IOP Publishing Phys. Med. Biol. 65 (2020) 085017 https://doi.org/10.1088/1361-6560/ab7b2e

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

RECEIVED
2 October 2019

REVISED
13 February 2020

ACCEPTED FOR PUBLICATION
28 February 2020

PUBLISHED
23 April 2020

PAPER

RDI—a regression detectability index for quality assurance in: x-ray imaging

M Anton[1], **W J H Veldkamp**[2], **I Hernandez-Giron**[2] **and C Elster**[1]

[1] Physikalisch-Technische Bundesanstalt Braunschweig and Berlin, Berlin, Germany
[2] Radiology Department, Leiden University Medical Center, Leiden, Netherlands

E-mail: mathias.anton@ptb.de

Keywords: x-ray CT, quality assurance, task-specific quality assessment, model observer, low contrast detectability

Abstract

Novel iterative image reconstruction methods can help reduce the required radiation dose in x-ray diagnostics such as computed tomography (CT), while maintaining sufficient image quality. Since some of the established image quality measures are not appropriate for reliably judging the quality of images derived by iterative methods, alternative approaches such as task-specific quality assessment would be highly desirable for acceptance or constancy testing. Task-based image quality methods are also closer to tasks performed by the radiologists, such as lesion detection. However, this approach is usually hampered by a huge workload, since hundreds of images are usually required for its application. It is demonstrated that the proposed approach works reliably on the basis of significantly fewer images, and that it correlates well with results obtained from human observers.

1. Introduction

Diagnostic examinations making use of x-rays are among the most frequently used in Radiology. Especially for x-ray computed tomography(CT), the number of examinations is steadily increasing. Although the fraction of CT examinations is small compared to conventional x-ray examinations or dental ones, its contribution to the overall population dose is significant with an increasing tendency (see, for example, European Commission 2015).

Suppliers have developed improvements of the CT procedure, such as automatic exposure control and novel image reconstruction methods, with the aim of reducing the radiation dose while maintaining a sufficiently high quality of the diagnostic images. Especially in the case of so-called iterative image reconstruction methods, suppliers claim to be able to reduce the required dose significantly (Vaishnav *et al* 2014, Viry *et al* 2018, Willemink and Noël 2019). Iterative reconstruction methods violate the assumptions of system linearity and shift-invariance, required to apply some of the methods traditionally accepted to quantify image quality (Illers *et al* 2005, ICRU 2006, Samei *et al* 2019).

In order to overcome this problem, task-specific image quality assessment methods, including model observers, have been investigated and applied for quite some time (Barrett and Myers 2003, ICRU 2006, He and Park 2013, Barrett *et al* 2015, Verdun *et al* 2015). These methods have picked a simplified diagnostic task, for example, classifying an image as belonging either to class 0 (no lesion visible) or to class 1 (lesion visible), and have used the probability of correct classification as a quantitative measure of image quality. Uncertainties associated with the results of the obtained quality assessment have typically been characterised by methods from classical statistics (see, for example, Wunderlich *et al* 2015), but also Bayesian methods have been proposed lately (Reginatto *et al* 2017, Khanin *et al* 2018). A model observer is a mathematical object which attributes a number (a test statistic) to an image and performs a classification according to that test statistic and thus mimicks the classification by a radiologist. In recent years, two model observers have gained more attention than others: the channelized Hotelling observer (CHO) due to its optimality (Myers and Barrett 1987, Abbey and Barrett 2001, Gifford *et al* 2000, Barrett and Myers 2003, Wunderlich *et al* 2015, Racine *et al* 2018, Ba *et al* 2018), and the non-prewhitening matched filter with eye filtering (NPWE) due to its good correlation with human observers which have been successfully applied to several imaging modalities (Burgess 1994, Barten 1999, Abbey and Barrett 2001, Bouwman *et al* 2017, Balta *et al* 2018). However, both types of model observer require quite a large number of image pairs (of the order of several hundreds) for a reliable image quality assessment.

For the NPWE, the number of required images can be reduced by using a template which is constructed based on prior knowledge of the image instead of using training data (see, for example, Hernandez-Giron *et al* 2014). For images of technical phantoms, like the low-contrast module of the Catphan (The Phantom Laboratory, Salem, NY, USA), the size and the shape of the target to be detected are precisely known, so a model of the image, the template, can be easily generated. Recently, the parametric modified simple observer for pooled images (MSOpi) has been proposed (Anton *et al* 2018), which is applicable to images from simple technical phantoms. The MSOpi compares the mean signal with the mean background, assuming a common white noise model. While the MSOpi observer has been shown to be very sensitive and able to reduce efforts in mammography quality assurance (Kretz *et al* 2019), it does not appear to correlate sufficiently well with results obtained by human observers in CT. The reason for this probably lies in its simplified white noise model, which neither reflects realistic noise characteristics in images, nor accounts for human perception. These findings prompted a further, and more universal, development, which accounts for realistic noise modelling and more general settings of noise and background models with the aim that results correlate well with results from human observers.

The basic idea of the new observer is to approximate the images as well as their covariance matrix by few-parameter models and to derive a detectability index from the fit parameters, which is why it is called the regression detectability index (RDI). The approach reduces the number of unknowns to be determined from the image data significantly, compared to established observers like the CHO and the NPWE. Therefore, only approximately an order of magnitude fewer images are required for the new observer than for established ones. The signal model is built by incorporating some blurring, which better accounts for mispositioning of the phantom and the limited resolution of human perception. Likewise, modelling covariances appears to yield a detectability index, which is closer to human perception than simple white noise models.

This publication is organized as follows: section 2 gives a short introduction to task-specific quality assessment in general, we briefly review the CHO and the NPWE, which are used for comparisons with our newly developed observer.

Our proposed method (RDI, regression detectability index) as a new image quality measure will be described in section 3: first, the scientific background related to object detectability in phantoms, second, the outline of the method to construct a model of the images, followed by the estimation method of the covariance matrix, and ending with practical calculation issues and the estimation of the uncertainty of the RDI. Section 4 shows the application of the RDI to CT images of an in-house phantom for a range of dose levels and two reconstruction methods to depict its general performance. In section 5, RDI results are compared to human scorings and model observer results from a previously published study with a different phantom, for benchmarking. Section 6 closes the paper with a discussion of the results and a brief outlook. A detailed description of the CT imaging data is shifted to the appendix, as well as details of the covariance matrix estimation procedure.

2. Task-specific quality assessment

Task-specific quality assessment quantifies the image quality by emulating a situation encountered by a radiologist. A typical simple example is the detection of a lesion in an x-ray image. In order to make this task accessible to a mathematical model observer, it is usually greatly simplified. In the following, we treat the so-called signal-known-exactly (SKE) / background-known-statistically (BKS) paradigm.

The task of detecting a lesion can be formulated as a binary classification. Assume an image is stored as a column vector $\mathbf{y}$ of length n, where n is the number of pixels. A test statistic z can then be obtained by multiplying $\mathbf{y}$ by the transpose of a vector $\mathbf{x}$ of the same length, which may be a template or a signal obtained from a set of training images:

$$z = \mathbf{x}^T \cdot \mathbf{y}. \tag{1}$$

The image $\mathbf{y}$ is classified as belonging to class 1 if z exceeds a chosen threshold value γ, and as belonging to class 0 otherwise:

$$\begin{aligned} z \geq \gamma \quad &\rightarrow \quad \mathbf{y} \in \text{class 1: with the signature of a lesion} \\ z < \gamma \quad &\rightarrow \quad \mathbf{y} \in \text{class 0: with no signature of a lesion} \end{aligned}$$

When the threshold γ is varied, the fraction of correctly classified images changes accordingly. A plot of the correctly 'positive' (i.e. with the signature of a lesion) classified fraction of images (TPF, true positive fraction) as a function of the incorrectly 'positive' classified fraction (FPF, false positive fraction) is called the ROC—or receiver operating characteristic—curve. For a perfect classifier, the ROC curve immediately jumps to $\mathrm{TPF} = 1$ at $\mathrm{FPF} = 0$. When the classification is equivalent to pure guessing, the ROC curve is a straight line connecting the origin with the point (1,1) in the plot. The area under the ROC curve (AUC) is a useful figure of merit for the quantification of image quality, and it is equivalent to the probability that a randomly drawn image from class 1 will result in a higher value of the test statistic than a randomly drawn image from class 0 (see, for example, Barrett and Myers 2003, Pepe 2003, ICRU 2006). If the test statistics are normally distributed, the AUC is connected to the signal-to-noise ratio SNR through

$$\mathrm{AUC} = \Phi\left(\frac{\mathrm{SNR}}{\sqrt{2}}\right), \tag{2}$$

where Φ is the cumulative distribution function of the standard normal distribution and the SNR is given by

$$\mathrm{SNR}^2 = \frac{\left(\bar{z}_1 - \bar{z}_0\right)^2}{0.5 \cdot \left(\sigma^2(z_1) + \sigma^2(z_0)\right)}. \tag{3}$$

In equation (3), $\bar{z}_i$ is the expectation of z-values obtained from images $\mathbf{y}$ belonging to class i, and the $\sigma^2(z_i)$ are the variances of the two distributions (see, for example, Barrett *et al* 1998, Barrett *et al* 2015). The SNR represents the distance of the centres of the two distributions relative to their mean width. In the following, we will mainly use the SNR or the detectability index d' (which are synonyms) as a quantifier for the image quality. The reason for this is that d'-values may well be distinguishable while the corresponding AUC values might all be close to unity and hence close to each other. Furthermore, the relations of d' to other parameters such as the target size can be simpler.

2.1. NPW and NPWE

One of the simplest observers is the so-called non-prewhitening matched filter (NPW). If a set of images without a lesion and a set of images with a lesion are available, its test statistic is constructed as

$$z_{\mathrm{NPW}} = \left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0\right)^T \cdot \mathbf{y}, \tag{4}$$

where $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_0$ denote the means of all images with and without a lesion, respectively, and $\mathbf{y}$ is a single image that shall be classified. However, better correlation with the performance of a human observer is achieved when the images are transformed by a so-called eye filter, which reflects the contrast sensitivity of the human eye. The eye filter was introduced by Burgess (1994), and an elaborated model of the contrast sensitivity of the human eye was developed by Barten (1992) (see also Barten 1999). For the test statistic, the image vectors in (4) have to be replaced by their filtered versions. In this work, the eye filter implemented by Hernandez-Giron *et al* (2014) was used which is based on the one proposed by Burgess (1999):

$$E(f_\alpha) = f_\alpha \cdot \exp\left(-\frac{f_\alpha}{r_\alpha}\right); \tag{5}$$

E quantifies the contrast sensitivity of the eye, the maximum of which is reached at $f_\alpha = r_\alpha$, where f_α is the spatial frequency in deg^{-1}. Usually, $r_\alpha = 4\ \mathrm{deg}^{-1}$ is assumed.

