Toward a comparison and an optimization of CT protocols using new metrics of dose and image quality part I: prediction of human observers using a model observer for detection and discrimination tasks in low-dose CT images in various scanning conditions

Nadia Othman, Anne-Catherine Simon, Thierry Montagu, Laureline Berteloot, David Grévent, Bouchra Habib Geryes, Mohamed Benkreira, Emeline Bigand, Sophie Capdeville, Julie Desrousseaux, et al.

Toward a comparison and an optimization of CT protocols using new metrics of dose and image quality Part I: Prediction of human observers using a model observer for detection and discrimination tasks in low-dose CT images in various scanning conditions

Nadia Othman[1], Anne-Catherine Simon[1], Thierry Montagu[1], Laureline Berteloot[2],David Grévent[2], Bouchra Habib Geryes[2], Mohamed Benkreira[3], Emeline Bigand[3], Sophie Capdeville[3], Julie Desrousseaux[3], Bardia Farman[3], Eloise Garnier[3], Stephanie Gempp[3],Jean-Marc Nigoul[3], Natacha Nomikossoff[3], and Marion Vincent[3]

[1] Université Paris-Saclay, CEA, List, F-91120 Palaiseau, France
[2] Necker-Enfants Malades University Hospital, Paediatric Radiology Department, Paris, France
[3] AP-HM, Marseille, France

E-mail: nadia.othmane@gmail.com

Received xxxxxx
Accepted for publication xxxxxx
Published xxxxxx

Abstract

In the context of reducing the patient dose coming from CT scanner (Computed Tomography) examinations without penalizing the diagnosis, the assessment of both patient dose and image quality (IQ) with relevant metrics is crucial. The present study represents the first stage in a larger work, aiming to compare and optimize CT protocols using dose and IQ new metrics. We proposed here to evaluate the capacity of the Non-PreWhitening matched filter with an eye (NPWE) model observer to be a robust and accurate estimation of IQ.

We focused our work on two types of clinical tasks: a low contrast detection task and a discrimination task. We designed a torso-shaped phantom, including Plastic Water® slabs with cylindrical inserts of different diameters, sections and compositions. We led a human observer study with 13 human observers on images acquired in multiple irradiation and reconstruction scanning conditions (voltage, pitch, slice thickness, noise level of the reconstruction algorithm, energy level in dual-energy mode and dose), to evaluate the behavior of the model observer compared to the human responses faced to changing conditions. The model observer presented the same trends as the human observers with generally better results. We rescaled the NPWE model on the human responses by scanning conditions (kVp, pitch, slice thickness) to obtain the best agreement between both observer types, estimated using the Bland-Altman method.

The impact of some scanning parameters was estimated using the correct answer rate given by the rescaled NPWE model, for both tasks and each insert size. In particular, the comparison between the dual-energy mode at 74 keV and the single-energy mode at 120 kVp showed that, if the 120 kVp voltage provided better results for the smallest insert at the lower doses for both tasks, their responses were equivalent in many cases.

Keywords: Model Observer, Diagnostic Imaging, Image Quality, Computed Tomography, Low Dose Level, Dual-Energy

1. Introduction

The number of medical imaging exams yearly performed has greatly increased for many years with the development of imaging technologies and the ageing of the population. In particular, although Computed Tomography (CT) examinations only correspond to about 10 % of medical imaging procedures, they are credited with about two thirds of the total imaging collective dose [1][2]. Reducing the dose due to CT examinations is therefore a major issue. However, dose reduction should not be undertaken without considering the clinical objective and the associated imaging tasks.

If relevant assessment of image quality (IQ) is crucial to ensure correct diagnosis while maintaining patient dose at the lowest as possible, despite the wide use of CT scanners in diagnosis, evaluation of their performances in terms of IQ/dose remains poorly documented. In particular, the medical community is confronted with a lack of robust indicators of IQ. Indeed, commonly traditional physical metrics measurements such as the Contrast-to-Noise Ratio (CNR) or the Modulation Transfer Function (MTF) are nowadays unsuitable for evaluating the IQ in diagnostic radiology, especially with the development of iterative reconstruction algorithms (IR) [3][4][5][6][7][8][9][10][11]. Indeed, the step of regularization in IR methods introduces different levels of non-linearity into the imaging systems. More importantly, such IQ metrics do not take into account a complete description of the IQ as they do not include some information about the diagnostic accuracy of a given clinical task, which represents the ultimate purpose of the image's acquisition.

Therefore, task-based image quality metrics have been developed these last decades, linked to a given clinical task such as the detection of a pathology or the discrimination between several types of lesions [12]. To measure the performance of such metrics, Receiver Operating Curve (ROC) studies including human experts (e.g. radiologists) are usually conducted [13] to evaluate the diagnostic accuracy of a given task. However, such evaluations are onerous and time-consuming [14]. Consequently, model observer approaches have been developed and applied to medical images [15][16], based on the decision theory. The main types of observers are the ideal observer (Bayesian) [17], the non-prewhitening matched filter (NPW) proposed by Wagner in [18] and its modified version with an eye filter called NPWE (later suggested by Burgess in [19]) and the Hotelling Observer and its improving version: Channelized Hotelling Observers (CHO) proposed by Barrett [13].

In the recent state-of-the-art, several studies have shown that NPWE models [6][7][10][20][21][22][23] and CHO models [7][24][25][26][27][28][29][30] among others, are highly correlated with human performance for a specific task such as the detection of a lesion, its localization, or the discrimination between benign and malignant lesions in CT images. In fact, several researchers have investigated human performance in detecting signals in CT images taken from anthropomorphic phantoms with embedded low-contrast objects [7][23][24], customized phantoms with different inserts [28][31], or control quality phantoms with several objects of different sizes and contrasts [20][21]. In contrast, fewer studies have been dedicated for rating human performance in the case of discriminating two signals of different shapes [26] or different textures [32] in CT images acquired from customized phantoms. In a recent work, the localization of liver lesions, digitally inserted in CT images, was also studied [30]. In all these works, the influence of only one or two acquisition and reconstruction parameters on IQ has been investigated. Among these parameters, the reconstruction algorithm has been probably the most described, when other features, such as dual-energy mode (also called spectral mode), have never been studied in terms of clinical task.

