BJR

Received: 31 December 2013 | Revised: 9 April 2014 | Accepted: 14 May 2014

doi: 10.1259/bjr.20140014

Cite this article as:
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.

FULL PAPER

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

[1,2]I HERNANDEZ-GIRON, MSc, [2]A CALZADO, PhD, [3]J GELEIJNS, PhD, [3]R M S JOEMAI, PhD and [3]W J H VELDKAMP, PhD

[1]Física Mèdica, Universitat Rovira i Virgili, Tarragona, Spain
[2]Departamento de Radiología, Universidad Complutense de Madrid, Madrid, Spain
[3]Radiology Department, Leiden University Medical Center, Leiden, Netherlands

Address correspondence to: Miss Irene Hernandez-Giron
E-mail: *irene.debroglie@gmail.com*

Objective: To compare low-contrast detectability (LCDet) performance between a model [non-pre-whitening matched filter with an eye filter (NPWE)] and human observers in CT images reconstructed with filtered back projection (FBP) and iterative [adaptive iterative dose reduction three-dimensional (AIDR 3D; Toshiba Medical Systems, Zoetermeer, Netherlands)] algorithms.
Methods: Images of the Catphan® phantom (Phantom Laboratories, New York, NY) were acquired with Aquilion ONE™ 320-detector row CT (Toshiba Medical Systems, Tokyo, Japan) at five tube current levels (20–500 mA range) and reconstructed with FBP and AIDR 3D. Samples containing either low-contrast objects (diameters, 2–15 mm) or background were extracted and analysed by the NPWE model and four human observers in a two-alternative forced choice detection task study. Proportion correct (PC) values were obtained for each analysed object and used to compare human and model observer performances. An efficiency factor (η) was calculated to normalize NPWE to human results.
Results: Human and NPWE model PC values (normalized by the efficiency, $\eta = 0.44$) were highly correlated for the whole dose range. The Pearson's product-moment correlation coefficients (95% confidence interval) between human and NPWE were 0.984 (0.972–0.991) for AIDR 3D and 0.984 (0.971–0.991) for FBP, respectively. Bland–Altman plots based on PC results showed excellent agreement between human and NPWE [mean absolute difference $0.5 \pm 0.4\%$; range of differences (−4.7%, 5.6%)].
Conclusion: The NPWE model observer can predict human performance in LCDet tasks in phantom CT images reconstructed with FBP and AIDR 3D algorithms at different dose levels.
Advances in knowledge: Quantitative assessment of LCDet in CT can accurately be performed using software based on a model observer.

CT has become one of the most used techniques in radiology departments. Its progressive introduction in healthcare services and the increasing number of CT scans performed worldwide per year has raised the concern about the related radiation dose.[1,2] Several improvements have been incorporated in the scanners to obtain images at the lowest achievable dose without losing relevant diagnostic information. Among them, iterative reconstruction techniques are promising. Several studies have shown that, with these algorithms, the image noise can be decreased and that higher contrast-to-noise ratios (CNRs) can be obtained compared with traditional filtered back projection (FBP) and thus a significant dose reduction can be achieved.[3–6]

A wide variability in dose and image quality has been found between different CT scanners to perform similar diagnostic tasks.[7] To assess image quality, low-contrast detectability (LCDet) is determined as the smallest object visible for certain contrast value at a given dose level. LCDet can be subjectively assessed by several observers scoring the visibility of objects on CT phantom images. These studies are time consuming and expensive owing to the large required number of observers and observations.[8] The range of available protocols and custom parameters for each application adds complexity to optimization too.[9] Furthermore, the results might be biased if the observers know beforehand the location of the objects in the phantom. Tests of statistical significance are controversial to

obtain average results based on human observer studies, as a great inter- and intra-observer variability may appear.[10,11] Computer model observers, intended to predict the performance of human observers in image analysis, can be an alternative to objectively assess image quality. They can be a useful tool when investigating the influence of acquisition and reconstruction parameters on image quality or the effect of object size, shape and contrast in detection tasks.[12–15]

In a previous work, an objective statistical method using a specific model observer [non–pre-whitening matched filter with an eye filter (NPWE)] was presented to investigate the influence of different CT acquisition parameters on LCDet.[16]

The main goal of this work is to compare the model observer LCDet performance in CT images acquired at different dose levels with human observers. Images reconstructed with two algorithms (FBP and iterative) were used in this study. Two-alternative forced choice (2-AFC) experiments, in which the observers scored samples containing signals or background (Bg) extracted from the images, were carried out. The results were presented at the Medical Imaging Perception Society XV Conference held in Washington DC during 14–16 August 2013, which is focused on observer performance analysis and diagnostic quality of imaging technique improvements.

METHODS AND MATERIALS

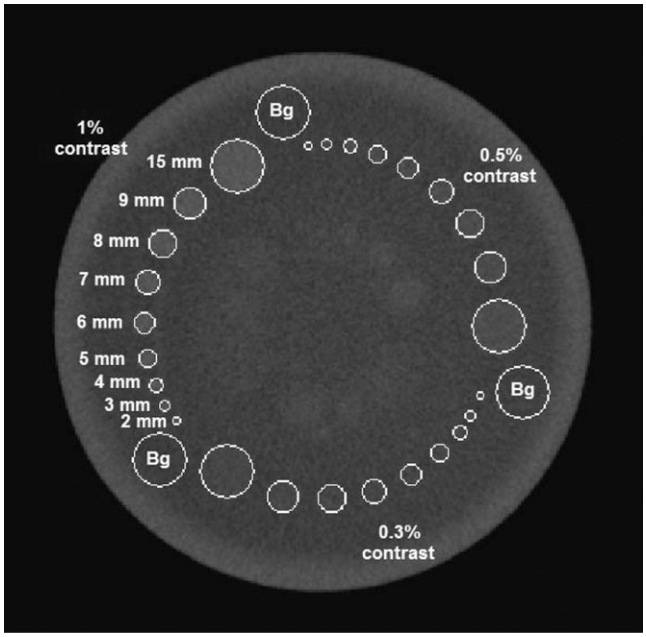
Image acquisition

Throughout this study, images of the Catphan® 500 phantom (Phantom Laboratories, New York, NY), which is dedicated to quality control tasks on CT scanners, were used. The low-contrast module (CTP515) contains three groups of cylindrical rods of various diameters (2–15 mm) and three contrast levels (0.3%, 0.5% and 1.0% nominal contrast), as shown in Figure 1. The nominal contrast (expressed as a percentage) is defined by the Catphan manufacturer as the difference in CT number between the target object and the background divided by 10.