For the practical calculations, the image data set is either split in training and testing groups, or a template accompanying the image data is used. In the first case, $\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0$ is obtained from the training images, and test statistics z_i are obtained from two groups of test images $\mathbf{y}_0$ and $\mathbf{y}_1$. In the second case, a template $\mathbf{x}_T$ is used instead of $\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0$, as described by Hernandez-Giron *et al* (2014) and all available images are used for testing. If z_i denotes the test statistic according to equation (4) for test images $\mathbf{y}_i$ of class i, the detectability index d' is obtained by

$$d' = \frac{\left|\overline{z_1} - \overline{z_0}\right|}{\sqrt{0.5 \cdot \sigma_1^2 + 0.5 \cdot \sigma_0^2}}, \tag{6}$$

where the overline denotes averaging and σ_i is the standard deviation of z_i (compare equation (3)). It is connected to the AUC via $d' = \sqrt{2} \cdot \Phi^{-1}(\mathrm{AUC})$, where Φ^{-1} is the inverse of the standard cumulative normal distribution function.

In order to establish the relation of the detectability of the average human observer d'_{human} to the detectability of the NPWE d'_{NPWE}, Hernandez-Giron *et al* (2014) have defined an efficiency η as the slope of an assumed linear relationship between the squares of the d'-values:

$$(d'_{\text{human}})^2 = \eta \cdot (d'_{\text{NPWE}})^2. \tag{7}$$

It was shown that a value of $\eta = 0.44$ resulted from the data used in their publication.

The uncertainty of d'_{NPWE} in this work has been evaluated by applying a bootstrap technique (Efron 1982) as follows. Subsets were drawn repeatedly from the set of all images and used to estimate d'_{NPWE} through the application of equation (6). From the resulting distribution of values of d'_{NPWE}, an interval was selected that covered 95% of these results. In our case, we chose a size of 40 images for the subsets. An approximate 95% confidence interval associated with the detectability of the NPWE when applied to all images was then constructed by rescaling the obtained interval with the factor $\sqrt{40/N}$, where N denotes the number of all images.

2.2. CHO

The optimum linear classifier is the Hotelling observer (HO) that can be formulated as

$$z_{\text{HO}} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)^T \cdot \mathbf{V}^{-1}\, \mathbf{y}, \tag{8}$$

where $\mathbf{V}$ is the covariance matrix of the images. The effect of multiplying an image $\mathbf{y}$ by $\mathbf{V}^{-1}$ is that the noise in the image gets decorrelated. It is assumed that the covariances for the two classes of images do not differ (small signal approximation). In that case, an estimate $\hat{\mathbf{V}}$ of the covariance matrix is obtained as

$$\hat{\mathbf{V}} = \frac{1}{N_0 + N_1 - 2} \cdot \left(\sum_{k=1}^{N_0} (\bar{\mathbf{x}}_0 - \mathbf{x}_0(k)) \cdot (\bar{\mathbf{x}}_0 - \mathbf{x}_0(k))^T + \sum_{l=1}^{N_1} (\bar{\mathbf{x}}_1 - \mathbf{x}_1(l)) \cdot (\bar{\mathbf{x}}_1 - \mathbf{x}_1(l))^T \right), \tag{9}$$

where N_i are the number of images $\mathbf{x}_i$ in each class, and $\bar{\mathbf{x}}_i$ denotes the average over all N_i images within class i. However, even for a very moderately sized square image with an edge length of 64 pixels (px), $\mathbf{V}$ has the size 4096×4096. Therefore, a channelized version is usually implemented. The images are projected to a lower dimension p by using a channel matrix $\mathbf{U}$ of size $n \times p$:

$$\tilde{\mathbf{x}}_i = \mathbf{U}^T \mathbf{x}_i, \tag{10}$$

$$\tilde{\mathbf{y}} = \mathbf{U}^T \mathbf{y}, \tag{11}$$

$$\tilde{\mathbf{V}} = \mathbf{U}^T \mathbf{V}\, \mathbf{U} \tag{12}$$

$$z_{\text{CHO}} = (\bar{\tilde{\mathbf{x}}}_1 - \bar{\tilde{\mathbf{x}}}_0)^T \cdot \tilde{\mathbf{V}}^{-1}\, \tilde{\mathbf{y}} \tag{13}$$

The ˜ (tilde) denotes the channelized versions of the image vectors and covariance matrix. Usually, $\tilde{\mathbf{V}}$ is not calculated as given in (12), but is rather calculated from the channelized images $\tilde{\mathbf{x}}_i$ directly. The detectability and its 95% coverage interval were calculated using the routine exactCI_CHO from a software toolbox published by Wunderlich (2015). We have used $p = 40$ Gabor channels as described by Wunderlich *et al* (2015); the size of the covariance matrix is therefore reduced to a size of 40×40 which is tractable.

3. RDI—a regression detectability index

Common phantoms used for quality assurance in CT share two important features: the targets to be detected by a human or a mathematical model observer exhibit a circular cross section, and the background material is homogeneous. Two examples are the low-contrast module CTP515 of the Catphan phantom or the MITA body phantom CCT189 (both by The Phantom Laboratory, Salem, NY, USA). This prior knowledge can be exploited for the construction of a detectability index which assesses image quality on the basis of significantly fewer images than the application of conventional model observers requires.

We propose a new detectability index called the RDI, which exploits the known image structure by fitting a regression model to the image. Apart from the fact that this approach requires fewer images, RDI does not need two classes of images. An estimate of the detectability index is obtained using only images with the signature of the target, and no separate background images are needed.

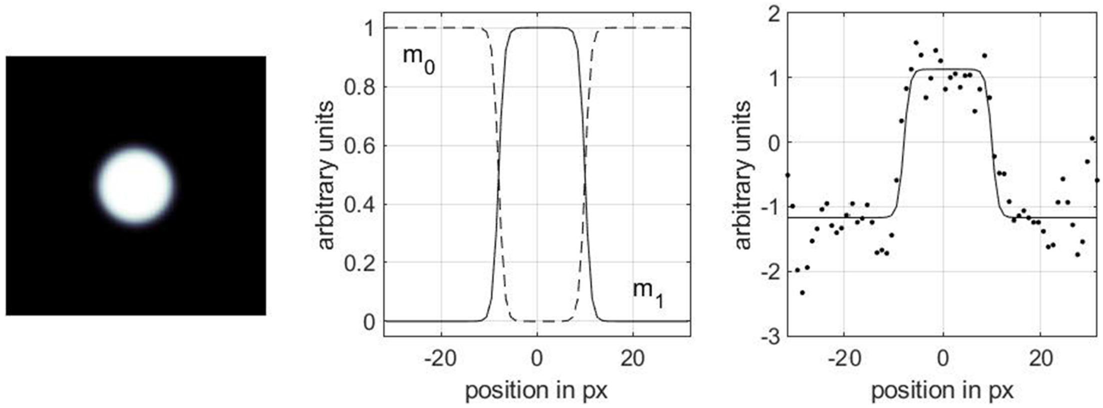

Figure 1. A model of the mean image. Left: a pseudocolour plot of the signal model $\mathbf{m}_1$, the bright zone indicates the region where the signal of the target is expected, the dark region is where only background is found. Middle: cross section of the background model $\mathbf{m}_0$ and $\mathbf{m}_1$ for $r_0 = 9$ px and a square image of 64 px edge length. Right: cross section of a fit of the model to an average of 100 images from data set EXP1 (see appendix A.1), FBP reconstruction at 80 mAs. The fit parameters in this case are $\theta_0 = -1.17$ and $\theta_1 = 1.12$.

Previous investigations (see, for example, references in the review by Barrett *et al* (2015), section 3.3.6, but also Verdun *et al* (2015) or Bouwman *et al* (2017)) have shown that the correlation of established observers—such as the CHO or the NPWE—with the performance of human observers can be related to the fact that both implement properties of the human visual system. The CHO uses channels—such as Gabor channels—to replicate the human cortex visual perception: the information of the visual stimulus is decomposed into 'channels', that when certain thresholds are reached, triggers detection. In addition, it is assumed that the human visual system also has the capability to—at least partly—decorrelate the noise structure of the images. The NPWE, on the other hand, includes a model of the contrast sensitivity of human vision. The detectability index proposed in this paper implements some of these aspects: decorrelation is achieved by employing a simplified model of the covariance matrix. In this way, the incomplete decorrelation by the human visual system is emulated. The limited contrast sensitivity is taken into account by softening edges in the regression model of the image.

In the following, we will describe the novel detectability index, including the design of the regression function emulating the limited contrast sensitivity of human observers, the estimation of a simplified model for the covariance matrix, the evaluation of uncertainty, as well as practical issues in the implementation of the approach.

3.1. Regression detectability index

The x-ray image is assumed to consist of a signal and a background part, which is illustrated on the left of figure 1. While the signal and background are separated in the technical phantom, we model the image by a smooth transition between the two regions to account for the limited contrast sensitivity of human vision. The transition chosen is described below. The image is then modelled as a superposition of these two parts. Specifically, the vector of image pixels $\mathbf{y} = (y_1, \ldots, y_n)^T$ is modelled as

$$\mathbf{y}|\theta, \mathbf{V} \sim \mathcal{N}(\mathbf{M}\theta, \mathbf{V}) \,, \tag{14}$$

where n denotes the number of pixels, $\mathcal{N}(\mu, \mathbf{V})$ stands for a multivariate Gaussian distribution with mean vector μ and covariance matrix $\mathbf{V}$, $\theta = (\theta_0, \theta_1)^T$ includes the two regression parameters involved, and the $n \times 2$ matrix $\mathbf{M} = (\mathbf{m}_0^T, \mathbf{m}_1^T)$ is the design matrix. The first column vector $\mathbf{m}_0^T$ of $\mathbf{M}$ relates to the background part, and the second column vector $\mathbf{m}_1^T$ to the signal part. Figure 1 illustrates the choice of these two functions in comparison with the known structure of the technical phantom. Note that within a transition zone, both regression functions are non-zero. The $n \times n$ matrix $\mathbf{V}$ is the covariance matrix. $\mathbf{V}$ is non-diagonal in general, which reflects the presence of correlations.

The new idea implemented as the RDI is to assess image quality in terms of how well the contrast between the two regions is distinguished within model (14). To this end, the detectability index

$$d' = \sqrt{\frac{(\theta_1 - \theta_0)^2}{\mathrm{var}(\theta_0) + \mathrm{var}(\theta_1) - 2 \cdot \mathrm{cov}(\theta_0, \theta_1)}} \tag{15}$$

is taken. The detectability index is of a similar form to the SNR in (3). The nominator is large when the mean signal and mean background are well separated. The denominator equals the variance of $(\theta_0 - \theta_1)$ and models the squared uncertainty associated with the difference $(\theta_0 - \theta_1)$. Large detectability indices are favoured by large mean differences between the signal and the background and by small variances of this difference.