The present study represents the first stage in a larger work that aims to objectively compare performances of new modes of scanning protocols, such as dual-energy mode, in terms of IQ/dose, and more generally, to pave the way for a standardized method to compare and optimize protocols in clinics. For that purpose, we propose to base our assessment of protocol performances on two metrics simultaneously: an estimation of IQ and an estimation of patient dose. The present paper deals with the use of a model observer as an estimation of IQ, with human observers as a reference. In this context, the question of the accuracy of the model observer, together with its robustness according to different scanning conditions, is raised, the goal being that the model observer matches the humans in all cases. Because of its reported capacity to reproduce human observer skills [20], we decided to use the Non Pre-Whitening Eye filter model (NPWE) in our study.

Two types of diagnostic tasks were addressed here: detection of small lesions with low contrast, and benign and malignant lesions discrimination in CT images of a customized phantom. A human observer study including a wide group of experimental observers was conducted to validate the model observer, by showing correlations between the human and model observers' responses for several cases of lesions and protocols. To show the model trend and performance in various situations, a large set of scanning and reconstruction parameters such as single and dual energy modes, slice thickness, IR algorithm, and helical pitches were explored in order to cover most as possible a wide number of scenarios at different doses and IQ levels. Some results were presented in an oral session of the World Congress on Medical Physics and Biomedical Engineering held in Prague, Czech Republic (June 2018), which was focused on how to improve diagnostic imaging by IQ measure.

2. Materials and methods

2.1 Data acquisition and reconstruction

For the purpose of the study, a customized phantom was designed as a torso-shaped phantom of size 26 x 35 x 30 cm³, filled by water, in order to simulate the average radiation attenuation of a standard size abdomen of an adult patient. In the middle of the phantom two slabs of size 20 x 20 x 6 cm³ could be inserted as illustrated in **Figure 1**.

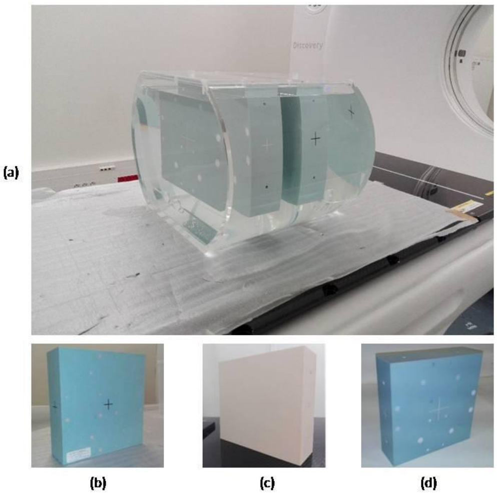


Figure 1: A torso shaped water phantom. (a) the tank filled by water, (b) the detection slab, (c) slab with no inserts to generate background images (d) the discrimination slab.

As explained before, two types of diagnostic tasks were addressed in this study: detection of small lesions with low contrast, and the discrimination between benign and malignant lesions. For that purpose, we designed several slabs, manufactured by CIRS company. In the case of detection task, two Plastic Water® LR (Low Energy Range) slabs were used. The first one contained 4 series of 4 cylindrical inserts each, made of epoxy resin materials, of diameters 2.5, 3.5, 5 and 7 mm, with contrast about 30 Hounsfield units (HU) at 120 kVp between the inserts and the background. The second slab was free of inserts, in order to generate the uniform background images (no-lesion). For the discrimination task, only one Plastic Water® slab was used. It contained 16 Teflon® (900 HU) and polystyrene (-25 HU) cylindrical inserts, with circular or hexagonal cross-section shapes in order to simulate respectively benign and malignant lesions (irregular boundaries). The diameters of the circular sections were 6.35 mm and 12.7 mm and the size of the hexagonal inserts were fixed to have the same cross-section area as the circular inserts in order to obtain the same signal power. All the inserts of the slabs were parallel to the z-axis of the scanner, and the distances between the different locations of the rods in the axial plan (x,y) were maximized much as possible (at least 4 cm) to avoid potential interference between the inserts in the reconstruction of the signal images.

The images of the phantom were acquired with a GE Discovery CT750 HD scanner, available at our research platform. The scan parameters were based on a clinical abdominal CT protocol available on the scanner, that we modified by varying several scanning and reconstruction parameters that influence IQ and dose level such as the tube voltage or the slice thickness. More precisely, the study covered all the available kilo-voltages on the system i.e. 80, 100, 120, and 140 kVp in the single-energy mode. The dual-energy mode (called GSI on the GE CT scanner) was also used for several acquisitions. In addition to the standard value of the helical pitch (1.375), two other values (0.516 and 0.984) were investigated. The detector collimation was kept to 40 mm and the used scan field of view (SFOV) was the "Large Body". The rotation time was set to 1 second. All the resulting images were reconstructed with the Adaptive Statistical Iterative Reconstructed algorithm (ASIR) using the "STD Kernel", at two levels of noise, with the standard slice-thickness (1.25 mm) and another higher value (5 mm).

The two ASIR levels were 30 %, used for osteo-articular examinations, and 70 %, used for digestive tract examinations. The images acquired under the GSI mode were reconstructed at two energies, 60 keV and 74 keV. The first energy level was the standard value proposed in the GSI abdomen protocol. The second one was obtained by using the energy search functionality of the GE scanner, which indicates the energy level that gives the optimal Contrast-to-Noise Ratio (CNR) between the insert and the background.

For each combination of those parameters, we changed the tube current (mA) to obtain 5 different CT dose indices (CTDIvol), given by the scanner. A wide range of CTDIvol values were investigated, from very low dose values to higher ones. In the case of the single-energy operating mode, the CTDIvol were fixed to: ~1, ~5, ~10, ~15 and 20 mGy for all the kilo-voltages (except for 80 kVp: ~1, ~3, ~5, ~7 and 10 mGy, because 10 mGy was the maximum reachable value with the studied irradiation parameters). In the case of the dual-energy operating mode, we tried to achieve similar CTDIvol values with the same investigated parameters but it was impossible because the tube current and the rotation speed cannot be modified by the user in this mode. Therefore, the CTDIvol for this mode were set to: ~6.5, ~7.6, ~10.7, ~15.6, and ~20.7 mGy.