CT images of the phantom were acquired on a 320-detector row CT scanner (Aquilion ONE™; Toshiba Medical Systems, Tokyo, Japan) by selecting the following parameters: 64×0.5-mm beam collimation, 240-mm field of view, helical acquisition (pitch, 0.828), 120 kVp tube voltage, 0.5 s rotation time and five different tube current levels (20, 40, 80, 300 and 500 mA). Images of 0.5-mm slice thickness were reconstructed with a soft-body kernel (FC13), which enhances low frequencies in the image, reduces high-frequency noise and smooths the appearance of the image in general. Two reconstruction algorithms were selected: FBP and an iterative algorithm [adaptive iterative dose reduction three dimensional (AIDR 3D); Toshiba Medical Systems]. The latter is an iterative algorithm that performs calculations in the raw data domain using statistical models, scanner characteristics and projection noise estimation to decrease the electronic noise and, afterwards, applies an iterative technique in the image domain to decrease image noise.[5]

The phantom was scanned two times for each tube current–time product (mA) value. To avoid possible artefacts owing to the nearby modules, only the 42 central axial images of the LC module were taken into account from each scan. Thus, ten image series (considering the five mA values and two reconstruction algorithms used), composed by 84 images each, were available for the model and human observer tests in this study.

Figure 1. A constructed 40-mm thick slice of the Catphan® (Phantom Laboratories, New York, NY) low-contrast module. The contrast groups and the object diameters are tagged for the supraslice region. The mask for the objects and the background (Bg) sample locations are overlaid in the figure.



Model observer (NPWE) and low-contrast detectability software

A software program dedicated to automated LC objects detection on CT, implemented in MATLAB® (MathWorks®, Natick, MA), was described in a previous work.[16] The improvements implemented in the methodology are explained in detail in this section.

To locate the LC objects in the CT images, a mask of the distribution of the disks in the phantom was created. The manufacturer specifications (size, shape, position and contrast) were used to generate templates to match the objects in the real CT images (Figure 1). The object templates were blurred to model the modulation transfer function in each case, which was obtained as the full width at half maximum of the point spread function (PSF).[17] Images of a phantom containing a 0.18-mm diameter tungsten bead were acquired for the different mA values and reconstruction algorithms to measure the PSF values. A thick slice is automatically created for each mA set by averaging all the available related images. To optimize the detection of the objects in the CT images, the templates were individually shifted 3×3 pixels around the initial location estimated using Catphan specifications.

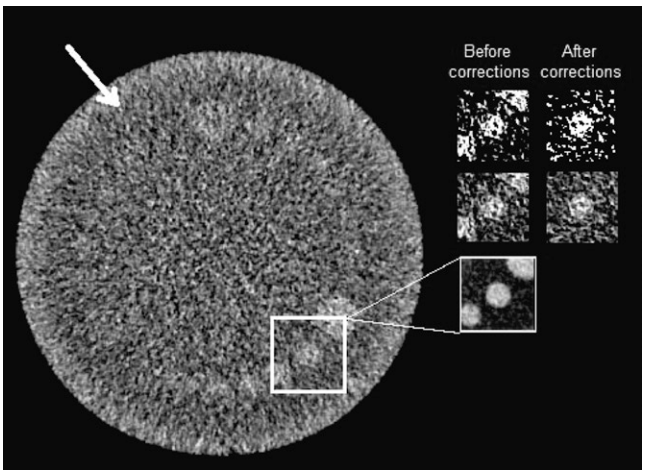
A circular white band was found close to the outer rim of the phantom images (Figure 2). These background inhomogeneities

may affect LCDet. To correct them, in the thick slices previously created, for each of the image sets individually, an annular-shaped region of interest (ROI) was taken around the 15-mm object of each contrast group, and another circular ROI was taken on these objects. The signal difference in Hounsfield units was measured between these regions. Based on these values, artificial signals were created, blurred by the measured PSF value and subtracted from the thick slice image. The resulting thick slice (equivalent to the LC module without objects in it) was then subtracted from the individual CT images.

To avoid any bias in the human observer study, the samples taken from the CT images should have the same size, independently of object diameter. The geometrical distribution of the LC objects in the Catphan phantom was a limitation for this purpose, as nearby objects could be included in the samples. To overcome this, an additional image correction was performed, using the templates previously created, to wipe out, from each object sample, the nearby objects in its corners.

The effect of these corrections (Bg inhomogeneities and object wipe out) in the images was analysed comparing the noise and contrast in the original and corrected images. For each mA and FBP/AIDR 3D series (for either the corrected or original set), the mean pixel value and the standard deviation σ (used as a measure of noise) were measured in ROIs of size $26.7 \times 26.7\,mm^2$ taken in the Bg sample locations (Figure 1). A relative difference value (%) was calculated for each condition as $(\sigma_{original\ images} - \sigma_{corrected\ images})/\sigma_{corrected\ images}$. Regarding the effect on contrast, a ROI was defined at the exact location of the 15-mm object for the three contrast groups. Contrast (C) was measured, averaged for each set (original or corrected image), and relative difference values were obtained as $(C_{original\ images} - C_{corrected\ images})/C_{corrected\ images}$.

Figure 2. An example of the wiping out of nearby object processes in the Catphan phantom CT images for the 150 mA filtered back projection series. Inside the white square, a crop of the thick slice is shown before the correction. For one of the images in the set, the object samples are shown with different window settings before and after the corrections. The arrow highlights the band background inhomogeneities.



For the 2-AFC experiment, Bg samples were extracted from an area located close to the smallest disk of each contrast group but positioned farther from the module centre (Figure 1). Object (signal) samples were extracted following the process explained above. Both types of samples had the same size ($26.7 \times 26.7\,mm^2$) for all the objects in the module with independence of their diameter.