In addition to the regression parameters θ, also the covariance matrix $\mathbf{V}$ in model (14) is unknown. We estimate $\mathbf{V}$ through a simple parametric model in advance and ignore the uncertainty associated with that estimate. Next we will describe the selection of the design matrix $\mathbf{M}$, followed by the estimation of the covariance matrix $\mathbf{V}$. Then we describe the estimation of d' in (14), including the assignment of an associated uncertainty. Finally, we give some practical guidance for carrying out the calculations.

3.2. Modelling the mean image

The matrix $\mathbf{M}$ in (14) describes a model of the mean image. The most common phantoms used for quality assurance in CT imaging have simple targets of a circular cross section. Therefore, we restrict the model to a square region of interest (ROI) with a circular target of variable diameter in its centre, see figure 1. We will use a logistic function to model a smooth transition between the two regions. The reason for this is twofold: first, positioning will almost never be exact, i.e. boundaries that are too sharp may cause systematic deviations of fitting a non-smooth model to the image. Second, there are two processes which will cause the sharp edges of the target object to soften in the perceived image, the finite resolution of the imaging device, commonly modelled by its modulation transfer function (MTF), and the contrast sensitivity of the human eye. The latter is probably the stronger limiting factor. By modelling the mean image through a smooth transition between the signal and the noise region, it is expected to better resemble human perception.

More precisely, we model the mean image as a superposition of the columns of the $n \times 2$ matrix $\mathbf{M}$ as follows. The first column, denoted by m_0, is taken as

$$m_0(r_i) = (1 + \exp\left[-k(r_i - r_0)\right])^{-1}, i = 1, \ldots, n, \tag{16}$$

where r_i denotes the distance of the ith pixel from the centre of the image, and r_0 is the radius of the inner disc containing the signal area. All distances are given in units of pixels (px). For pixels i with values of $r_i \gg r_0$ in the background region, this function takes a value of approximately 1, while for $r_i \ll r_0$, its value approaches zero. The selection of k is described below. The second column of $\mathbf{M}$, m_1, is taken as

$$m_1(r_i) = 1 - (1 + \exp\left[-k(r_i - r_0)\right])^{-1}, i = 1, \ldots, n \tag{17}$$

and it describes the complement of (16). For pixels with values of $r_i \ll r_0$ in the signal region, this function takes a value of approximately 1, while for $r_i \gg r_0$, its value approaches zero.

For the selection of k in (16) and (17), the following procedure is proposed. Let $f_\alpha^{50\%}$ in $\deg^{-1}$ denote the spatial frequency, where the contrast sensitivity in (5) drops to half its maximum. The corresponding width on a display can be calculated as

$$w_{\mathrm{px}}^{50\%} = \frac{\pi}{180\deg} \cdot R \cdot S \cdot \frac{1}{f_\alpha^{50\%}}, \tag{18}$$

where R is the resolution of the display and S is the viewing distance. For our calculations we used $f_\alpha^{50\%} = 10.7\ \deg^{-1}$, $R = 4.3$ px$\cdot$ mm^{-1} and $S = 50$ cm, yielding $w_{\mathrm{px}}^{50\%} = 3.54$ px. The width $w_{\mathrm{px}}^{50\%} = 3.54$ px is identified with the width of the transition region in the models (16) and (17). For the latter, we define the width as the distance Δ_{95} between the points where each of the model functions m_i attain the values of 0.05 and 0.95, that is

$$\Delta_{95} = \frac{2}{k} \ln\left(\frac{0.95}{0.05}\right). \tag{19}$$

For our choice of $w_{\mathrm{px}}^{50\%} = 3.54$ px, setting $\Delta_{95} = w_{\mathrm{px}}^{50\%}$ resulted in the value $k = 1.66$ px^{-1}. The second parameter r_0 can be directly derived from the known scale of the CT image and the known sizes of the target, or from templates that are available with the data. We refer to figure 1 for an illustration of the chosen mean model and to the appendix for further details concerning our image data.

3.3. Estimation of the covariance matrix

The elements $V_{ij}, i, j = 1, \ldots, n$, of the covariance matrix $\mathbf{V}$ are modelled as a Gaussian kernel

$$V_{ij} = f(d_{ij}) = a_1 \cdot \exp\left(-a_2 \cdot d_{ij}^2\right) \tag{20}$$

with $a_1, a_2 > 0$, where

$$d_{ij}^2 = \|\mathbf{x}_i - \mathbf{x}_j\|_2^2\,. \tag{21}$$

In equation (21), $\mathbf{x}_i$ is the 2×1 vector of the spatial position of the ith pixel, and $\mathbf{x}_j$ that of the jth pixel. $\|\cdot\|_2$ stands for the Euclidean norm. Hence, (20) models the covariance as a simple function decaying with distance, allowing for a possible correlation even between remote pixels. Underlying assumptions are that the pixel covariances are shift invariant and rotationally symmetric, which can be viewed as a natural assumption in this context (see also appendix A.4).

The unknowns a_1 and a_2 in (20) are obtained by a least-squares fit to the observed covariances

$$\widehat{V}_{ij} = \frac{1}{N-1}\sum_{l=1}^{N}\left(\overline{y}_i - y_i^{(l)}\right)\left(\overline{y}_j - y_j^{(l)}\right) \tag{22}$$

from a set of N images $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(N)}$. Please recall that only one class of images containing the signal is used. In (22), $\overline{y}_i$ denotes the mean across images of the ith pixel. In what follows, $\mathbf{V}$ always denotes the fitted model matrix according to equation (20). After having obtained estimates for a_1 and a_2, it is checked whether the obtained matrix $\mathbf{V}$ is numerically positive-definite and has a sufficiently small condition number. If these conditions are not met, the matrix is modified as described in section 3.5.

3.4. Estimation of the detectability index

We propose estimating d' for a *single image* through

$$\widehat{d'}^{(1)} = \sqrt{\frac{\left(\widehat{\theta}_1 - \widehat{\theta}_0\right)^2}{(1,-1)\,\mathbf{C}\,(1,-1)^T}}\,, \tag{23}$$

where $\mathbf{C}$ is the covariance matrix of the estimate $\widehat{\theta}$ of the parameters θ

$$\mathrm{cov}(\widehat{\theta}) = \mathbf{C} = \left(\mathbf{M}^T\mathbf{V}^{-1}\mathbf{M}\right)^{-1} \tag{24}$$

and

$$\widehat{\theta} = \mathbf{C}\,\mathbf{M}^T\mathbf{V}^{-1}\mathbf{y}\,. \tag{25}$$

We refer to subsection 3.5 below for guidance as to how (24) and (25) are actually calculated.

Let N denote the number of images taken for estimating d', and $\widehat{d'}^{(l)}, l = 1, \ldots, N$, the corresponding estimates obtained for each single image. Following the Guide to the expression of uncertainty in measurement (GUM) (JCGM100 2008), we recommend taking the final estimate as the mean

$$\widehat{d'} = \frac{1}{N}\sum_{l=1}^{N}\widehat{d'}^{(l)}\,, \tag{26}$$

along with its associated squared standard uncertainty

$$u^2(d') = \frac{1}{N}\frac{1}{(N-1)}\sum_{l=1}^{N}\left(\widehat{d'} - \widehat{d'}^{(l)}\right)^2\,, \tag{27}$$

and the 95% coverage interval[1]

$$I_{0.95} = \left[\widehat{d'} - t_{(N-1),0.975}u(d'), \widehat{d'} + t_{(N-1),0.975}u(d')\right]\,, \tag{28}$$

where $t_{\nu,\alpha}$ denotes the α-quantile of a t-distribution with ν degrees of freedom.

3.5. Practical calculation issues

Fitting the model for the covariances (20) to the observed covariances (22) was done in a least-squares sense using a Levenberg-Marquardt algorithm (Gill *et al* 1981). In order to stabilize the fit, and to ease the choice of starting values, scaling of the observed covariances is recommended, for example, using

---

[1] Following the notation of the GUM (JCGM100 2008), the term coverage interval is used.

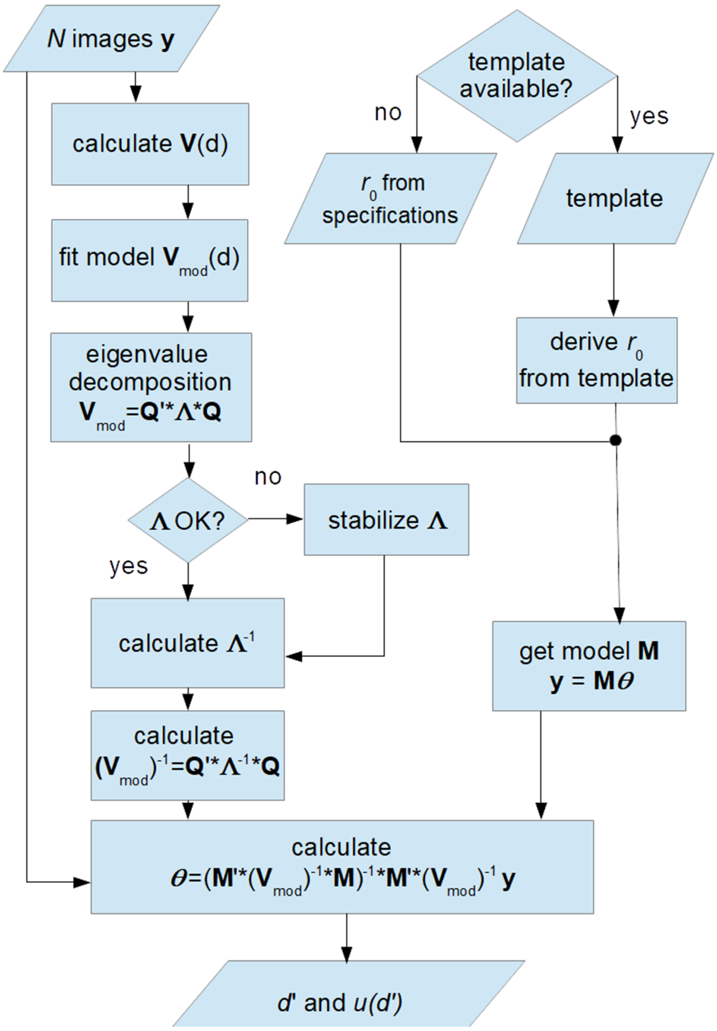

Figure 2. Flow chart summarising the necessary steps to obtain d' and its uncertainty for the RDI observer.

$$\lambda = \left(\frac{1}{n}\sum_{j=1}^{n}\widehat{V}_{jj}\right)^{-1}. \tag{29}$$

After having fitted the model covariance matrix to the scaled observed covariances, the obtained model covariance matrix is re-scaled by $1/\lambda$. Examples are shown in figures A6 and A7 in appendix D below.