We repeatedly scanned the phantom: first, for each set of parameters, the phantom was scanned three times with the detection and discrimination slabs in order to obtain the signals images (images with inserts). In a second time, to obtain the background images, only the slab free of inserts was inserted in the phantom in the location of the previous scanned detection slab (same x, y and z position), and the whole phantom was scanned again three times in the same conditions.

2.2 Data preparation for the human observer study

Before the extraction of the images for the human observer study, for each set the scanning parameters, and at all dose levels, a visual inspection of several images was carried out, leading to the following observations:

- The smallest inserts of the detection slab (2.5 mm) were visible only at very high CTDIvol (20 mGy) whatever of the scanning parameters.
- The visibility of the insert of size 3.5 mm in the reconstructed images with a slice thickness of 1.25 mm was roughly limited regardless of the different scanning parameters and the dose levels.
- It was impossible to distinguish between the circular and hexagonal inserts of the discrimination slab, made of polystyrene, in the majority of the explored scanning and reconstruction parameters.
- We also noticed that the discrimination between the hexagonal and circular Teflon® inserts in the images reconstructed with a thickness of 5 mm was too easy, even at very low dose.

These observations led us to discard all the cases described above from the study.

As explained before, various scanning and reconstruction parameters were involved in this study, leading to a high number of possible combinations. In order to study the influence of specific parameters on the detectability at several dose levels, we decided to keep constant the other ones to better analyze the results. Table 1 summarizes the various conditions (scanning and reconstruction parameters, doses, and lesion profiles) investigated for the detection task, in the single and dual source modes. At final, four sets of experiments were conducted for this task: kVp, pitch, slice thickness and GSI experiments. The studied conditions for the discrimination task, in the single and dual source modes, are given in Table 2. As shown, three sets of experiments were conducted: kVp, pitch, and GSI experiments.

Table 1: Summary of the scanning and reconstruction parameters, and doses for the detection of lesions of diameters 3.5, 5 and 7 mm.

	Tube voltage (kVp)	Pitch	Slice thickness (mm)	ASIR (%)		CTDIvol (mGy)				
kVp experiments	80	1.375	5	30	70	~1	~3	~5	~7	~10
	100	1.375	5	30	70	~1	~5	~10	~15	~20
	120	1.375	5	30	70	~1	~5	~10	~15	~20
	140	1.375	5	30	70	~1	~5	~10	~15	~20
GSI experiments	GSI - 60 keV	1.375	5	70		~6.5	~7.6	~10.7	~15.6	~20.7
	GSI - 74 keV	1.375	5	70		~6.5	~7.6	~10.7	~15.6	~20.7
Thickness experiments*	120	1.375	1.25	30	70	~1	~5	~10	~15	~20
	120	1.375	5	30	70	~1	~5	~10	~15	~20
Pitch	120	0.516	5	30	70	~1	~5	~10	~15	~20

experiments	120	0.984	5	30	70	~1	~5	~10	~15	~20
	120	1.375	5	30	70	~1	~5	~10	~15	~20

*Only for the lesions of diameters 5 and 7 mm.

Table 2: Summary of the scanning and reconstruction parameters, and doses for the discrimination of Teflon lesions of size 6.35 and 12.7 mm.

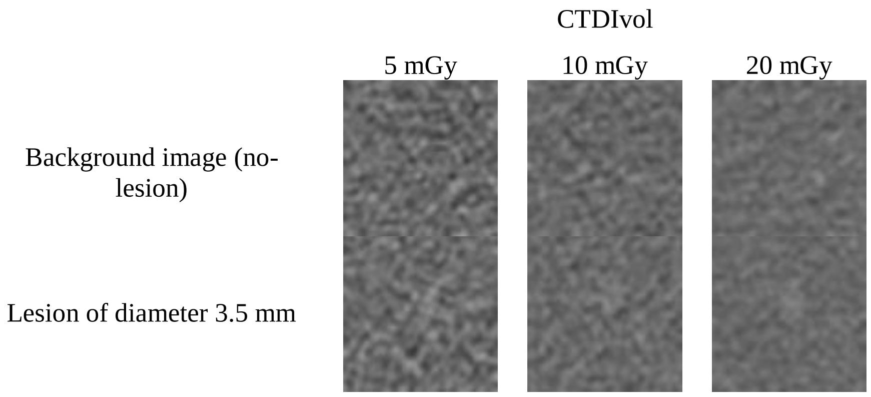
	Tube voltage (kVp)	Pitch	Slice thickness (mm)	ASIR (%)		CTDIvol (mGy)				
kVp experiments	80	1.375	1.25	30	70	~1	~3	~5	~7	~10
	100	1.375	1.25	30	70	~1	~5	~10	~15	~20
	120	1.375	1.25	30	70	~1	~5	~10	~15	~20
	140	1.375	1.25	30	70	~1	~5	~10	~15	~20
GSI experiments	GSI - 60 keV	1.375	1.25	30		~6.5	~7.6	~10.7	~15.6	~20.7
	GSI - 74 keV	1.375	1.25	30		~6.5	~7.6	~10.7	~15.6	~20.7
Pitch experiments	120	0.516	1.25	30	70	~1	~5	~10	~15	~20
	120	0.984	1.25	30	70	~1	~5	~10	~15	~20
	120	1.375	1.25	30	70	~1	~5	~10	~15	~20

For each given condition (scanning and reconstruction parameters + lesion profile), several images were extracted from the repeated scans in order to have a large number of samples for the statistical analysis of the detectability. Since the phantom was scanned at first with the detection and discrimination slabs, the images used for the detection and discrimination experiments were extracted from the same scans at the same time. As the inserts of the two slabs were carefully aligned along the x and y axis, and parallel along the z axis through the acquisition, the location of all the inserts was fully known. Regions of interest (ROIs) of 24 x 24 mm² were automatically extracted around each insert thanks to a program written in Python language. In fact, since the position of the phantom did not change during the acquisition campaign, the extraction of the ROIs was achieved according to a mask template, with the pre-established inserts coordinates based on the slabs plans and phantom position. These coordinates were visually verified at high doses to avoid possible mistakes. All the ROIs had the same size regardless to the size, shape and material of the insert, and the signal was always located in the center of the image. Only ROIs coming from successive slices inside the slabs were exploited in order to prevent from eventual reconstruction artifacts coming from the transition between the Plastic Water and the water.