The software automatically calculated LCDet using an NPWE model observer for each object and the three contrast groups present in the LC module of the phantom. This model is based on the assumption that the human observer uses templates of the expected signals for cross-correlation in the images and that it is unable to modify the template to pre-whiten correlated noise. The addition of an eye filter (E) takes into account the spatial frequency (f) response of the human eye. We selected the eye filter proposed by Burgess $E(f) = fe^{-bf}$, with b chosen such that $E(f)$ peaked at four cycles per degree and assuming a fixed viewing distance of 50 cm from the monitor.[18] Different studies have shown that human performance lies between pre-whitening and non–pre-whitening, depending on the spectral distribution of the image noise.[19,20]

For each object in the phantom, the model cross-correlates the samples (signal or Bg) taken from the 84 images of the set with the appropriate template (blurred expected signal), after filtering them by an eye filter (E).[18] This results in T_1 (correlations of the template and object samples) and T_2 (correlations of the template and Bg samples). Based on distributions of the test statistics of the correlation results, a discrimination index d' was calculated applying Equation (1):[18]

$$d' = \frac{\langle T\rangle_1 - \langle T\rangle_2}{\sqrt{\frac{1}{2}\sigma_1^2 + \frac{1}{2}\sigma_2^2}}, \quad (1)$$

where $< >$ refers to the mean and $\sigma(\ )$ is the standard deviation; subindexes 1 and 2 are related to the object and to the Bg distributions of test statistics, respectively.

This procedure was performed for all the contrast groups in the phantom and repeated for the five selected mA values and two reconstruction techniques sets. The detectability index d' was expressed as a function of object diameter for the three contrast groups and each condition. Then, d' values were transformed into proportion correct (PC) using Equation (2):[16,18]

$$PC = 0.5 + 0.5\,\mathrm{erf}\left(\frac{d'}{2}\right) \quad (2)$$

where erf (x) is the error function given by Equation (3):

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^{\infty} e^{-x^2}dx \quad (3)$$

This method was applied for the 10 CT image series, and thus, d' and PC profiles as a function of the object diameter were obtained for each mA and FBP or AIDR 3D sets. As, just by chance, in a 2-AFC experiment, a default PC = 50% value can

be obtained, the detectability threshold (λ) was fixed at PC = 75%. Thus, when PC $\geq$ 75% in the analysis, the related object diameter was considered visible.

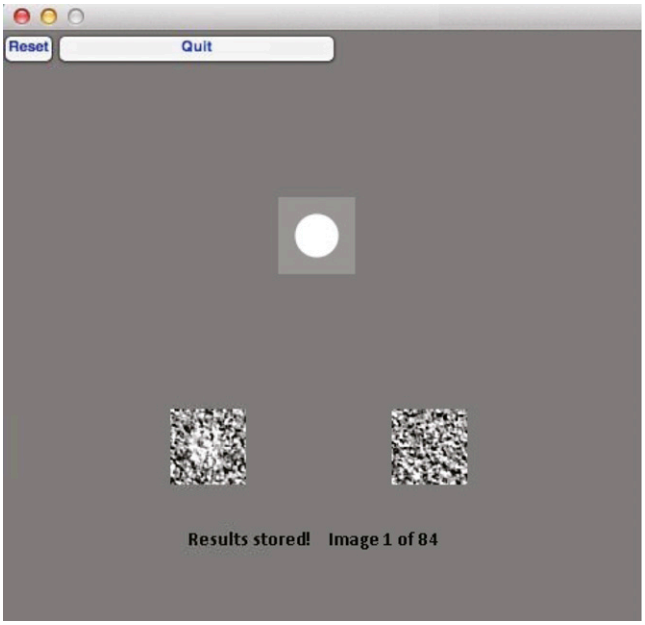
Human observer two-alternative forced choice study

To validate the trends shown by the NPWE, a 2-AFC human observer study was carried out by four medical physicists, each of them scoring pairs of ROIs (signal or Bg samples) extracted from the different sets of images for the 1% contrast group. To analyse intra-observer variability, each observer scored twice the 84 pairs of images (the same used for the NPWE model) related to a given object diameter acquired at certain mAs and reconstructed with FBP or AIDR 3D. Thus, each observer scored 84 (pairs of images) $\times$ 9 (diameters) $\times$ 5 (mA values) $\times$ 2 (FBP or AIDR 3D) $\times$ 2 (intra-observer variability), which makes 15,120 images in total.

For the signal known exactly and background known exactly (SKE/BKE) task performed in the human observer study, an application was created in MATLAB. In Figure 3, an example of the 2-AFC software interface is shown: two images are displayed together with the template on a grey canvas; the one which contains the object must be clicked on and scoring results are automatically stored in an output file. The object always appears in the centre of the sample, as shown in the template. The images with or without object were displayed randomly at left or right. Each set of images (for a given diameter, mA and reconstruction method) was independently scored and images related to different conditions were not mixed in this study.

The scoring was performed on an i-MAC 27″ (Apple Inc., cupertino, CA) monitor using recommended visualization conditions, with fixed values for window level and width (taken as 3σ, where σ is the average standard deviation of pixel values of the Bg samples for each series). The quotient between the maximum and minimum luminance that the monitor can deliver or luminance ratio was 491, and the measured ambient luminance was kept $<$10 lux.[21]

Figure 3. Interface of the two-alternative forced choice software used for the human observer experiment. The template (above) is shown together with signal and background samples extracted from the images.



One training session was programmed for the observers to get used to the software features and the task. All observers scored the images twice (without any time limitation to review them) in four different sessions (two for each reconstruction method to analyse the intra-observer variability), which lasted approximately 2 hours each. There was a gap of at least 2 weeks between them, to avoid learning effects. The viewing distance was fixed at 50 cm, and the observers were allowed to rest whenever they wanted to avoid fatigue.

An analysis of the intra-observer consistency was performed using the Wilcoxon signed rank test for matched-pair samples (consistent results if $p \geq 0.05$) by comparing the scores for each object size and mA separately obtained in each session for AIDR 3D and FBP.[22] If one observer was inconsistent in his results between both sessions for a given condition, that scoring was ruled out.[16] The average human observer performance was obtained as the mean of the PC values that passed the intra-observer tests for each condition. Finally, PC curves, as a function of the object diameter, were obtained for each mA and either FBP or AIDR 3D.

Efficiency (η) calculation and agreement between human and model observer

To obtain an efficiency (η) between the human observers and the model in our experiments, PC values had to be transformed into d' applying Equation (4):[12,23,24]

$$d' = \sqrt{2}\Phi^{-1}(\mathrm{PC}) \quad (4)$$

where $\Phi^{-1}(\mathrm{PC})$ is the inverse of the standard cumulative normal distribution function.