It may happen that the modelled covariance matrix V in (20) is numerically close to singular, in which case it is modified. Therefore, first the decomposition

$$\mathbf{V} = \mathbf{Q}^T\Lambda\mathbf{Q} \tag{30}$$

is formed, where the $n \times n$ matrix $\mathbf{Q}$ is orthonormal and the $n \times n$ diagonal matrix Λ contains the eigenvalues of $\mathbf{V}$. If any of the eigenvalues of $\mathbf{V}$ are negative, or if the condition number of Λ is larger than the selected upper bound for the condition number of $\mathbf{V}$, Λ is modified by adding the identity matrix to it ε times , where ε is chosen such that all $\Lambda_{ii} > 0$ and the condition number of the modified Λ equals the chosen upper bound. In our calculations, we used 10^3 as an upper bound for the condition number. (The value of ε is usually small. ε relative to the maximum diagonal element of $\mathbf{V}$ is of the order of 0-0.1% for the FBP reconstructions and approximately 1% for the iterative reconstructions investigated below.) The inverse $\mathbf{V}^{-1}$ is given through

$$\mathbf{V}^{-1} = \mathbf{Q}^T\Lambda^{-1}\mathbf{Q}, \tag{31}$$

where Λ denotes the possibly modified matrix of eigenvalues of $\mathbf{V}$. Expression (31) can then be used to calculate the 2×2 matrix

$$\mathbf{M}^T\mathbf{V}^{-1}\mathbf{M} = \mathbf{M}^T\mathbf{Q}^T\Lambda^{-1}\mathbf{Q}\mathbf{M}, \tag{32}$$

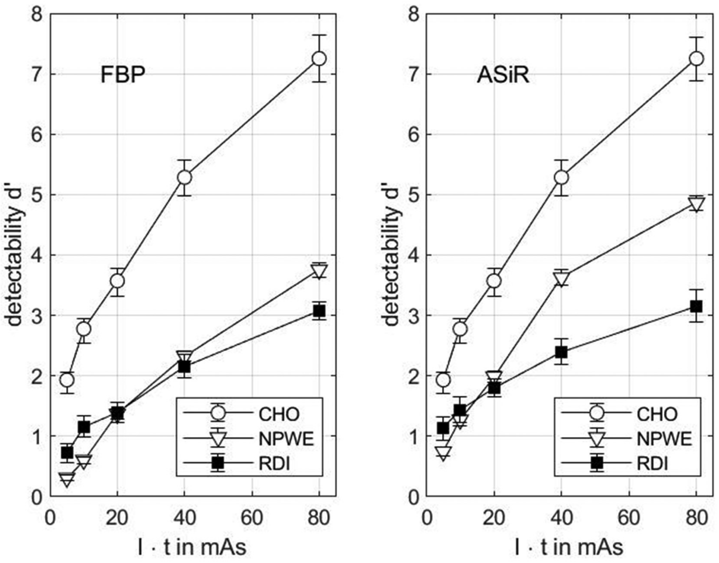

Figure 3. Detectability d' as a function of the time-current product; data set EXP1 (see appendix A.1). Open circles: CHO using 800 images; open triangles: NPWE using 400 images for training and 400 images for testing; filled squares: RDI using 40 images. Error bars represent the 95% coverage intervals, lines are shown only to guide the eye.

and subsequently its inverse (24), the covariance matrix of θ. The estimate $\widehat{\theta}$ in (26), is finally obtained by solving the linear system of equations for θ

$$\left(\mathbf{M}^T\mathbf{V}^{-1}\mathbf{M}\right)\theta = \mathbf{M}^T\mathbf{V}^{-1}\mathbf{y}. \tag{33}$$

The whole procedure is summarized in a flow chart in figure 2.

4. Application of the RDI—general performance

In order to explore the general performance of the RDI, we have applied it to two experimental CT image data sets that were obtained using the PTB's CT scanner, a General Electric Optima™ CT660 and a custom-built low-contrast phantom. In both cases, the time-current product is varied while other settings are held constant; image reconstructions using filtered back projection (FBP) and the iterative reconstruction ASiR of GE are compared. For the sake of the readability of this article, we have shifted detailed descriptions of the data sets to appendices A.1 and A.2.

The general behaviour of the RDI is shown in figure 3: CHO, NPWE and RDI have been applied to data set EXP1. The detectability d' is shown as a function of the time-current product for the standard filtered-back-projection (FBP) reconstruction (left panel) and the adaptive statistical iterative reconstruction (ASiR) of General Electric (GE, right panel; GE Molecular Imaging & Computed Tomography, USA). Error bars represent the 95% coverage interval. The results for the CHO with 40 Gabor channels are shown as open circles and were obtained as described in section 2.2. 400 images with and without a signal were supplied to the CHO. Gabor channels were chosen to obtain highest possible d'-values, in order to contrast the RDI with what could ideally be achieved[2]. The open triangles represent the results obtained using the NPWE as described in section 2.1. Here, 200 images with and without a signal each were used for training; the same number was used for testing. The RDI results are shown as filled squares. The parameters were set as explained in section 3, and only 40 images were used. Note that the size of the RDI error bars is comparable to that of the NPWE and the CHO, although only 5% of the images were necessary. The NPWE and the CHO both used 800 images in total.

The detectability obtained by the RDI appears to be rather lower than the one of the NPWE. This is not critical, since the investigations of Hernandez-Giron *et al* (2014) showed that the corresponding detectability of a human observer would approximately amount to $\sqrt{\eta}$ =66% of the NPWE's (see equation (7)), albeit for CT images of the Catphan phantom and a different CT manufacturer and iterative reconstruction algorithm. Both NPWE and RDI estimate a higher d' and hence a better image quality for the iterative reconstruction at the same time-current product, compared to the FBP reconstruction. For the RDI, this effect is smaller than

---

[2] DDOG channels yielded almost identical results.

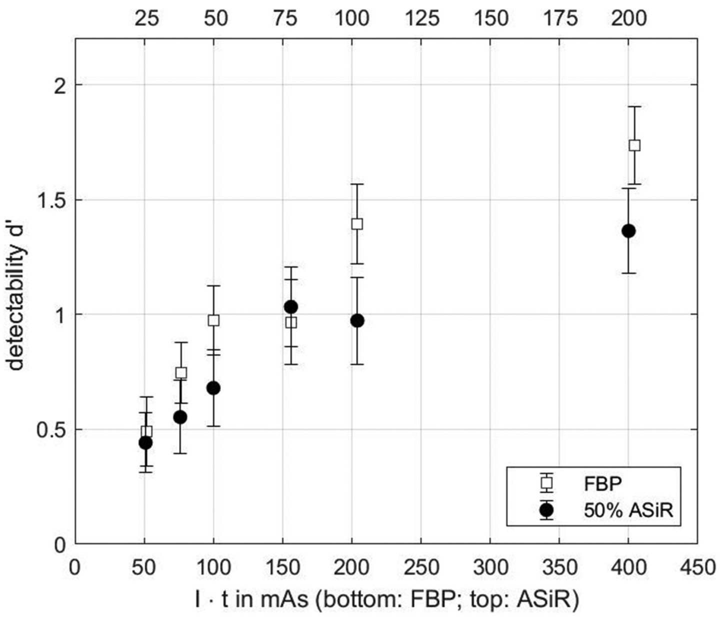

Figure 4. Detectability d' as a function of the time-current product in mAs obtained using the RDI with 40 images for each setting of data set EXP2 (see appendix B). The lower x-axis refers to the open symbols, which represent the results using FBP-reconstructed images, the upper x-axis refers to the filled symbols, which show the results from GE's iterative algorithm ASiR. Error bars represent the 95% coverage intervals.

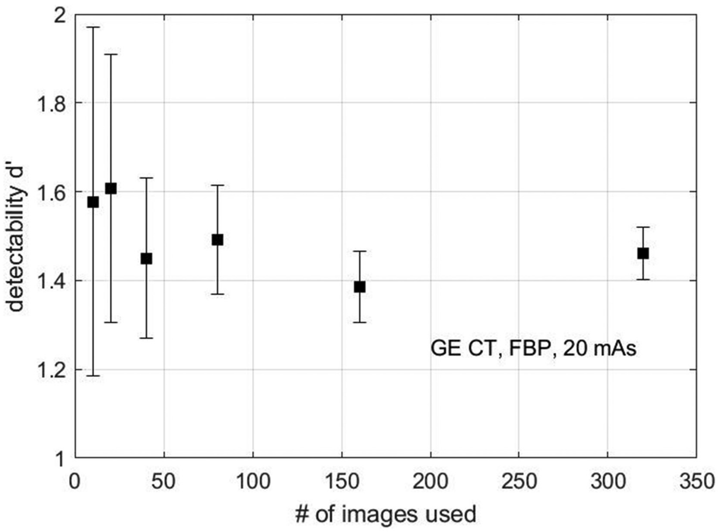

Figure 5. RDI detectability d' as a function of the number N of images used. Error bars indicate the 95% coverage interval. Data from set EXP1 (see appendix A.1), FBP at 20 mAs, target radius of approximately 9 px.

for the NPWE. The NPWE detectability is—on average—higher by 75% for the iterative reconstruction; for the RDI the increase is only 24%. The difference is smaller for the highest time-current products. The comparatively low performance of the RDI may appear surprising in view of its prewhitening property compared with the NPWE; note, however, that the detectability index defined by the RDI is different from that underlying the NPWE. The intersection of the NPWE and the RDI curves is due to the fact that for the NPWE a training/testing approach was used, where the quality of the template also degrades with decreasing time-current product, whereas the model template of the RDI is independent of $I \cdot t$.

For the CHO, no dependence of d' on the reconstruction method is observed, which may be due to its optimality: by accounting for all covariances, the noise is disentangled, independent of the image reconstruction method used. In what follows, we compare the new observer to the NPWE alone.