At final, 48 signal images, coming from repeated scans and adjacent slices, were used for each set of the scanning parameters and lesion profile (size/shape/material). More precisely, 4 central successive slices, with 4 identical lesions per slice, from 3 repeated scans, were used for the detection task, whereas for the discrimination task, the 48 trials came from 8 central successive slices, with two identical lesions per slice, from 3 repeated scans.

Background images were also obtained from the free-insert slab at the scanning and reconstruction parameters used in the detection experiments. The ROIs of same size (24 x 24 mm²) were extracted at the same coordinates x, y and z, previously used for the detection slab, resulting in 48 images per condition as for the detection images.

At final, all these signal and background images were given to the human observers for the human observer study. Some examples of these images for the detection and discrimination tasks are respectively shown in Figure 2 and Figure 3.



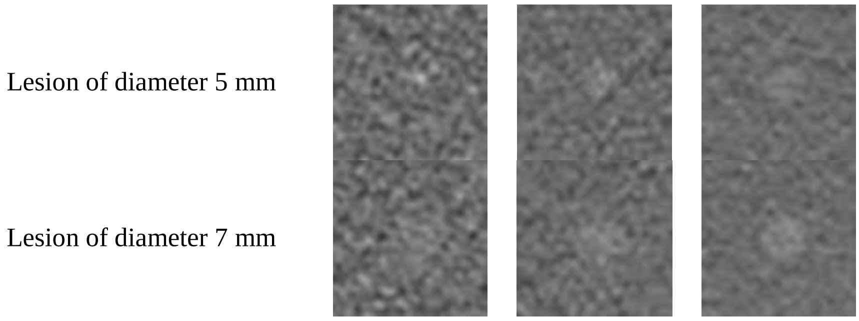


Figure 2: Regions of interest for the detection task at different dose levels. The images were acquired at 120 kVp, with a pitch of 1.375, and reconstructed with ASIR 30 % with a slice thickness of 5 mm, at three different CTDIvols: 5, 10 and 20 mGy.

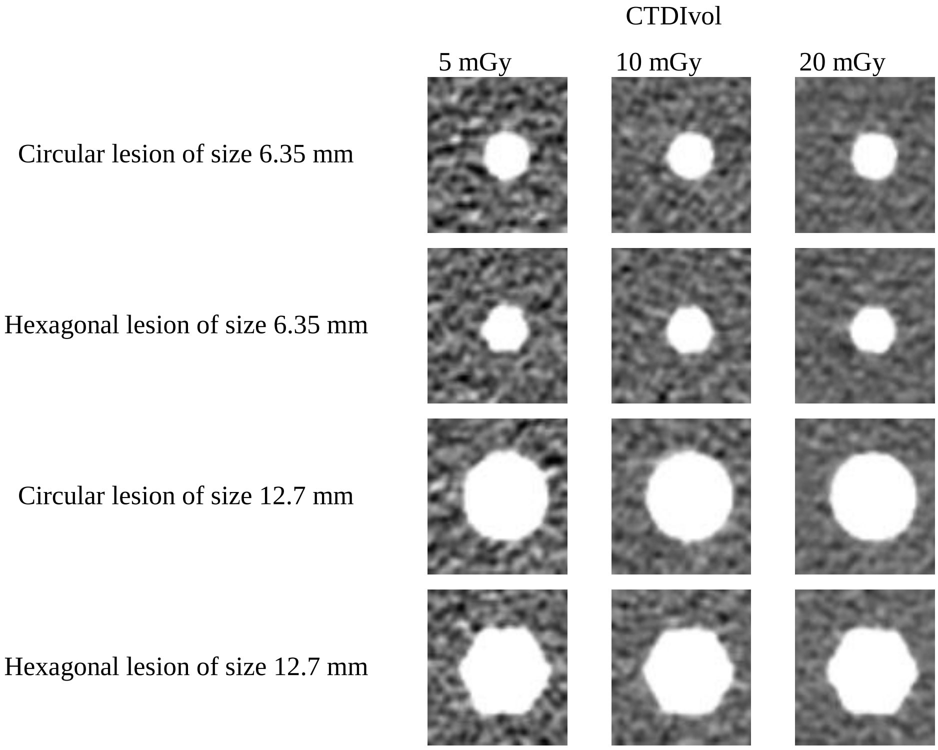

CTDIvol
5 mGy
10 mGy
20 mGy
Circular lesion of size 6.35 mm
Hexagonal lesion of size 6.35 mm
Circular lesion of size 12.7 mm
Hexagonal lesion of size 12.7 mm

Figure 3: Regions of interest for the discrimination task at different dose levels. The images were acquired at 120 kVp, with a pitch of 1.375, and reconstructed with ASIR 30 % with a slice thickness of 1.25 mm, at three different CTDIvols: 5, 10 and 20 mGy.

2.3 Human observer study

A large number of experimented human observers were recruited from two clinical institutes in France. In total, two senior radiologists and eleven medical physicists were asked to perform both clinical tasks. The study was based on two alternative forced choice (2-AFC) experiments. Series of couples of images were presented side by side to the observers in a randomized order. In the case of the detection task, the couple was composed of a signal-image and a background image (absence of signal). The observers were independently asked to choose the image that contained the lesion, i.e. the signal image. In contrast, in the discrimination task, the couple of images included one image with a hexagonal signal and one image with a circular signal. This time, the observer was asked to select the malignant lesion, i.e. to identify the image with the hexagonal signal.

Two Graphical User Interfaces (GUI), written in Python language, were developed to separately achieve the 2-AFC tests for the detection and discrimination tasks. When a program was launched, a couple of images was randomly displayed side by side, without any information on the acquisition parameters or on the lesion profile to be identified. The GUI allowed the observer to select the image of interest (i.e. the signal image in the detection case, or the hexagonal signal image in the discrimination case) by clicking on it with the mouse. The observer could change its decision until the validation of the image's selection. No time limit was imposed on the observer to take the decision. Figure 4(a) and Figure 4(b) show respectively an example of a 2-AFC test for the detection and the discrimination tasks.

