semin Nucl Med 2011;41:419–36. doi:10.1053/j.semnuclmed.2011.06.005.
[75] Tapiovaara M. Efficiency of low-contrast detail detectability in fluoroscopic imaging. Med Phys 1997;24:655–64. doi:10.1118/1.598076.
[76] Eckstein MP, Abbey CK, Bochud FO. Visual signal detection in structured backgrounds. IV. Figures of merit for model performance in multiple-alternative forced-choice detection tasks with correlated responses. J Opt Soc Am A Opt Image Sci Vis 2000;17:206–17. doi:10.1364/JOSAA.17.000206.
[77] Zhang Y, Leng S, Yu L, Carter RE, McCollough CH. Correlation between human and model observer performance for discrimination task in CT. Phys Med Biol 2014;59:3389–404. doi:10.1088/0031-9155/59/13/3389.
[78] Hernandez-Giron I, Geleijns J, Calzado A, Veldkamp WJH. Automated assessment of low contrast sensitivity for CT systems using a model observer. Med Phys 2011;38:S25–35. doi:10.1118/1.3577757.
[79] Yu L, Leng S, Chen L, Kofler JM, Carter RE, McCollough CH. Prediction of human observer performance in a 2-alternative forced choice low-contrast detection task using channelized Hotelling observer: impact of radiation dose and reconstruction algorithms. Med Phys 2013;40:041908. doi:10.1118/1.4794498.
[80] Monnin P, Marshall NW, Bosmans H, Bochud FO, Verdun FR. Image quality assessment in digital mammography: part II. NPWE as a validated alternative for contrast detail analysis. Phys Med Biol 2011;56:4221–38. doi:10.1088/0031-9155/56/14/003.
[81] Borasi G, Samei E, Bertolini M, Nitrosi A, Tassoni D. Contrast-detail analysis of three flat panel detectors for digital radiography. Med Phys 2006;33:3580. doi:10.1118/1.2337636.
[82] Rivetti S, Lanconelli N, Bertolini M, Acchiappati D. A new clinical unit for digital radiography based on a thick amorphous selenium plate: physical and psychophysical characterization. Med Phys 2011;38:4480–8. doi:10.1118/1.3605471.
[83] Wagner RF, Brown DG, Pastel MS. Application of information theory to the assessment of computed tomography. Med Phys 1979;6:83–94.
[84] Myers KJ, Rolland JP, Barrett HH, Wagner RF. Aperture optimization for emission imaging: effect of a spatially varying background. J Opt Soc Am A 1990;7:1279–93.
[85] Burgess AE, Ghandeharian H. Visual signal detection. II. Signal-location identification. J Opt Soc Am A 1984;1:906–10. doi:10.1364/JOSAA.1.000906.
[86] Burgess A. Visual signal detection. III. On Bayesian use of prior knowledge and cross correlation. J Opt Soc Am A 1985;2:1498–507.
[87] Wagner RF, Myers KJ, Tapiovaara MJ, Brown DG, Burgess AE. Schneider RH, editor. Maximum a-posteriori detection and figures of merit for detection under uncertainty. 1990. p. 195–204 doi:10.1117/12.18797.
[88] Chesters MS. Human visual perception and ROC methodology in medical imaging. Phys Med Biol 1992;37:1433. doi:10.1088/0031-9155/37/7/001.
[89] Brown DG, Insana MF, Tapiovaara M. Detection performance of the ideal decision function and its McLaurin expansion: signal position unknown. J Acoust Soc Am 1995;97:379–98.
[90] Bochud FO, Abbey CK, Bartroff J, Vodopich D, Eckstein MP. Effect of the number of locations in MAFC experiments performed with mammograms 1999.
[91] Burgess AE. Evaluation of detection model performance in power-law noise, vol. 4324. 2001. p. 123–32 doi:10.1117/12.431180.
[92] Kotre CJ. The effect of background structure on the detection of low contrast objects in mammography. Br J Radiol 1998;71:1162–7. doi:10.1259/bjr.71.851.10434911.
[93] Bochud FO, Valley JF, Verdun FR, Hessler C, Schnyder P. Estimation of the noisy component of anatomical backgrounds. Med Phys 1999;26:1365–70.