Figure 4 shows the results of the RDI for data set EXP2 (see appendix A.2). The RDI detectability d'—obtained from 40 images—is plotted as a function of the time-current product. Error bars indicate the 95% coverage intervals. Open squares represent the detectability obtained from FBP-reconstructed images whereas the filled circles show the detectability for the ASiR method. Note that there are two x-axes: the

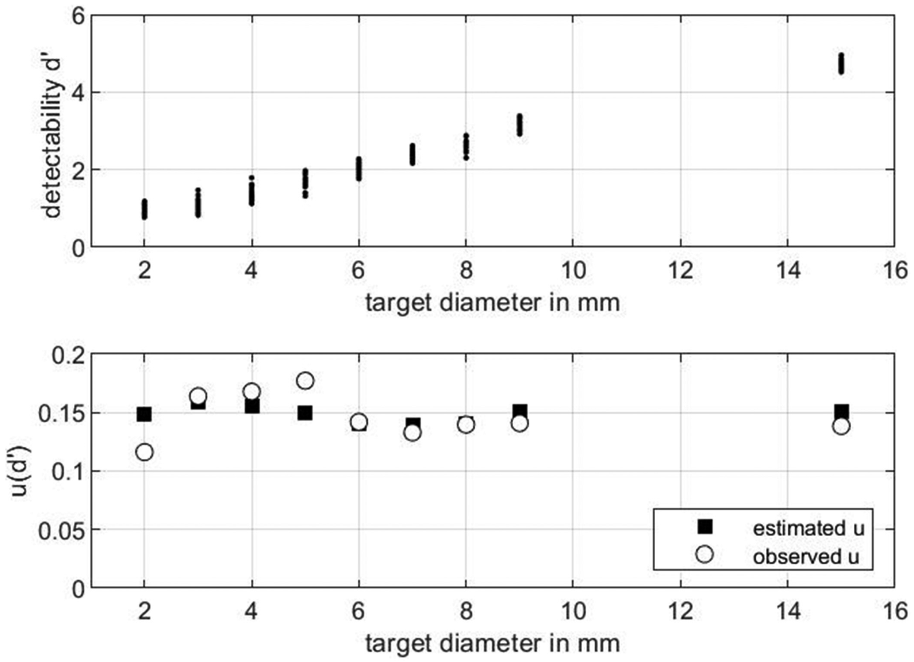

Figure 6. Top frame: RDI detectability d' as a function of the target diameter using experimental data (EXP3, see appendix C, 40 mAs with FBP reconstruction). The dots represent the 20 different d'-values obtained using 21 images every time. The bottom frame shows the standard uncertainty as a function of the target diameter. The open circles represent the standard deviation u obtained from the 20 repetitions, whereas the filled squares indicate the mean of the 20 uncertainty estimates obtained from the scatter of d' over 21 images each, calculated according to (27).

bottom axis refers to the FBP data, the top one to the ASiR results. The RDI detectability with the ASiR method is—on average—only (20 ± 15)% lower than the RDI detectability for FBP at twice the time-current product and hence twice the dose.

These results lead to the question: how far can we go?—i.e. how many images are really necessary? figure 5 demonstrates the influence of the number of images on the RDI results using a part of the experimental data set EXP1 (see appendix A.1). The detectability obtained from the RDI is displayed as a function of the number N of images used. The error bars represent 95% coverage intervals. Figure 5 illustrates that the expectation value of the estimate d' for the image quality does not appear to depend on N, as expected, only its uncertainty decreases with an increasing number of images used. The number of images has to be chosen according to the requirements on the uncertainty. When using $N = 40$ images, for example, a relative uncertainty of d' of less than 15% (half-width of the 95% coverage interval) is achieved. This number (of course) depends on the type of image under investigation.

Another open question is: how consistent are estimation of detectability and its uncertainty evaluation? In order to investigate this, we have used data set EXP3 (see appendix C), where images with different target sizes are available. For every target size, the evaluation of the detectability index according to equation (26) was repeated 20 times, each time using 21 randomly drawn images. In figure 6, d' is displayed as a function of the target diameter in mm (top panel). The black dots represent the 20 different d'-values per target diameter. The scatter of these 20 values is quantified by its standard deviation, which is shown in the lower panel as the open symbols. The filled symbols in the bottom panel represent the mean of the 20 uncertainty estimates calculated using equation (27) that were obtained from the scatter of the 21 images per repetition. As a result, the uncertainty estimate as evaluated from (27) appears to reliably predict the observed standard deviation of the mean.

5. Application of the RDI—comparison to human observer data

For a comparison of the RDI to human observer data and to a different NPWE evaluation, data that had been used for the publication by Hernandez-Giron *et al* (2014) were partly re-used. A description of the dataset that was obtained at the Leiden University Medical Center on a Toshiba Aquilion ONE™ CT with the low-contrast module of the Catphan Phantom (The Phantom Laboratory, Salem, NY, USA) is given in appendix C. There are nine different target sizes of the Catphan phantom with a nominal contrast of 1%. For every target size, images at five different time-current products are available (10 mAs to 250 mAs). For every setting (diameter and time-current product) two reconstructions were obtained, one with FBP and one with

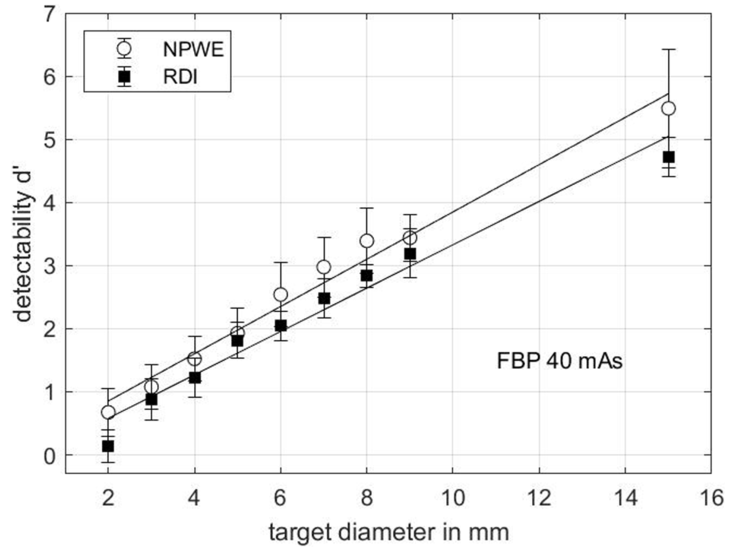

Figure 7. Detectability d' as a function of the target diameter in mm, data set EXP3 (see appendix A.3), FBP reconstruction at 40 mAs time-current product. NPWE results (shown as open symbols) were obtained using a template and 84 images with and without the target, RDI results (filled symbols) were obtained using only 21 images showing the signature of the target. Error bars represent the 95% coverage intervals. Straight-line fits are shown to guide the eye.

Table 1. Slope a in mm^{-1} and offset b of a straight line $d' = a \cdot D + b$ fitted to d' (NPWE) as a function of target diameter D in mm for different values of the time-current product in mAs (leftmost column). $u(\ )$ denotes the standard uncertainty. The column χ^2_{red} denotes the reduced χ^2 of the fit as in equation (35).

NPWE	FBP			AIDR 3D		
$I \cdot t$ in mAs	$a \pm u(a)$ in mm^{-1}	$b \pm u(b)$	χ^2_{red}	$a \pm u(a)$ in mm^{-1}	$b \pm u(b)$	χ^2_{red}
10	0.16 ± 0.02	0.07 ± 0.11	0.59	0.21 ± 0.02	0.08 ± 0.13	0.89
20	0.26 ± 0.02	0.17 ± 0.15	0.09	0.29 ± 0.02	0.10 ± 0.15	0.16
40	0.40 ± 0.03	-0.02 ± 0.21	0.39	0.49 ± 0.04	-0.13 ± 0.23	0.91
150	0.76 ± 0.05	0.01 ± 0.33	0.14	0.81 ± 0.05	0.12 ± 0.35	0.10
250	1.07 ± 0.07	0.43 ± 0.44	1.68	1.20 ± 0.08	0.39 ± 0.48	2.36

Table 2. Slope a in mm^{-1} and offset b of a straight line $d' = a \cdot D + b$ fitted to d' (RDI) as a function of target diameter D in mm for different values of the time-current product in mAs (leftmost column). $u(\ )$ denotes the standard uncertainty. The column χ^2_{red} denotes the reduced χ^2 of the fit as in equation (35).

RDI	FBP			AIDR 3D		
$I \cdot t$ in mAs	$a \pm u(a)$ in mm^{-1}	$b \pm u(b)$	χ^2_{red}	$a \pm u(a)$ in mm^{-1}	$b \pm u(b)$	χ^2_{red}
10	0.15 ± 0.01	-0.02 ± 0.11	0.80	0.16 ± 0.01	-0.17 ± 0.11	1.05
20	0.23 ± 0.02	0.14 ± 0.13	0.50	0.28 ± 0.02	-0.28 ± 0.13	1.18
40	0.35 ± 0.02	-0.17 ± 0.15	1.39	0.38 ± 0.02	-0.15 ± 0.17	0.34
150	0.67 ± 0.03	-0.23 ± 0.26	0.80	0.68 ± 0.03	0.08 ± 0.26	0.53
250	0.94 ± 0.05	0.03 ± 0.35	0.64	1.02 ± 0.05	-0.20 ± 0.37	0.63

Toshiba's adaptive iterative dose reduction algorithm (AIDR 3D, Toshiba Medical Systems Corporation, now a subsidiary of Canon Medical Electronics, Otawara, Tochigi, Japan).

We analysed all available data with the NPWE as described above, using a template instead of $(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)$ in equation (4) and all 84 available image pairs for testing ($\mathbf{y}_1$ with a 'lesion', $\mathbf{y}_0$ without, see Hernandez-Giron *et al* 2014). The RDI was applied as detailed above, using only 21 images $\mathbf{y}_1$ per setting. As an example, the results for one setting, i.e. for FBP reconstruction at 40 mAs are shown in figure 7, where d' is displayed as a function of the target diameter. The NPWE results are shown as open circles, the RDI results are given as filled squares. Error bars indicate 95% coverage intervals. Straight-line fits are shown in addition to guide the eye. Since the number ν of pixels inside the target scales with the square of the diameter, and on the other hand, the signal is expected to scale with $\sqrt{\nu}$, a linear relation between the detectability and the target diameter is plausible.

Table 3. The threshold visibility D_{lim} in mm (related to AUC = 75%, equivalent to $d' = 0.954$) as determined from straight-line fits to NPWE and RDI results for d' along with its standard uncertainty u as a function of the time-current product in mAs.

	NPWE		RDI	
	FBP	AIDR 3D	FBP	AIDR 3D
$I \cdot t$ in mAs	$D_{\text{lim}} \pm u$ in mm			
10	5.7 ± 0.4	4.3 ± 0.3	6.4 ± 0.3	7.3 ± 0.3
20	3.1 ± 0.4	2.9 ± 0.3	3.5 ± 0.4	4.5 ± 0.3
40	2.5 ± 0.4	2.2 ± 0.3	3.2 ± 0.3	2.9 ± 0.3
150	1.2 ± 0.4	1.0 ± 0.4	1.8 ± 0.3	1.3 ± 0.3
250	0.5 ± 0.4	0.5 ± 0.4	1.0 ± 0.3	1.1 ± 0.3

For all settings (reconstruction method, time-current product), linear fits to d' as a function of the target diameter D were obtained using a weighted total least-squares fit (TLS, see Krystek and Anton 2007) and the relation

$$d' = a \cdot D + b. \tag{34}$$

The results for the fit parameters a and b are listed in table 1 and table 2 for the NPWE and the RDI, respectively. In addition, χ^{red} is given, which is defined as

$$\chi^2_{\text{red}} = \frac{\chi^2}{k-l}, \tag{35}$$

where k is the number of data pairs used for fitting and $l = 2$ is the number of fit parameters of the straight line. χ^2 is given by

$$\chi^2 = \sum_{j=1}^{k} \left[\frac{(D_j - D_j^{(fit)})^2}{u^2(D_j)} + \frac{(d'_j - d_j'^{(fit)})^2}{u^2(d'_j)} \right], \tag{36}$$

where (D_j, d'_j) are the data points, $u(D_j)$ and $u(d'_j)$ are their respective standard uncertainties and $(D_j^{(fit)}, d_j'^{(fit)})$ are the corresponding points obtained from the TLS fit. For $u(D_j)$, a constant value of 0.5 mm was assumed, which corresponds to the size of one pixel. Ideally, χ^2_{red} should be close to unity.