Finally, η could be calculated to relate the average human observer performance (d'_{human}) to the model observer (d'_{NPWE}) by applying Equation (5) and using a least-squares procedure to fit the data.[12,19] The error bars used as weights in the linear fit were estimated as 2σ, where σ is the standard deviation of the d'_{human} squared values. The efficiency η tallied the linear fit slope.

$$\left(d'_{\mathrm{human}}\right)^2 = \eta \left(d'_{\mathrm{NPWE}}\right)^2 \quad (5)$$

To study the agreement of the NPWE and human observers, their related PC values were compared using Bland–Altman plots using EpiDat software.[25] Additionally, Pearson's product–moment correlation coefficients (r) between human and model PC scorings were calculated for both reconstruction methods and each mA separately (perfect correlation if the absolute value of $r = 1.0$).[14]

Psychometric fits and visibility thresholds

Psychometric fits were performed for the obtained PC profiles as a function of the object diameter.[26,27] For this 2-AFC experiment, fitting curves according to Equation (6) were applied for

each mA and reconstruction set independently, for both the average human and model observer.[16] For the average human observer, the error bars related to the PC values, previously calculated, were used as weights in the fitting process based on a least-squares procedure. The range of the fitting curves runs from 0.5 (pure guessing) and 1.00 (certain detection).

$$PC = \frac{0.5}{1 + e^{-f\log\left(\frac{d}{\lambda}\right)}} + 0.5 \quad (6)$$

where d represents the object diameter and f and λ are the fitting parameters. The steepness of the psychometric curve is determined by f. The smallest object diameter, which matches the proposed visibility threshold (PC = 75%) is λ itself.

In the case of the NPWE, additional psychometric fits were performed using the PC values corrected by the efficiency value η.

Image quality comparison between both reconstruction algorithms

To analyse the effect of selecting FBP or AIDR 3D in LCDet performance, two-tailed paired *t*-tests ($\alpha = 0.05$) were performed comparing d' values obtained with NPWE model and all contrast groups for the different mAs. Similar tests were also performed using the PC values obtained for the 1% contrast group and all mAs, by the human observers and the model observer, respectively.[22]

Additionally, an estimation of the average noise value was obtained for each mA and reconstructed image set. Pixel noise was measured as the standard deviation (σ) of the pixel values, in three circular ROIs taken at the same locations as the Bg samples, and the average noise value was calculated. A relative difference value (%) between FBP and AIDR 3D sets was obtained for each mA as $(\sigma_{FBP} - \sigma_{AIDR\,3D})/\sigma_{FBP}$. A repeated measures analysis of variance (ANOVA) test was performed between the noise measurements calculated for both algorithms and each mA separately (significant differences if $p \leq 0.05$) in the original images.

RESULTS

Analysing the signal and Bg samples before and after applying the Bg corrections (to suppress undesired Bg trends and to wipe out nearby objects), it was found that contrast varied <5% in all cases. The standard deviation of pixel values, which reflects the combined effect of inhomogeneities and noise in the images, was also reduced after applying these corrections in the range 4–10%. To depict the effect on the images, in Figure 2, it can be seen on one of the signal samples before and after this correction.

Model observer results

The NPWE model observer obtained higher d' values with increasing object contrast. Detectability also increased approximately linearly with object diameter. In Table 1, the slopes for the linear fits performed for all the sets of d' as a function of the object diameter and the three contrast groups are summarized [95% confidence interval (CI)]. The range of R_2 for the linear fits was 0.907–0.995 for FBP and 0.890–0.993 for AIDR 3D sets, respectively.

The influence of contrast and mAs in LCDet is shown in Table 1: higher slopes are obtained with increasing contrast and mAs for both FBP and AIDR 3D. Two-tailed paired *t*-tests ($\alpha = 0.05$) were performed comparing the d' values related to the contrast groups for both reconstruction methods and each mA separately. A significant improvement in the detection of objects as contrast increased was found ($p \leq 0.05$ in all cases). Similar tests were performed to determine the differences in d', with increasing mAs indicating that NPWE showed a significant improvement in LCDet as tube current increased for all contrast groups and both reconstruction algorithms ($p \leq 0.05$).

Human observer results

To study human observers LCDet performance, 60,480 pairs of images (for 1% contrast group in the Catphan phantom) were analysed [15,120 (images scored by 1 observer) × 4 (4 observers)]. From now on we will use the term "scoring" to refer to the series of results for a given diameter, mAs and reconstruction obtained by an observer.

The intra-observer variability test led to discard ($p < 0.05$) eight individual pairs of scorings, three for FBP and five for AIDR 3D (2.2% of all the scorings). The distribution of discarded scorings by the four observers was 4, 3, 1 and 0, respectively. After filtering the results, removing the inconsistent data, no significant differences were found between the human scorings ($p \geq 0.05$).

The psychometric fits obtained for the average human observer based on the AIDR 3D scoring data (1% contrast) are illustrated in Figure 4. The related R^2 fitting values were in the ranges 0.743–0.945 for FPB reconstruction and 0.710–0.955 for AIDR 3D. The error of the mean PC value for the average human observer for the different mA series (10, 20, 40, 150 and 250 mAs) were in the ranges 0.2–18%; 0.8–12.5%; 0.8–17.8%; 0.7–16.7%; and 0.5–5% for AIDR 3D and 6.5–12.8%; 1.1–8.2%; 0.8–9.8%; 0.6–16.5% and 0.7–6.7% for FBP, respectively.

Efficiency calculation

Owing to the shape of the curve of d' as a function of PC, it is difficult to measure d' when its value is above three, approximately (PC ≈ 0.98) in a 2-AFC experiment.[28,29] Only the human PC values below this threshold were used to determine the efficiency of the NPWE model observer. In Figure 5, the d' values for the average human observer are plotted as a function of NPWE models (both squared). The data related to all the mA series for AIDR 3D and FBP for 1% contrast were taken into account in this graph. The linear fit slope, which tallies the efficiency, η, was 0.44 (0.42–0.46, 95% CI).

Visibility thresholds non-pre-whitening matched filter with an eye filter and average human observer

The visibility thresholds λ (related to PC = 75%, 95% CI) for the 1% contrast group obtained by the average human observer in the 10–250 mA range are depicted in Table 2 together with the NPWE model values, after correcting them by the efficiency ($\eta = 0.44$). It can be seen that smaller objects could be detected as mAs increased for both reconstruction algorithms by the human and model observer.