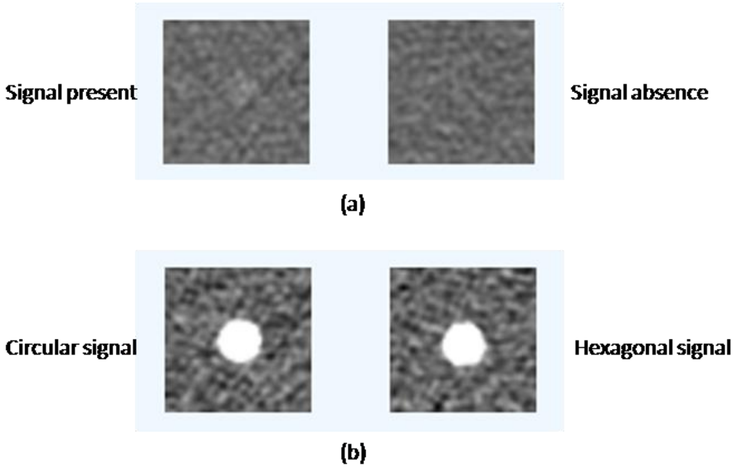


Figure 4: Examples of 2-AFC tests at 120 kVp and 10 mGy with a pitch of 1.375, using ASIR 30 % reconstruction: (a) Lesion detection of diameter 5 mm, (b) Lesions discrimination of size 6.35 mm.

All the observers were asked to perform the AFC tests in a dark room. The viewing distance between the reader and the monitor was set to approximately 40 cm. All the images were displayed with a window width of 400 HU and a window center of 40 HU, as recommended in our standard abdomen protocol. Sessions of 2 hours at maximum were planned to avoid fatigue.

Before the study, training sessions with all the observers were organized in the two hospitals to familiarize the observers with the two tasks and the GUI. The training set was only composed of high-quality images of the lesions under investigation, and they were not included in the 2-AFC tests.

In total, there were respectively 230 and 140 categories of 2-AFC experiments for the detection and discrimination tasks (for more details, see **Table 1** and **Table 2**), and each 2-AFC was composed of 48 trials, leading to 17760 decisions, which would have been very tedious and time consuming for the observers. Since we had a large number of observers for the study (13 observers), we decided to split the 2-AFC tests into subparts and we made four groups of human observers: three groups of 3 people and one group of 4. Each person of a given group was asked to undergo the same 2-AFC tests. Each group had to review images with respect to the variation of one or several given parameters, such as slice thickness or pitch. All groups had the same subpart relative to the reference standard protocol. In order to have about the same number of decisions for each group, some subparts were given to several groups. In consequence, the number of recruited observers differed from one subpart of the 2-AFC test to another.

For each category of the 2-AFC experiments, i.e. for each condition, the percentage of correct (PC) answers across the observers was computed according to the comparison between the truth and their answers. In addition, 95 % confidence intervals were also computed as $CI_{95\%} = 1.96 \cdot STD\ (X)/\sqrt{N_{obs}}$, where N_{obs} denotes the number of observers and X is the detectability accuracy of the observers [7](Gaussian hypothesis).

2.4 Model observer

2.4.1 NPWE. In this work, we used the non-prewhitening matched filter model with an eye (NPWE). The model's concept is to define a "template" that matches exactly the signal under investigation (for example the detection of a signal in a CT image). It also integrates physical measures of the imaging system such as the resolution and the noise of the images. In [33], Burgess has proposed a modified version of the NPW model by adding a front-filter called "eye filter" to the task template, in order to account for the human visual system's sensitivity to different spatial frequencies.

It allows to compute in the Fourier domain a scalar d', usually called detectability index. This index d' reflects the prediction score of a signal's detection in some given conditions. The computation of d' for the NPWE model that we used is as follows, for each set of the scanning and reconstruction parameters and for each lesion size:

$$d'_{NPWE}{}^{2} = \frac{[\iint MTF^{2}{}_{task}(u,v).W^{2}{}_{task}(u,v)E^{2}(u,v)dudv]^{2}}{\iint NPS(u,v)\cdot MTF^{2}{}_{task}(u,v)\cdot W^{2}{}_{task}(u,v)E^{4}(u,v)dudv} \quad (1)$$

with *u* and *v* the spatial frequencies, MTF_{task} the task-based Modulation Transfer Function, *NPS* the Noise Power Spectrum and *E(u,v)* the eye filter. The entity W_{task} represents the task template, which is the representation of the imaging task under investigation in the spatial domain. It is expressed as the Fourier Transform of the difference between two hypotheses *h1* and *h2* as follows [15]:

$$W_{task}(u,v) = |FT[h1(x,y) - h2(x,y)]| \quad (2)$$

with *FT* the Fourier Transform, and *h1* and *h2* the hypothesis functions on the signal description in the spatial domain for the two hypotheses. The different quantities in (1) are supposed to be spatially stationary.

In this study, two clinical tasks with different lesions sizes were investigated, leading to several W_{task}. In the case of the detection task, the considered hypotheses were the following: a uniform background (no signal, h1(x,y)=0) and a circular signal (h2(x,y)=2D projection of a circular cross-section cylinder, corresponding to the lesion profile in the axial plan). For the discrimination case, the hypotheses were as follows: h1(x,y)= 2D projection of an hexagonal cross-section cylinder and h2(x,y)= 2D projection of a circular cross-section cylinder, both corresponding to the lesion profiles under investigation in the axial plan. The projections were blurred with a Gaussian filter as done in [34] in order to add some noise induced by the imaging system. In total, three and two W_{task} were respectively generated for the detection and discrimination tasks, corresponding to the five lesion sizes.

The eye filter E(u,v) represents the contrast sensitivity function (CSF) of the human eye in the spatial domain. The CSF that we used in this study was the same as proposed in [35]:

$$A(f) = 2.6\cdot(0.0192 + 0.114f)\cdot e^{-(0.114f)^{1.1}} \quad (3)$$

where *f* is the spatial frequency (cycle/deg).

The physical measures, i.e. the MTF_{task} and the NPS, were computed for all the acquisition parameters under investigation. Following the conclusions of [9], the MTF_{task} was calculated differently depending on the object contrast. For the high contrast inserts, the MTF_{task} was taken equal to the MTF generally used in the NPWE model. To assess the different MTFs, the CTP528 High-Resolution Module of the CATPHAN® 503 phantom (Phantom Laboratories, New York, USA) was used, following [36]. For the low contrast inserts, the MTF_{task} was considered equal to the TTF (Task Transfer Function) and computed using the method described by [9] and [11] on the acrylic target of the CTP404 module of the CATPHAN® 503 phantom.