From the straight-line fit, a visibility threshold D_{lim} is calculated by determining the intersection of the straight line with the horizontal at d'_{lim}, where

$$d'_{\text{lim}} = \sqrt{2} \cdot \Phi^{-1}(\text{AUC}_{\text{lim}}) \tag{37}$$

is obtained from AUC_{lim} (see equation (2) and the comment underneath equation (6)). AUC_{lim} was chosen as 75%, which defines the threshold visibility D_{lim} as the size of the target where the AUC attains the middle between guessing (AUC = 50%) and certainty (AUC = 100%). Results for the threshold visibility D_{lim} are tabulated in table 3, along with the standard uncertainty of D_{lim}.

The D_{lim} results can be directly compared to the data for a contrast of 1% listed in table 3 of the publication by Hernandez-Giron *et al* (2014). For 10 mAs, 20 mAs and 40 mAs, the newly evaluated NPWE results agree with the published data within the limits of uncertainty. It has to be kept in mind that the published data were obtained from fits of a psychometric curve to the AUC as a function of the target diameter D (Hernandez-Giron *et al* 2014), not from a straight-line fit to d' as in the present work. Only for the highest time-current products, is the threshold visibility derived from the straight-line fit significantly lower than the published data. This statement is valid for both FBP- and AIDR 3D–reconstructed images.

The evaluation of the RDI, on the other hand, yields overall slightly higher values of D_{lim} than the NPWE. For the FBP-reconstructed images, the visibility thresholds are quite close to the ones predicted by the NPWE. In the case of the AIDR 3D reconstruction, the RDI predicts slightly higher visibility thresholds than for FBP, which is somewhat surprising since iterative reconstruction methods are generally believed to improve the visibility. However, the differences are not really significant. The expected behaviour is observed for the NPWE, which predicts a (slightly) better visibility for AIDR 3D for all time-current products except for the highest value of 250 mAs, although the differences are not significant in view of the uncertainties. This statement is in accordance with the published findings in Hernandez-Giron *et al* (2014), where only small differences between FBP and AIDR 3D were reported.

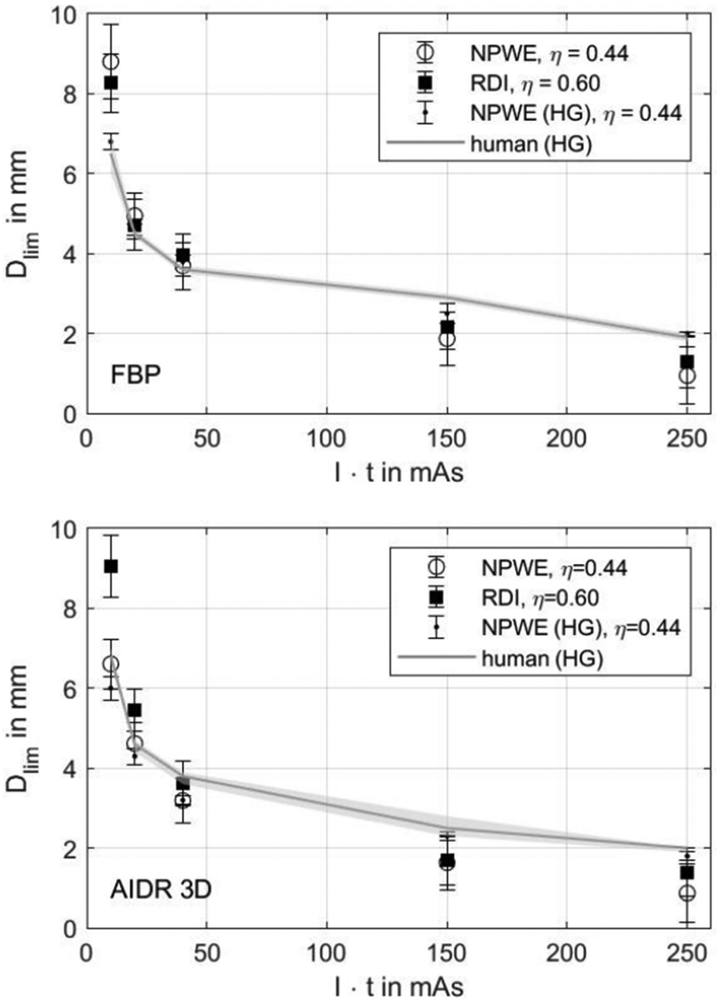

Figure 8. Threshold visibility diameter D_{lim} in mm as a function of the time-current product in mAs, evaluation of data set EXP3. Top panel: FBP; bottom panel: AIDR 3D. Open symbols: NPWE; filled symbols: RDI; black dots: NPWE (HG) from Hernandez-Giron *et al* (2014); grey curve: average human observer results (HG) from the same publication. The error bars and the shaded areas indicate the 95% coverage intervals. For details concerning the calculation of D_{lim}, please consult the text.

Table 4. Threshold visibility diameter D_{lim} (AUC = 75%) and its standard uncertainty u in units of mm for different values of the time-current product in mAs (leftmost column). The following columns display the results for the NPWE and the RDI observer, for both FBP- and AIDR 3D–reconstructed images. d' data were scaled using equation (7) and $\eta = 0.44$ for the NPWE results and $\eta = 0.60$ in the case of the RDI results. The right part of the table reproduces the results from Hernandez-Giron *et al* (2014) for the NPWE ($\eta = 0.44$) and for the human observers.

	this work				Hernandez-Giron *et al* 2014			
	NPWE ($\eta = 0.44$)		RDI ($\eta = 0.6$)		NPWE ($\eta = 0.44$)		Human	
	FBP	AIDR 3D	FBP	AIDR 3D	FBP	AIDR 3D	FBP	AIDR 3D
$I \cdot t$ in mAs	$D_{\text{lim}} \pm u$ in mm				$D_{\text{lim}} \pm u$ in mm			
10	8.8 ± 0.5	6.6 ± 0.3	8.3 ± 0.4	9.0 ± 0.4	6.8 ± 0.2	6.0 ± 0.3	6.5 ± 0.5	6.8 ± 0.2
20	4.9 ± 0.3	4.6 ± 0.3	4.7 ± 0.3	5.5 ± 0.3	4.6 ± 0.1	4.3 ± 0.2	4.5 ± 0.1	4.6 ± 0.1
40	3.7 ± 0.3	3.2 ± 0.3	4.0 ± 0.3	3.6 ± 0.3	3.8 ± 0.1	3.2 ± 0.1	3.6 ± 0.1	3.8 ± 0.1
150	1.9 ± 0.3	1.6 ± 0.3	2.2 ± 0.3	1.7 ± 0.3	2.5 ± 0.3	2.3 ± 0.1	2.9 ± 0.1	2.5 ± 0.3
250	0.9 ± 0.4	0.9 ± 0.4	1.3 ± 0.3	1.4 ± 0.3	2.0 ± 0.1	1.8 ± 0.1	1.9 ± 0.1	2.0 ± 0.1

The detectabilities obtained by the NPWE as well as those achieved through RDI will not directly match those of human observers. This is not an issue, as long as the automatically determined detectabilities and those obtained by the human observers are connected through relation (7), in which case the automatically determined detectabilities can be rescaled such that they do correspond to results of a human observer. For the NPWE, such a scaling was carried out by Hernandez-Giron *et al* (2014) using $\eta = 0.44$. For the RDI d', a value of η was determined such that a fair agreement between the threshold visibility D_{lim}—as derived from the scaled RDI d'—with the human observer data was achieved, leading to $\eta = 0.60$. The results obtained for the rescaled detectabilities of NPWE and RDI are shown in figure 8, where $D_{\lim}$ is displayed as a function of the time-current product. The top panel shows the results for the FBP reconstruction, and the bottom one shows the results for AIDR 3D, the iterative reconstruction. Open symbols indicate the present evaluation of the NPWE, whereas the filled symbols indicate the RDI results. The published data from Hernandez-Giron *et al* (2014) are also shown. Their rescaled NPWE results are indicated by the dots, whereas the average human observer results are shown as the continuous curve. Error bars and—for the human observer results—the shaded areas, represent the 95% coverage intervals.

The results are also tabulated in table 4, where the current results are given along with their standard uncertainties on the left and the published data with their uncertainties are reproduced on the right of the table, for the convenience of the reader[3]. The comparatively large differences between the NPWE results are explained by the different analysis methods: whereas the published data (Hernandez-Giron *et al* 2014) had been obtained from a fit of psychometric curves to the AUC as a function of target diameter, the present results were obtained using a straight line fit to d' as outlined above.

6. Summary and outlook

A novel regression detectability index for the task-specific image quality assessment of CT images has been proposed. The detectability index can be calculated reliably using as few as 20–40 images, and only images of one class containing the signature of a lesion or a target are needed. These properties of the detectability index are relevant practical advantages compared to current tools in task-specific image quality assessment.

The new detectability index has a prewhitening property, by the aid of a simplified model of the covariance matrix. Simplification of the covariance model turned out to be necessary in order to achieve robust results. Note that the quantity underlying the new detectability index differs from that of the prewhitening matched filter (CHO). Note further that the size of the image has an influence on the uncertainty of the proposed figure of merit, which may not be the case for the established observers. This means that some care has to be taken as to provide a sufficient number of pixels as well for the lesion part of the image as for the background part.

Applying the proposed detectability index to two sets of images from a GE CT indicated that the ASiR algorithm yields the same low-contrast detectability d' at a lower time-current product than the standard FBP reconstruction. For images obtained from a Toshiba Aquilion One CT, applying the proposed detectability index showed a high correlation with the detectabilities obtained from human observers. A comparison of the novel approach with human observers for images obtained from other CT imaging systems is highly desirable.

Currently, the novel regression detectability index is restricted to images of technical phantoms with circular targets and a homogeneous background. However, the approach may be generalized by using more complex models, consisting of more than the two regression functions of ‘background’ and ‘signal’. The proposed treatment carries over immediately regarding the estimation of the covariance matrix and the regression step; however, a new way of determining the final figure(s) of merit would have to be developed. It would also be of interest to see whether the new observer could be adapted to also work with more complex backgrounds such as anatomical ones, which could be investigated using images of anthropomorphic phantoms. Among other modifications, this would require to drop the assumption of a rotationally symmetric covariance matrix.