Table 1. Slopes of the linear fits of detectability index (d') as a function of object diameter for the tube current–time product (mA) range and filtered back projection (FBP)/adaptive iterative dose reduction three dimensional (AIDR 3D) algorithm reconstructed sets of images for the three contrast groups and non-pre-whitening matched filter with an eye filter model observer (values for confidence interval = 95%). The results of two-tailed paired *t*-tests (significant differences for $p \leq 0.05$) comparing FBP and AIDR 3D d' values for each condition are also shown

Contrast	Tube current–time product	10 mA	20 mA	40 mA	150 mA	250 mA
1%	FBP	0.21 (0.20–0.22)	0.37 (0.35–0.40)	0.48 (0.46–0.50)	0.81 (0.80–0.83)	1.03 (0.99–1.07)
	AIDR 3D	0.27 (0.25–0.28)	0.40 (0.37–0.43)	0.56 (0.54–0.58)	0.85 (0.83–0.86)	1.08 (1.03–1.13)
	p-value	0.005	0.052	<0.001	<0.001	0.002
0.5%	FBP	0.10 (0.09–0.10)	0.19 (0.18–0.19)	0.22 (0.22–0.23)	0.38 (0.37–0.38)	0.60 (0.58–0.62)
	AIDR 3D	0.11 (0.11–0.12)	0.18 (0.17–0.18)	0.25 (0.25–0.26)	0.39 (0.39–0.40)	0.63 (0.60–0.66)
	p-value	<0.001	<0.001	<0.001	<0.001	<0.001
0.3%	FBP	0.05 (0.05–0.05)	0.11 (0.11–0.11)	0.12 (0.11–0.12)	0.20 (0.19–0.20)	0.38 (0.37–0.39)
	AIDR 3D	0.06 (0.06–0.06)	0.09 (0.08–0.09)	0.13 (0.13–0.14)	0.20 (0.20–0.21)	0.39 (0.38–0.40)
	p-value	0.004	0.002	0.002	0.002	<0.001

For NPWE, the visibility threshold λ (related to PC = 75%) increased dramatically with decreasing contrast (Table 3, 95% CI) for both AIDR 3D and FBP. This effect was more evident below <150 mA.

Analysis of agreement between non-pre-whitening matched filter with an eye filter and human observer

The normalization of the NPWE results by the efficiency led to a high correlation with the average human observer, for all mAs and both reconstruction methods. The overall Pearson's product–moment correlation coefficients (considering all mAs) calculated for 95% CI were 0.984 (0.972–0.991) and 0.984 (0.971–0.991) for AIDR 3D and FBP, respectively. The correlations for 10, 20, 40, 150 and 250 mA are shown in Table 2. Figure 6 depicts the psychometric fits for the human observer and the NPWE model (after the efficiency correction) as a function of mAs for the FBP reconstructed sets.

Figure 7 shows the Bland–Altman plot performed for the PC values obtained by the average human observer and the NPWE model (after correction by efficiency) for AIDR 3D and FBP altogether. It showed an excellent agreement with a mean absolute difference, Δ, of 0.5 ± 0.4%. The range of the differences, given by ($\Delta - 2\sigma$, $\Delta + 2\sigma$) was (−4.7%, 5.6%), where Δ is the mean absolute difference and σ is the standard deviation of the differences between NPWE and human observers. For AIDR 3D images, the mean absolute difference (Δ) and the range of the differences were 0.4 ± 0.4% and −4.8%, 5.2%, respectively, whereas for FBP sets they were 0.4 ± 0.2% and −3.9%, 5.0%.

Image quality comparison between both reconstruction algorithms

The repeated measures ANOVA test performed to analyse the differences in the image noise when applying FBP or AIDR 3D in the original images showed significant differences for all the mA values ($F > 113{,}985$; $p < 0.001$). AIDR 3D produced a significant reduction of noise compared with FBP of 51%, 43%,

Figure 4. Psychometric fits [proportion correct (PC) as a function of the object diameter] for the average human observer and all tube current–time product (mA) for the images reconstructed with adaptive iterative dose reduction three dimensional algorithm and 1% contrast. The dots represent the average human observer PC values.

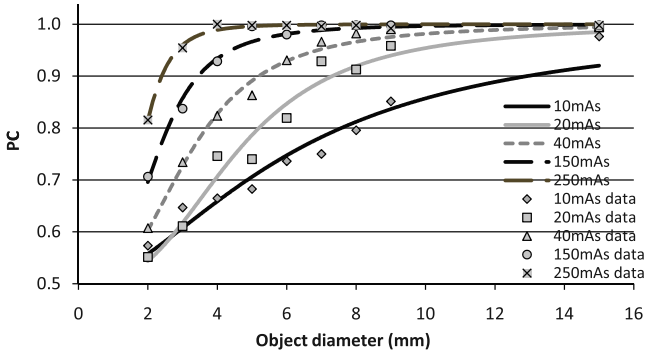


Figure 5. Squared detectability index (d'^2) for the average human observer as a function of the model observer [non-pre-whitening matched filter with an eye filter (NPWE)]. The efficiency η is given by the slope of the linear fit [95% confidence interval (CI)].

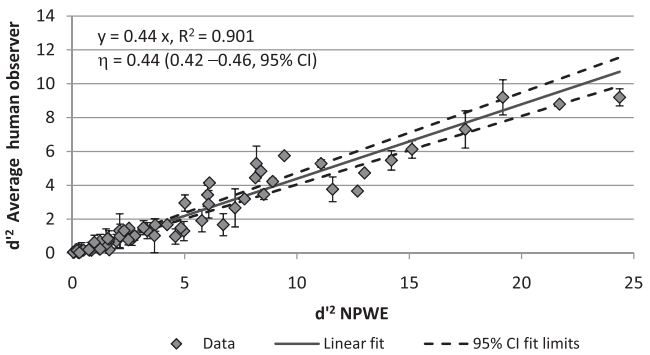


Table 2. Visibility thresholds [proportion correct (PC) = 75%] for the average human and non-pre-whitening matched filter with an eye filter (NPWE) (corrected by the efficiency) model observers and Pearson's product-moment correlation coefficients (r) for their PC values for all tube current–time products (mAs) and both reconstruction algorithms [filtered back projection (FBP) and adaptive iterative dose reduction three dimensional (AIDR 3D) algorithm] (95% confidence interval)