To measure the noise in the images, the NPS [37] was evaluated for all the conditions, using the previous images of the slab with no-insert of the torso-shaped phantom [38].

Details on the detectability index calculation are given in Appendix A.

In order to compare the model detectability to the human performance, d'_{NPWE} values were converted to the same metrics previously used in the 2-AFC experiments, i.e. the percentage of correct answers (PC) as follows [19][29] (assuming that d'_{NPWE} follows a Gaussian distribution):

$$PC = \frac{1}{2} + \frac{1}{2}erf\left(\frac{d'_{NPWE}}{2}\right) \quad (5)$$

whith *erf* the Gaussian error function given by the following formula:

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^{\infty} e^{-x^2} dx \tag{6}$$

2.4.2 Rescaled NPWE (rNPWE). In order to better mimic human detectability accuracy [39], model observers are usually modified by adding some internal noise [7][40] or by reproducing the efficiency of the visual detection performance of humans [34][20][21]. In this work, we decided to use the efficiency approach. The PC given by the NPWE model was rescaled thanks to coefficients η and δ to form the PC of the rescaled model rNPWE as follows:

$$PC_{rNPWE} = \eta \cdot PC_{NPWE} + \delta \tag{8}$$

For each lesion size, the coefficients η and δ were determined by a least-squares procedure to fit the human responses (PC_{Human}) acquired in all the scanning and reconstruction conditions. The rescaling of $PC_{NPWE}$was carried out following Eq. 8 according two different ways using data sets differently. In the first case, η and δ were determined using all the points available for an insert size ("Case 1", called rNPWE$_1$). In the second case, coefficients η and δ were established using all points of an insert size for each condition of tube voltage, pitch and slice thickness ("Case 2"), called rNPWE$_2$.

2.5 Agreement between human and model observers

To compare the performance of the model and the human observers (PC) for the detection and discrimination tasks, Bland–Altman test was used [41]. This tool enables to evaluate the degree of agreement between the two observers by exploiting all the PC values obtained in the different conditions. Bland–Altman plots the difference between the two observers as a function of the mean of the two observers as follows:

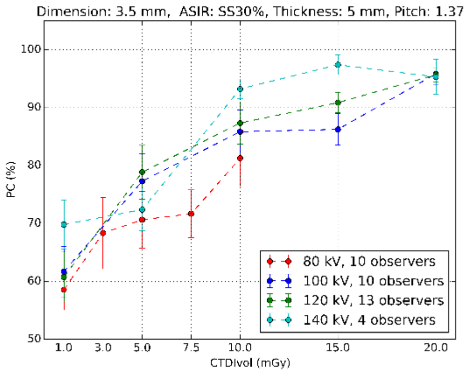
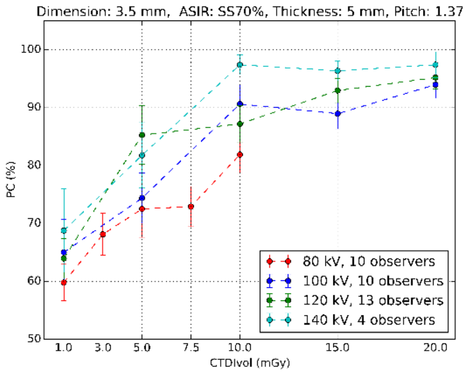
$$PC_{Human} - PC_{rNPWE} = f\left(\left(PC_{Human} + PC_{rNPWE}\right)/2\right) \tag{9}$$

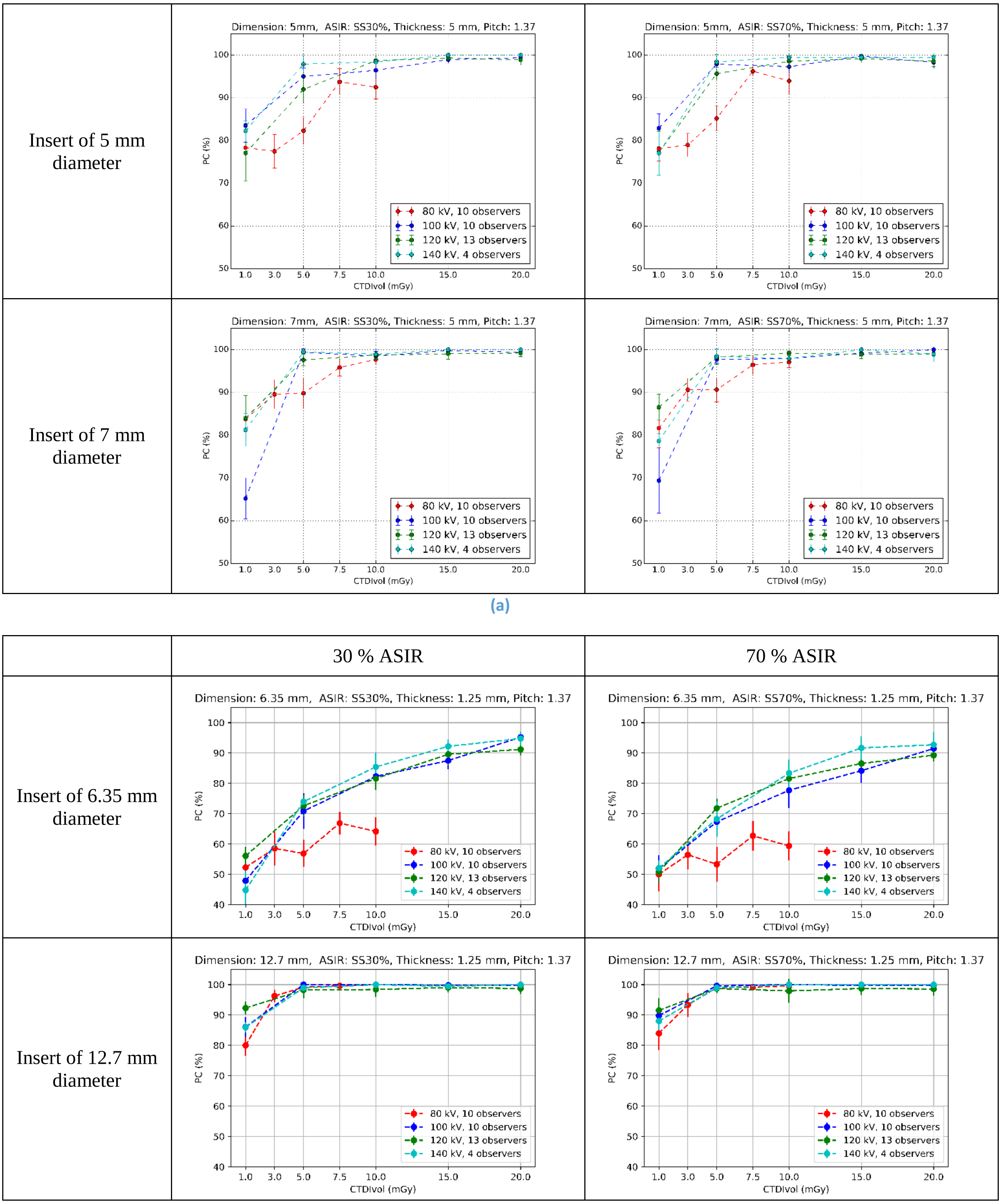
The limits of the observers agreement are fixed by the mean of the difference Δ and the standard deviation of the difference σ as follows: $[\Delta \pm 2\sigma]$.