Due to the widespread use of iterative image reconstruction methods, practical task-specific quality assessment methods are urgently needed for acceptance and constancy testing in CT imaging and elsewhere. Although only CT data were used for demonstrating the properties of the RDI observer, it may also be applicable to other imaging modalities, such as mammography. Applying current observers, such as the NPWE, is often challenging due to its huge workload. We are confident that the developed regression detectability index can facilitate the use of task-specific image quality in many routine quality assurance procedures.

Acknowledgment

The authors wish to thank their colleagues Reinulf Boettcher and Marcel Reginatto for acquiring CT images and for inspiring discussions.

---

[3] Details of the human observer study are given on page 4 of the publication by Hernandez-Giron *et al* (2014).

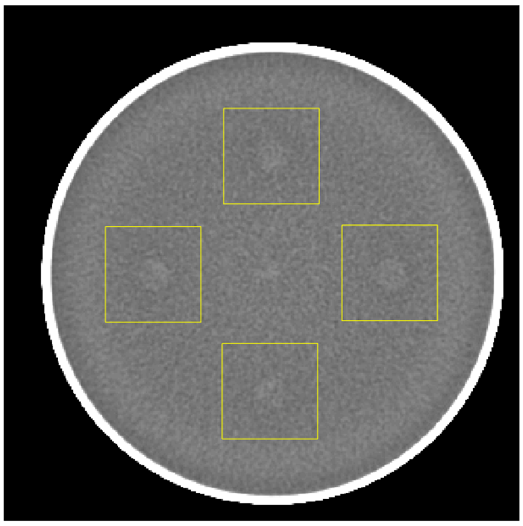

Figure A1. Image of the PTB low-contrast phantom. The interval of Hounsfield units displayed is [−25, 25], the contrast of the rods is approximately 2 Hounsfield units.

Appendix A. Data sets used

A.1. Data set EXP1: GE CT, PTB phantom

Experimental images were obtained using a low-contrast phantom designed and built at the Physikalisch-Technische Bundesanstalt (PTB), which is described in detail in section 4.2 of a previous publication (Anton *et al* 2018). The images were taken using the PTB's CT scanner, a General Electric Optima™ CT660. With a voltage of 120 kV, the product of current and exposure time was varied in five steps (5 mAs, 10 mAs, 20 mAs, 40 mAs and 80 mAs). The images were taken in 'axial scan' mode, with a slice width of 10 mm. For the reconstruction, the 'head' filter, corresponding to the phantom size, and the 'standard' convolution kernel were chosen. Every image was reconstructed using both the standard filtered back projection (FBP) procedure and GE's iterative algorithm (ASiR). The ASiR setting was chosen as 100%, which is not accessible for clinical praxis. There, a maximum setting of 50% is accessible, which was chosen for data set EXP2 (see appendix A.2).

Figure A1 shows an average of 100 images of the low-contrast phantom taken at a setting of 80 mAs, reconstructed using FBP. For this average, the signals are clearly visible. The four regions of interest (ROIs) that were used to produce images containing a signal are indicated by the squares. Images without a signal were produced by selecting ROIs that covered areas of the phantom that did not include the signal. The edge length of each of the eight square sub-images extracted from one slice was 64 px corresponding to 31.3 mm. The images of the rods had a radius of 10 pixels. In total, 400 images (ROI) with a signal and 400 images (ROI) without a signal were produced for each setting of the time-current product and constitute data set EXP1.

A.2. Data set EXP2: GE CT, PTB phantom

Data set EXP2 is similar to EXP1 (see appendix A.1), only the slice width—which is 0.625 mm in this case—and the time-current product are changed. The images that were reconstructed with GE's iterative algorithm, ASiR, were acquired at approximately half of the time-current product as the FBP images. ASiR was set to 50%. Following the GE CT's manual, ASiR images at a setting of 50% should yield an image quality equal to FBP-reconstructed images acquired at a 1/0.49 times larger time-current product. Therefore, FBP images were acquired at 51.5 mAs, 77 mAs, 100 mAs, 156 mAs, 204 mAs and 404 mAs, whereas ASiR images were acquired at 25.5 mAs, 38 mAs, 50 mAs, 78 mAs, 102 mAs and 200 mAs. Other settings were identical. Again, 400 square ROIs with an edge length of 64 px (31.3 mm), containing a target and 400 equivalent ROIs showing only background were extracted per time-current setting. Some characteristic features of the data are illustrated in figure A2 and figure A3, for FBP and ASiR. The examples are also typical of data set EXP1.

A.3. Data set EXP3: Toshiba Aquilion ONE, Catphan low-contrast module

Data set EXP3 comprises images of the low-contrast module of a Catphan phantom acquired at Leiden University Medical Center (LUMC) using a Toshiba Aquilion ONE CT, with both FBP and Toshiba's iterative algorithm AIDR 3D. The data set EXP3 is a subset of the data that were used in the publication by

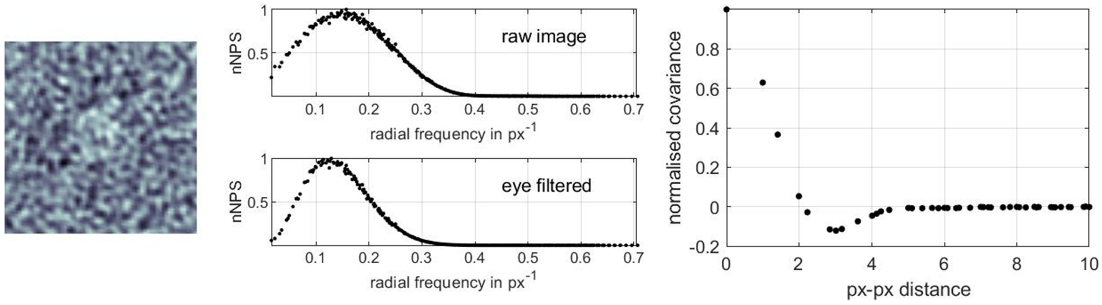

Figure A2. Data set EXP2, FBP, ≈50 mAs. Left: average over 25 images; Middle: normalized noise power spectrum (NPS) as a function of spatial frequency in px^{-1}, upper panel: raw data, lower panel: eye filtered data; Right: mean normalized covariance as a function of pixel-to-pixel distance. These graphics are also representative of data set EXP1.

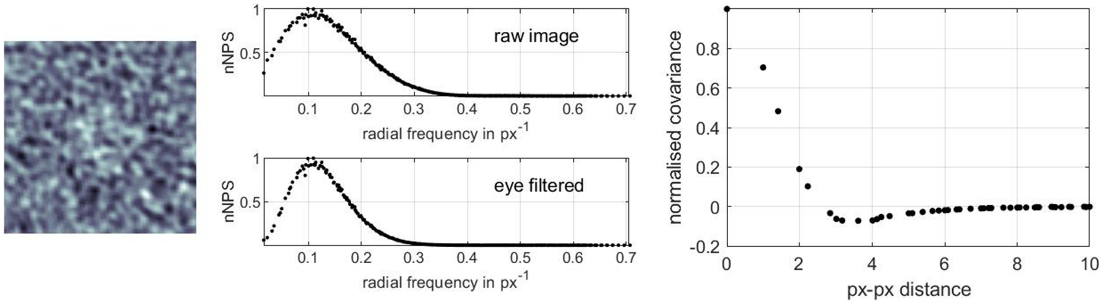

Figure A3. Data set EXP2, ASiR, ≈50 mAs. Left: average over 25 images; Middle: normalized NPS as a function of spatial frequency in px^{-1}, upper panel: raw data, lower panel: eye filtered data; Right: mean normalized covariance as a function of pixel-to-pixel distance. These graphics are also representative of data set EXP1.

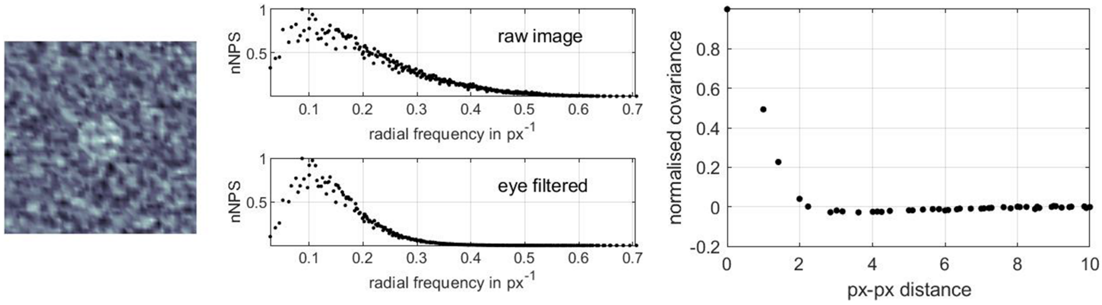

Figure A4. Data set EXP3, FBP, 40 mAs, target diameter 6 mm. Left: average over 25 images; Middle: NPS as a function of spatial frequency in px^{-1}, upper panel: raw data, lower panel: eye-filtered data; Right: mean normalized covariance as a function of pixel-to-pixel distance.

Hernandez-Giron *et al* (2014). The images are square ROIs with an edge length of 57 px corresponding to 26.7 mm, centred on the targets. The targets are the group of nine areas with a stated contrast of 1% with diameters between 2 mm and 9 mm (in steps of 1 mm) plus a 15 mm target. For each of the five settings of the time-current product (10 mAs, 20 mAs, 40 mAs, 150 mAs and 250 mAs), images reconstructed using FBP and AIDR 3D are available. For each time-current product setting, 84 images with a target, 84 images without a target and a template are available. The template is a noise-free image of the corresponding target, including the MTF of the CT. A detailed description of the data preparation procedure can be found in Hernandez-Giron *et al* (2014). An important reason for re-investigating this data set is that human observer data are available for comparison. Some characteristic features of the data are illustrated in figures A4 and A5, for FBP and AIDR 3D image reconstruction.

A.4. Fits of the covariance model to experimental data

The left part of figure A6 shows an example of the mean normalized covariance calculated from data set EXP2 (see appendix A.2) for the whole range of pixel-to-pixel distances for the square image with an edge

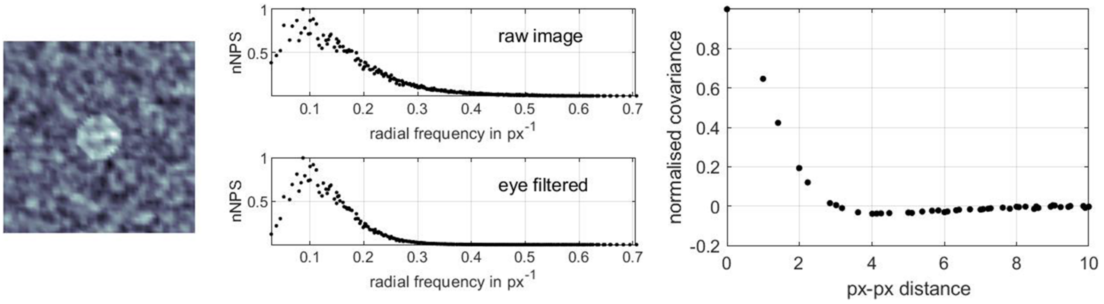

Figure A5. Data set EXP3, AIDR 3D , 40 mAs, target diameter 6 mm. Left: average over 25 images; Middle: NPS as a function of spatial frequency in px^{-1}, upper panel: raw data, lower panel: eye-filtered data; Right: mean normalized covariance as a function of pixel-to-pixel distance.