[94] Marshall NW, Kotre CJ, Robson KJ, Lecomber AR. Receptor dose in digital fluorography: a comparison between theory and practice. Phys Med Biol 2001;46:1283–96.
[95] Tseng H-W, Fan J, Kupinski MA, Sainath P, Hsieh J. Assessing image quality and dose reduction of a new x-ray computed tomography iterative reconstruction algorithm using model observers. Med Phys 2014;41:071910. doi:10.1118/1.4881143.
[96] Samei E, Richard S. Assessment of the dose reduction potential of a model-based iterative reconstruction algorithm using a task-based performance metrology. Med Phys 2015;42:314–23. doi:10.1118/1.4903899.
[97] Chen B, Ramirez Giraldo JC, Solomon J, Samei E. Evaluating iterative reconstruction performance in computed tomography. Med Phys 2014;41:121913. doi:10.1118/1.4901670.
[98] Tapiovaara MJ, Wagner RF. SNR and noise measurements for medical imaging: I. A practical approach based on statistical decision theory. Phys Med Biol 1993;38:71. doi:10.1088/0031-9155/38/1/006.
[99] Rolland JP, Barrett HH. Effect of random background inhomogeneity on observer detection performance. J Opt Soc Am A 1992;9:649–58. doi:10.1364/JOSAA.9.000649.
[100] Tapiovaara MJ. SNR and noise measurements for medical imaging. II. Application to fluoroscopic X-ray equipment. Phys Med Biol 1993;38:1761. doi:10.1088/0031-9155/38/12/006.
[101] Boedeker KL, Cooper VN, McNitt-Gray MF. Application of the noise power spectrum in modern diagnostic MDCT: part I. Measurement of noise power spectra and noise equivalent quanta. Phys Med Biol 2007;52:4027–46. doi:10.1088/0031-9155/52/14/002.
[102] Burgess AE. Statistically defined backgrounds: performance of a modified nonprewhitening observer model. J Opt Soc Am A Opt Image Sci Vis 1994;11:1237–42.
[103] Loo LN, Doi K, Metz CE. A comparison of physical image quality indices and observer performance in the radiographic detection of nylon beads. Phys Med Biol 1984;29:837–56.
[104] Zhang Y, Pham BT, Eckstein MP. Evaluation of internal noise methods for Hotelling observer models. Med Phys 2007;34:3312–22. doi:10.1118/1.2756603.
[105] Brankov JG. Evaluation of channelized Hotelling observer with internal-noise model in a train-test paradigm for cardiac SPECT defect detection. Phys Med Biol 2013;58:7159–82. doi:10.1088/0031-9155/58/20/7159.
[106] Leng S, Yu L, Zhang Y, Carter R, Toledano AY, McCollough CH. Correlation between model observer and human observer performance in CT imaging when lesion location is uncertain. Med Phys 2013;40:081908. doi:10.1118/1.4812430.
[107] Brankov JG. Optimization of the internal noise models for channelized Hotelling observer. 2011 IEEE Int Symp Biomed Imaging Nano Macro 2011;1788–91. doi:10.1109/ISBI.2011.5872753.
[108] Hernandez-Giron I, Calzado A, Geleijns J, Joemai RMS, Veldkamp WJH. Comparison between human and model observer performance in low-contrast detection tasks in CT images: application to images reconstructed with filtered back projection and iterative algorithms. Br J Radiol 2014;87:20140014. doi:10.1259/bjr.20140014.
[109] Myers KJ, Barrett HH. Addition of a channel mechanism to the ideal-observer model. J Opt Soc Am A 1987;4:2447–57. doi:10.1364/JOSAA.4.002447.
[110] Abbey CK, Barrett HH. Human- and model-observer performance in ramp-spectrum noise: effects of regularization and object variability. J Opt Soc Am A Opt Image Sci Vis 2001;18:473–88.
[111] Barrett HH, Yao J, Rolland JP, Myers KJ. Model observers for assessment of image quality. Proc Natl Acad Sci U S A 1993;90:9758–65.
[112] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.
[113] Bland JM, Altman DG. Agreement between methods of measurement with multiple observations per individual. J Biopharm Stat 2007;17:571–82. doi:10.1080/10543400701329422.