3. Results

3.1 Human observer study

The correct answer rates for the human observers were calculated for the different values of irradiation and reconstruction parameters of Table 1 and Table 2. Figure 5(a) and Figure 5(b) illustrate the variations of PC_{Human} as a function of CTDIvol for the four tube voltages (80 kVp, 100 kVp, 120 kVp and 140 kVp), respectively for the detection task and the discrimination task, for all insert sizes, and both levels of ASIR.

	30 % ASIR	70 % ASIR
Insert of 3.5 mm diameter		



(a)

(b)

Figure 5: Comparison of percent correct (PC) calculated for the human observers according to the CTDIvol for the different tube voltages, using 30 % and 70 % ASIR settings . (a) Detection task, inserts of 3.5 mm, 5 mm and 7 mm diameter (epoxy resin), (b) discrimination task, inserts of 6.35 mm and 12.7 mm size (Teflon).

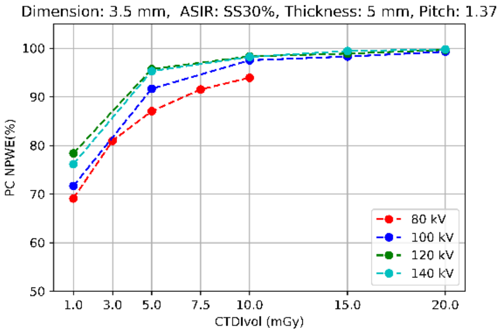
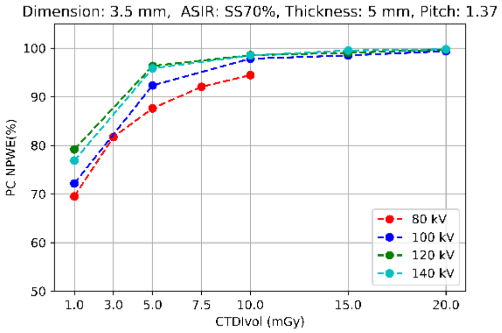
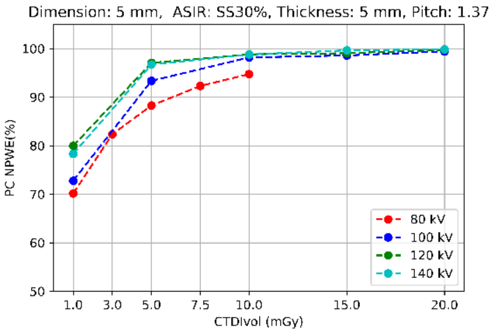
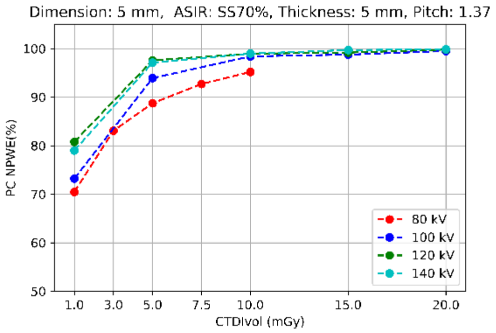
First, it can be noticed that PC_{Human} does not increase with the CTDIvol in a regular manner, due to a high variability in the human responses. The lowest results are generally obtained for the 80 kVp tube voltage, for both tasks, while the results for 100 kVp, 120 kVp and 140 kVp are rather similar. Second, the results given by the human observers are globally better when the insert size increases, except for the point corresponding to 100 kVp and 1 mGy for the 7 mm insert (detection task). Thus, for the largest sizes (7 mm for the detection task and 12.7 mm for the discrimination task), PC is close to 100 % for CTDIvol higher than 5 mGy, except for 80 kVp for the task detection. At the opposite, for the smallest size of the discrimination task, the answers are completely random (~50 %) for 1 mGy. Finally, no particular difference between the two levels of ASIR is observed, although high values (higher than 50 %) of ASIR level are recommended to improve soft tissues contrast (detection task).

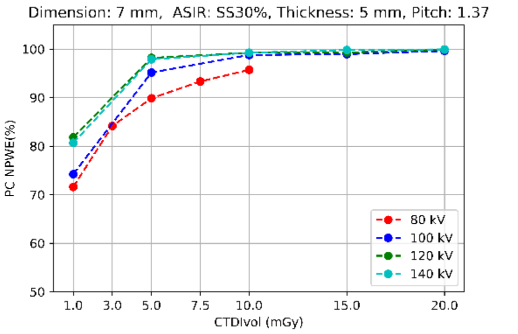
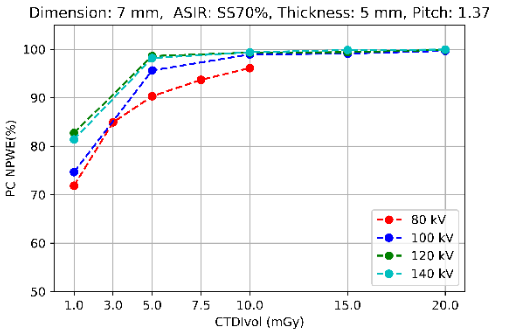
3.2 Model observer before rescaling