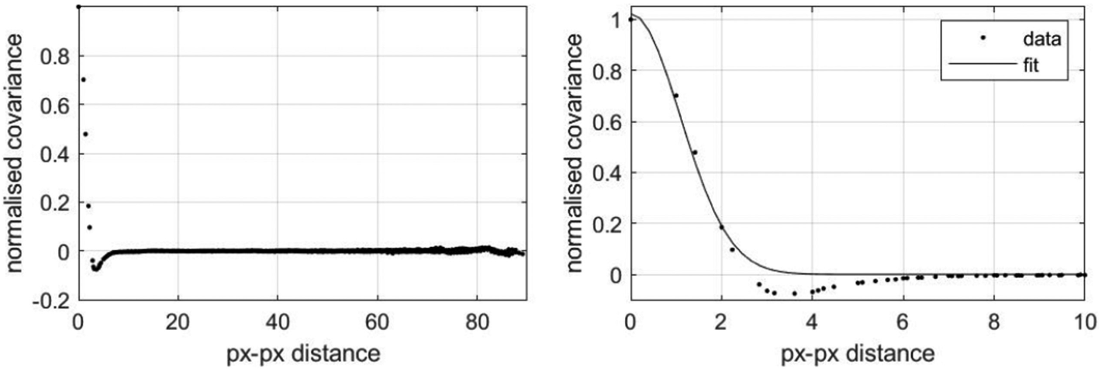

Figure A6. Example of data set EXP2 (see appendix A.2), GE CT, PTB phantom and ASiR. Left: mean normalized covariance as a function of pixel-to-pixel distance; Right: close-up of the mean normalised covariance for short pixel-to-pixel distances, the continuous line represents a simplified fitting curve according to (20).

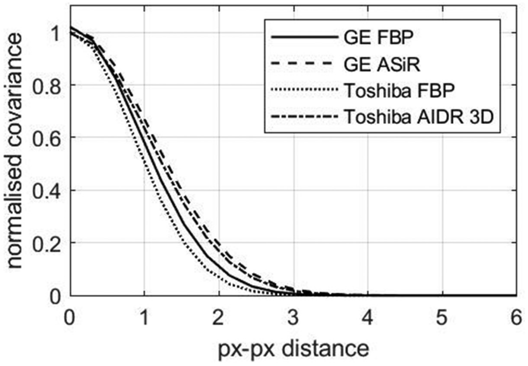

Figure A7. Normalized covariance as a function of pixel-to-pixel distance, fit function equation (20), continuous line: GE CT, PTB phantom, FBP reconstruction; dashed line: GE CT, PTB phantom, ASiR image; dotted line: Toshiba Aquilion ONE CT, FBP reconstruction; dash-dotted line: Toshiba Aquilion ONE CT, AIDR 3D reconstruction.

length of 64 px, where 'normalized' means that a scaling as in equation (29) is applied. The mean is over all pixels with identical distances. The disturbances visible for large distances are artefacts caused by the finite image size. Therefore, the fit is usually restricted to a shorter pixel-to-pixel distance (≈25–35 for an image size of 64× 64). A close-up of the fit for the nearest pixels is shown in the right of figure A6. figure A6 also shows that negative covariances can emerge and that hence the chosen covariance model yields an approximation of the true covariances only. However, according to our experience, the results achieved by using the approximate covariance model are significantly more robust than those employing a refined and more flexible covariance model.

Fitting results for data from PTB's GE CT (data sets EXP1 and EXP2) are compared to fits to CT data from the LUMC's Toshiba Aquilion ONE (data set EXP3) in figure A7, where the normalized covariance is shown as a function of the pixel-to-pixel distance. The continuous curve represents the fit to data from EXP2 (see appendix A.2) with FBP reconstruction, whereas the dashed line shows the fit to the corresponding data obtained using ASiR. The dotted curve shows the fit to data from set EXP3 (see appendix A.3) reconstructed by FBP, whereas the dash-dotted line represents the fit to the corresponding iterative reconstruction (AIDR 3D). The results are representative in the sense that the parameters of the fit essentially depend on the type of CT (GE vs. Toshiba) and the type of reconstruction algorithm. Other parameters such as the time-current product have no noticeable influence. For both CTs, the normalized covariance drops faster in the case of FBP reconstruction. The main differences between Toshiba and GE are different noise power spectra (see figures A2–A5) and that the undershoot seen in the normalized covariance is much less pronounced for the data from the Toshiba CT.

References

Abbey C K and Barrett H H 2001 Human- and model-observer performance in ramp-spectrum noise: effects of regularization and object variability *J. Opt. Soc. Am.* A **18** 473–88

Anton M, Khanin A, Kretz T, Reginatto M and Elster C 2018 A simple parametric model observer for quality assurance in computer tomography *Phys. Med. Biol.* **63** 075011

Ba A *et al* 2018 Inter-laboratory comparison of channelized hotelling observer computation *Med. Phys.* **45** 3019–30

Balta C, Bouwman R W, Sechopoulos I, Broeders M J, Karssemeijer N, Engen R E and Veldkamp W J 2018 A model observer study using acquired mammographic images of an anthropomorphic breast phantom *Med. Phys.* **45** 655–65

Barrett H H, Abbey C K and Clarkson E 1998 Objective assessment of image quality. III. ROC metrics, ideal observers and likelihood-generating functions *J. Opt. Soc. Am.* A **15** 1520–35

Barrett H H and Myers K J 2003 *Foundations of Image Science* (New York: Wiley)

Barrett H H, Myers K J, Hoeschen C, Kupinski M A and Little M P 2015 Task-based measures of image quality and their relation to radiation dose and patient risk *Phys. Med. Biol.* **60** R1–R75

Barten P G 1999 *Contrast Sensitivity of the Human eye and its Effects on Image Quality* (Bellingham, WA: SPIE Optical Engineering Press) vol 19

Barten P G J, 1992 Physical model for the contrast sensitivity of the human eye *Proc. SPIE* **1666** 57–72

Bouwman R W, Mackenzie A, van Engen R E, Broeders M J M, Young K C, Dance D R, Heeten den and Veldkamp W J H 2017 Toward image quality assessment in mammography using model observers: Detection of a calcification-like object *Med. Phys.* **44** 5726–39

Burgess A E 1994 Statistically defined backgrounds: performanceof a modified nonprewhitening observer model *J. Opt. Soc. Am.* A **11** 1237–42

Burgess A E 1999 Visual signal detection with two-component noise: low-pass spectrum effects *J. Opt. Soc. Am.* A **16** 694–704

Directorate-General for Energy (European Commission) 2015 *Medical radiation exposure of the European population* (Luxembourg: European Union)

Efron B 1982 *The Jackknife, the Bootstrap and Other Resampling Plans* (Philadelphia, PA: SIAM) vol 38

Gifford H C, King M A, de Vries D J and Soares E J 2000 Channelized Hotelling and human observer correlation for lesion detection in hepatic SPECT imaging *J. Nucl. Med.* **41** 514–21

Gill P E, Murray W and Wright M H 1981 *Practical Optimization* (London: Academic)

He X and Park S 2013 Model observers in medical imaging research *Theranostics* **3** 774

Hernandez-Giron I, Calzado A, Geleijns J, Joemai R M S and Veldkamp W J H 2014 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 *The British Journal of Radiology* **87** 20140014

ICRU 2006 ICRU Report No. 54: Medical imaging—the assessment of image quality *Journal of the ICRU* **6**

Illers H, Buhr E and Hoeschen C 2005 Measurement of the detective quantum efficiency (DQE) of digital x-ray detectors according to the novel standard IEC 62220-1 *Radiat. Prot. Dosim.* **114** 39–44

JCGM100 2008 Evaluation of measurement data—Guide to the expression of uncertainty in measurement. GUM 1995 with minor corrections Technical report, BIPM, Working Group 1 of the Joint Committee for Guides in Metrology (JCGM/WG 1) (www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdfw)

Khanin A, Anton M, Reginatto M and Elster C 2018 Assessment of CT image quality using a Bayesian framework *IEEE Trans. Med. Imaging* **37** 2687–94

Kretz T, Anton M, Schaeffter T and Elster C 2019 Determination of contrast-detail curves in mammography image quality assessment by a parametric model observer *Phys. Medica.* **62** 120—128

Krystek M and Anton M 2007 A weighted total least-squares algorithm for fitting a straight line *Meas. Sci. Technol.* **18** 1–5

Myers K J and Barrett H H 1987 Addition of a channel mechanism to the ideal-observer model *J. Opt. Soc. Am.* A **4** 2447–57

Pepe M S 2003 *The Statistical Evaluation of Medical Tests for Classification and Prediction* (Oxford: Oxford University Press)

Racine D, Ryckx N, Ba A, Becce F, Viry A, Verdun F R and Schmidt S 2018 Task-based quantification of image quality using a model observer in abdominal CT: a multicentre study *Eur. Radiol.* **28** 5203–10

Reginatto M, Anton M and Elster C 2017 Assessment of CT image quality using a Bayesian approach *Metrologia* **54** S74

Samei E *et al* 2019 Performance evaluation of computed tomography systems: Summary of AAPM Task Group 233 *Med. Phys.* **46** e735–e756

Vaishnav J Y, Jung W C, Popescu L M, Zeng R and Myers K J 2014 Objective assessment of image quality and dose reduction in CT iterative reconstruction *Med. Phys.* **41** 071904

Verdun F *et al* 2015 Image quality in CT: From physical measurements to model observers *Phys. Medica: Eur. J. Med. Phys.* **31** 823–43

Viry A, Aberle C, Racine D, Knebel J-F, Schindera S T, Schmidt S, Becce F and Verdun F R 2018 Effects of various generations of iterative CT reconstruction algorithms on low-contrast detectability as a function of the effective abdominal diameter: A quantitative task-based phantom study *Phys. Medica* **48** 111–18

Willemink M J and Noël P B 2019 The evolution of image reconstruction for CT—from filtered back projection to artificial intelligence *Eur. Radiol.* **29** 2185–95

Wunderlich A 2015, IQmodelo: Statistical Software for Image Quality Assessment with Model Observers (https://github.com/DIDSR/IQmodelo)

Wunderlich A, Noo F, Gallas B D and Heilbrun M E 2015 Exact confidence intervals for channelized hotelling observer performance in image quality studies *Medical Imaging, IEEE Trans.* **34** 453–64