	λ (mm) FBP		Pearson coefficient (r)	λ (mm) AIDR 3D		Pearson coefficient (r)
	Average human	NPWE		Average human	NPWE	
10 mA	6.5 (6.0–7.0)	6.8 (6.6–7.0)	0.969 (0.857–0.993)	6.8 (6.6–7.0)	6.0 (5.7–6.3)	0.988 (0.943–0.997)
20 mA	4.5 (4.4–4.6)	4.6 (4.4–4.7)	0.983 (0.921–0.996)	4.6 (4.4–4.7)	4.3 (4.1–4.5)	0.978 (0.897–0.995)
40 mA	3.6 (3.5–3.7)	3.8 (3.6–3.9)	0.984 (0.925–0.996)	3.8 (3.6–3.9)	3.2 (3.1–3.3)	0.991 (0.953–0.998)
150 mA	2.9 (2.8–3.0)	2.5 (2.3–2.8)	0.996 (0.978–1.000)	2.5 (2.3–2.8)	2.3 (2.2–2.4)	0.989 (0.946–0.997)
250 mA	1.9 (1.8–2.0)	2.0 (1.9–2.0)	0.997 (0.986–1.000)	2.0 (1.9–2.0)	1.8 (1.7–1.9)	0.994 (0.971–0.998)

34%, 25% and 23% relative to FBP for 10, 20, 40, 150 and 250 mA, respectively.

For the NPWE model, two-tailed paired t-tests ($\alpha = 0.05$) were performed comparing the d' values obtained for FBP and AIDR 3D, each mA and all contrast groups. The related p-values for each mA are shown in Table 1. Significant improvement ($p \leq 0.05$) was shown with AIDR 3D for all mAs and contrast groups.

Figure 8 depicts the overall effect of selecting each reconstruction method on the NPWE LCDet performance showing the psychometric fits for the 0.3% contrast group. R^2 values were in the range 0.995–0.960 for FPB and 0.993–0.953 for AIDR 3D for all the contrast groups. This trend was the same for the human observer (1% contrast).

For NPWE, the results of the two-tailed paired t-tests ($\alpha = 0.05$) performed for the PC values related to each mA comparing both algorithms showed significant differences in all cases ($p \leq 0.05$). For the human observer, significant differences ($p < 0.05$) appeared for the lower mA series (10, 20 and 40 mA). No significant differences between both reconstruction methods were found for the 150 and 250 mA series (p-values of 0.05 and 0.06, respectively).

DISCUSSION

The selected model observer NPWE reproduced the LCDet performance trends of the average human observer as a function of mAs. In this study, the model and human observers scored the same sets of images (corrected to suppress undesired background trends). The model was more efficient than the human observer to detect LC objects in FPB and AIDR 3D reconstructed CT images. The calculated efficiency (0.44) is in the range obtained by other authors ($\eta \approx 0.5$) when applying the same model observer to other types of images.[14,18,29] The agreement between the model and human observer was excellent at the dose range considered in this work (10–250 mA) for both reconstruction algorithms after applying the η factor, as shown in Figure 6.

The efficiency was also calculated using all the human scorings (without discarding any values owing to intra-observer inconsistency), obtaining a slightly smaller η of 0.41 (0.39–0.43, 95% CI) in this case.

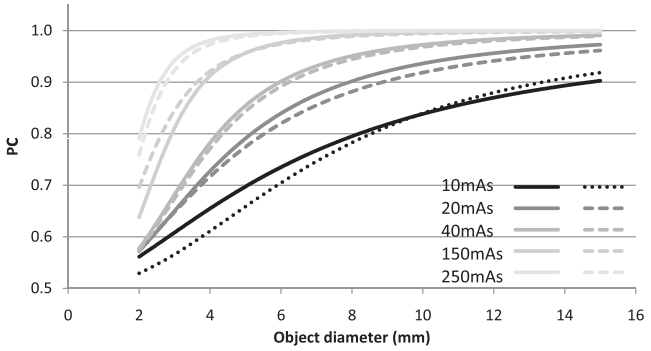
The Bland–Altman plot showed an excellent agreement ($\Delta = 0.5 \pm 0.4\%$) between the human and NPWE, the range of the differences being about ±5%. This analysis was also performed taking into account all the original human PC values to study the effect or discarding data (owing to intra-observer inconsistency) on the correlation between human and model. In this case, the differences increased on average $\Delta = -1.0\% \pm 0.7\%$ and also in range −11.2% to 9.1%.

By analysing the slopes of d' as a function of object diameter fits (Table 1), it was shown that the NPWE model LCDet

Table 3. Visibility thresholds (related to proportion correct = 75%) for the non-pre-whitening matched filter with an eye filter model and both reconstructions [filtered back projection (FBP) and adaptive iterative dose reduction three dimensional (AIDR 3D) algorithm] for all the tube current–time product (mA) series and contrast groups (confidence interval = 95%)

	λ (mm) 1% contrast		λ (mm) 0.5% contrast		λ (mm) 0.3% contrast	
	FBP	AIDR 3D	FBP	AIDR 3D	FBP	AIDR 3D
10 mA	5.1 (5.0–5.3)	4.3 (4.2–4.4)	9.6 (9.3–9.9)	8.2 (7.9–8.5)	18.4 (17.8–18.9)	14.7 (14.3–15.1)
20 mA	3.2 (3.2–3.3)	3.1 (3.0–3.1)	5.4 (5.2–5.6)	5.6 (5.5–5.8)	8.5 (8.2–8.7)	10.9 (10.5–11.4)
40 mA	2.8 (2.7–2.9)	2.6 (2.5–2.7)	4.7 (4.5–4.9)	4.0 (3.9–4.1)	8.2 (7.9–8.5)	7.3 (7.1–7.5)
150 mA	1.9 (1.9–1.9)	1.8 (1.8–1.8)	2.8 (2.7–2.9)	2.6 (2.5–2.7)	4.9 (4.7–5.1)	4.6 (4.5–4.8)
250 mA	1.7 (1.7–1.7)	1.7 (1.7–1.7)	2.1 (2.0–2.1)	1.9 (1.9–2.0)	2.9 (2.9–3.0)	2.8 (2.7–2.8)

Figure 6. Psychometric fits for the human (lines) and the non-pre-whitening matched filter with an eye filter model (dashed lines) based on the results for the filtered back projection reconstructed images and all tube current-time products (mAs) for 1% contrast objects. PC, proportion correct.



performance significantly improved for all mAs and contrast groups with AIDR 3D ($p \leq 0.05$). These trends were also reflected in the psychometric fits for both, humans and model (Figure 8), obtaining higher PC values with AIDR 3D. In general, AIDR 3D showed an overall improvement in detectability as object diameter increased, compared with FBP for the entire dose range. The two-tailed *t*-tests performed for the PC values and each mA showed significant improvement ($p \leq 0.05$) for the NPWE when using AIDR 3D in all the dose range. For the human observer, significant improvement was found only in the range 10, 20 and 40 mA when applying the iterative algorithm.

The visibility thresholds for 1% contrast showed differences between both reconstruction methods, with the same trends for the model and human observers, but they were very subtle for high mAs. It has been noted that for the human observers, no significant differences between the algorithms were found between the PC values obtained for the higher mAs (150–250 mA).

Figure 7. Bland–Altman plot of proportion correct (PC) difference between human and model observer (after correcting by efficiency) for filtered back projection (FBP) (○) and the adaptive iterative dose reduction three dimensional (AIDR 3D) algorithm (◆). The black line represents the average absolute difference Δ (0.5 ± 0.4%); the two dash lines represent $\Delta \pm 2\sigma$, where σ is the standard deviation of the differences, which are −4.7%, 5.6%. NPWE, non-pre-whitening matched filter with an eye filter.

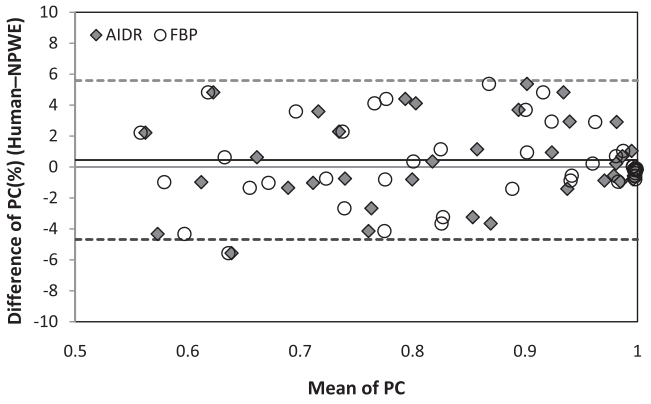
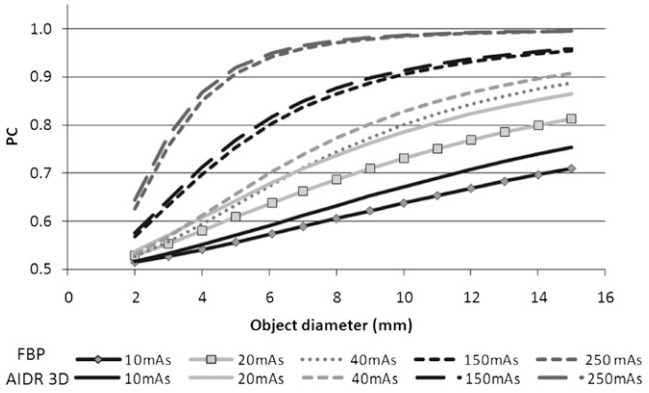


Figure 8. Psychometric fits of proportion correct (PC) as a function of object diameter for non-pre-whitening matched filter with an eye filter and both reconstructions filtered back projection (FBP)/adaptive iterative dose reduction three dimensional (AIDR 3D) algorithm in all the tube current-time products (mAs) range for 0.3% contrast.



Selecting only one threshold value may lead to missing relevant information related to LCDet performance, although it can be helpful as a rough estimate to compare different protocols where dose is changed significantly. As an alternative, the profiles shown in Figure 8 and Bland–Altman plots represent a good tool to study LCDet performance in CT.

In a previous study, a different and smaller set of images reconstructed with AIDR 3D was compared with FBP for the same mA range.[5] The visibility thresholds obtained with NPWE were slightly different then, but it has to be noted that a different psychometric fit was used. In the present work, the selected psychometric curve was a good candidate to be applied to both sets of data (human and NPWE model).[16]

The undesired Bg trends in the images (white band) were suppressed by applying a correction based on the creation of a thick slice image. The transformations applied to the images to correct these trends and to wipe out the nearby objects (to enable taking samples of a reasonable and equal size for the human observer study) did not affect substantially the CNR of the objects owing to the low noise level of the thick slice. Despite being promising, the effect was not of the same order for all the mA sets, and these processes can still be optimized. Other studies opted for a different strategy to perform 2-AFC human observer experiments based on the entire Catphan image and covering the objects that were not being scored by crops taken from the nearby background regions in the image.[30]

The performed human observer study has some limitations. The first one is the reduced number of observers (only four). To obtain a good average of the human LCDet performance for the proposed task, a statistical analysis was performed to remove the inconsistent data. Even so, the study was quite complex to carry out, owing to the high number of images and conditions analysed, although it was restricted only to the 1% contrast group.

The results shown in this work are based on geometrical phantom images, which were modified (Bg correction), and its

conclusions have to be taken cautiously and cannot be extrapolated directly to patient images. Model observers can be helpful tools to analyse image quality in an objective and fast way and to compare different CT scanners, protocols or reconstruction algorithms in terms of image quality. The increasing complexity and variety in the available CT protocols and reconstruction algorithms makes the development of these automated methods even more necessary.

CONCLUSIONS

The LCDet performance of human and a model observer (NPWE) has been compared in this study analysing phantom images reconstructed with AIDR 3D and FBP algorithms and a range of mAs. The A 2-AFC study was carried out to estimate the average human observer performance for an SKE/BKE task. The NPWE model was more efficient than the average human ($\eta = 0.44$) and showed an excellent agreement after the correction by the efficiency factor. Other alternatives to match the model observer results in order to reproduce the human observer performance are based on internal noise, which will be explored in the near future. The iterative algorithm (AIDR 3D) showed an overall improvement in LCDet, especially for low mAs and low-contrast objects. The methodology that we have developed for the human study can be used to perform analysis with different types of medical images, not necessarily CT. The proposed method can be adapted to other phantoms and other model observers will be implemented to assess image quality in an objective way. Applying the model observer to more realistic diagnostic images based on anthropomorphic phantoms or real patients will be one of the future applications to investigate.

FUNDING

The authors would like to warmly thank the Medical Imaging Perception Society (MIPS) for both scholarships that enabled the first author to attend the XIV and XV MIPS conferences. We would also like to acknowledge the Spanish Medical Physics Society (SEFM) for their funding and support under its international grant program. One of the authors RMSJ held a research grant from Toshiba Medical Systems.

ACKNOWLEDGMENTS

We would like to thank Maria Cros and Ramon Casanovas, from the Unitat de Física Mèdica at the Universitat Rovira i Virgili for their help in scoring the images in the human observer study.

REFERENCES

1. Brenner DJ, Hall EJ. Computed tomography–an increasing source of radiation exposure. *N Engl J Med* 2007; **357**: 2277–84.
2. Berrington de González A, Mahesh M, Kim KP, Bhargavan M, Lewis R, Mettler F, et al. Projected cancer risks from computed tomographic scans performed in the United States in 2007. *Arch Intern Med* 2009; **169**: 2071–7.
3. Bittencourt MS, Schmidt B, Seltmann M, Muschiol G, Ropers D, Daniel WG, et al. Iterative reconstruction in image space (IRIS) in cardiac computed tomography: initial experience. *Int J Cardiovasc Imaging* 2011; **27**: 1081–7. doi: 10.1007/s10554-010-9756-3
4. Leipsic J, Labounty TM, Heilbron B, Min JK, Mancini GB, Lin FY, et al. Adaptive statistical iterative reconstruction: assessment of image noise and image quality in coronary CT angiography. *AJR Am J Roentgenol* 2010; **195**: 649–54. doi: 10.2214/AJR.10.4285
5. Joemai RM, Veldkamp WJ, Kroft LJ, Hernandez-Giron I, Geleijns J. Adaptive iterative dose reduction 3D versus filtered back projection in CT: evaluation of image quality. *AJR Am J Roentgenol* 2013; **201**: 1291–7. doi: 10.2214/AJR.12.9780
6. Beister M, Kolditz D, Kalender WA. Iterative reconstruction methods in X-ray CT. *Phys Med* 2012; **28**: 94–108.
7. Imaging Performance and Assessment of CT scanners. 32 to 64 slice CT scanner comparison report version 14. ImPACT. Report 06013. NHS Purchasing and Supply Agency. NHS PASA, 2005.
8. International Commission on Radiation Units and Measurements. Receiver operating characteristic analysis in medical imaging. ICRU Report No. 79. Bethesda, MD: International Commission on Radiation Units and Measurements; 2008.
9. Ogden K, Huda W. Applications of AFC methodology in optimization of CT systems. In: Samei E, Krupinski E, eds. *Medical image perception and techniques*. New York, NY: Cambridge University Press; 2010. pp. 356–63.
10. Chesters MS. Human visual perception and ROC methodology in medical imaging. *Phys Med Biol* 1992; **37**: 1433–76.
11. Klein Zeggelink WF, Hart AA, Gilhuijs KG. Assessment of analysis-of-variance-based methods to quantify the random variations of observers in medical imaging measurements: guidelines to the investigator. *Med Phys* 2004; **31**: 1996–2007.
12. 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
13. Popescu LM, Myers KJ. CT image assessment by low contrast signal detectability evaluation with unknown signal location. *Med Phys* 2013; **40**: 111908. doi: 10.1118/1.4824055
14. 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
15. Eckstein MP, Abbey CK, Bochud FO. A practical guide to model observers for visual detection in synthetic and natural noisy images. In: Van Metter RL, Beutel J, Kundel HL, eds. *Handbook of medical imaging. Physics and psychophysics*. Vol. **1**. Bellingham, WA: SPIE-The International Society for Optical Engineering; 2000. pp. 595–628.
16. Hernandez-Giron I, Geleijns J, Calzado A, Veldkamp WJ. Automated assessment of low contrast sensitivity for CT systems using a model observer. *Med Phys* 2011; **38**: S25–35. doi: 10.1118/1.3577757
17. Mori S, Endo M, Nishizawa K, Murase K, Fujiwara H, Tanada S. Comparison of patient doses in 256-slice CT and 16-slice CT scanners. *Br J Radiol* 2006; **79**: 56–61. doi: 10.1259/bjr/39775216
18. Reiser I, Nishikawa RM. Identification of simulated microcalcifications in white noise and mammographic backgrounds. *Med Phys* 2006; **33**: 2905–11.
19. Burgess AE. Prewhitening revisited. In: Kundel HL, ed. *Proceedings of SPIE 3340,*

medical imaging 1998: image perception, 55; 21 April 1998; San Diego, CA.
20. Burgess AE, Jacobson FL, Judy PF. Human observer detection experiments with mammograms and power-law noise. *Med Phys* 2001; **28**: 419–37.
21. 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.
22. Woolson RF. Comparison of two groups: t-tests and rank tests. In: *Statistical methods of analysis of biomedical data*. New York, NY: Wiley; 1987. pp. 172–87.
23. MacMillan NA, Creelman CD. Comparison (two-distribuion) designs for discrimination. In: *Detection theory: a user's guide*. Mahwah, NJ: Lawrence Eribaum Associates; 2005. pp. 165–85.
24. Prins N, Kingdom FAA. (2009) Palamedes: matlab routines for analysing psychophysical data. [Cited 27 May 2014] Available from: http://www.palamedestoolbox.org
25. Hervada Vidal X, Santiago Pérez MI, Vázquez Fernández E, Castillo Salgado C, Epidat 3.0: Programme for epidemiological analysis of tabulated data. *Rev Esp Salud Pública* 2004; **78**: 277−80. [Cited 27 May 2014] Available from: http://www.sergas.es/MostrarContidos_N3_T01.aspx?IdPaxina=62714
26. Karssemeijer N, Thijssen MAO. Determination of contrast-detail curves of mammography systems by automated image analysis. In: Doi K, Giger ML, Nishikawa RM, Schmidt RA, eds. *Digital mammography*. Amsterdam, Netherlands: Elsevier; 1996. pp. 155–60.
27. Klein SA. Measuring, estimating, and understanding the phsychometric function: a commentary. *Percept Phsychophys* 2001; **63**: 1421–55.
28. Kingdom FAA, Prins N. Psychophysics: a practical introduction. London, UK: Academic press imprint of Elsevier; 2010.
29. Tapiovaara MJ. Efficiency of low-contrast detail detectability in fluoroscopic imaging. *Med Phys* 1997; **24**: 655–64.
30. Fan J, Madhav P, Sainath P, Cao X, Wu H, Nilsen R, et al. Evaluation of low contrast detectability performance using the two-alternative forced choice method on computed tomography dose reduction algorithms. In: Abbey CK, Mello-Thoms CR, eds. *Proceedings of the 2012 SPIE medical imaging: image perception, observer performance, and technology assessment*. San Diego, CA: Proc. SPIE 8318, 2012. 86731F1–7.