The model observer trends are quite similar to the human ones. We can see examples of this in Figure 6(a) and Figure 6(b), which represent the raw PCs calculated by the model observer before rescaling for the same conditions of irradiation and reconstruction as for the human observers, respectively for the detection and the discrimination tasks. However, without the subjective aspect of human answers, variations of PC_{NPWE} according to CTDIvol are more regular. The other differences are:

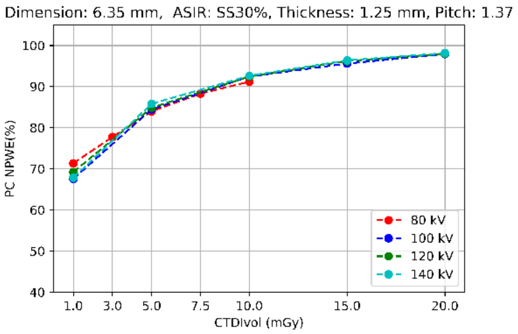
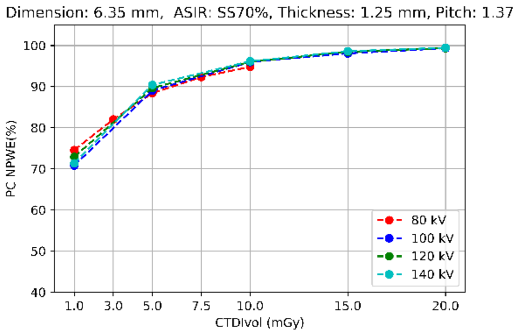
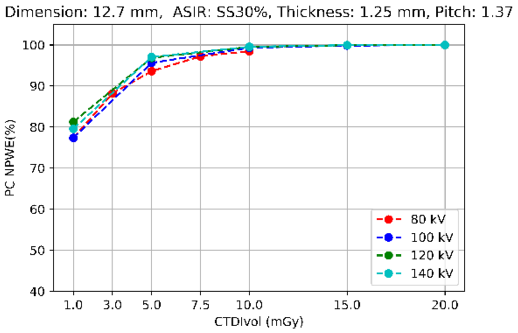
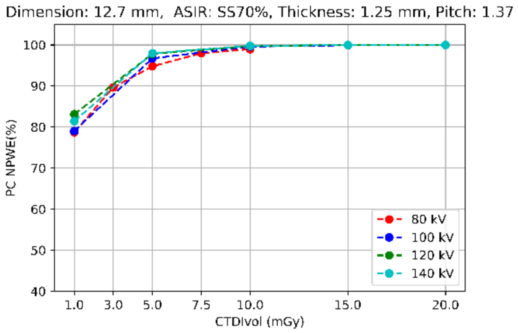
(1) The PC results are generally higher for the NPWE model, particularly in the difficult conditions (small inserts, low doses), except for a few occasional cases. For instance, PC_{NPWE} is about 70 % for the 6.35 mm insert at 1 mGy in the discrimination task, where the humans give a random response (about 50 %).

(2) The response curves for the different tube voltages are quite close to each other, whereas PC_{Human} is in general much lower for 80 kVp.

	30 % ASIR	70 % ASIR
Insert of 3.5 mm diameter		
Insert of 5 mm diameter		

Insert of 7 mm diameter		

(a)

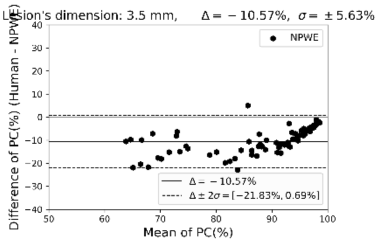
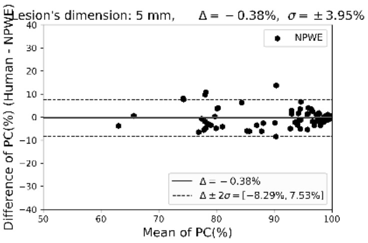
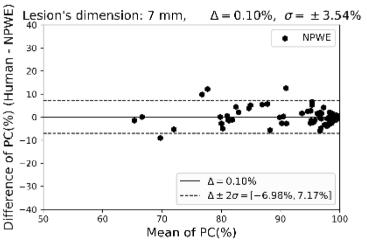
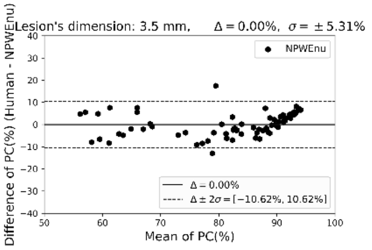
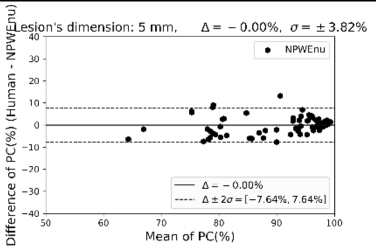
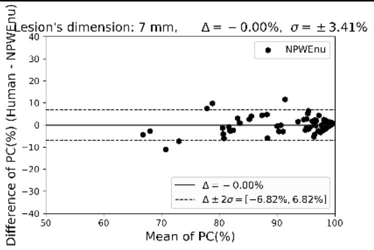
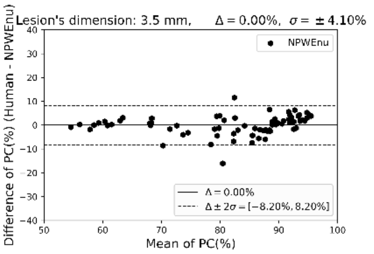
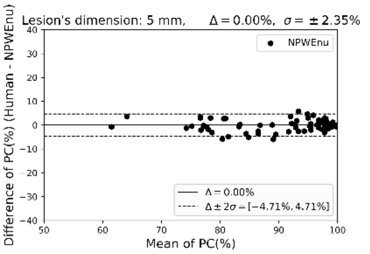
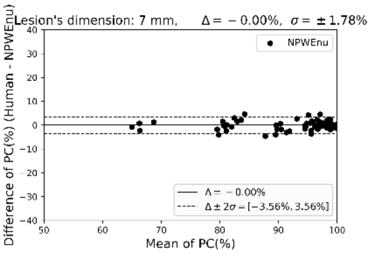
	30 % ASIR	70 % ASIR
Insert of 6.35 mm diameter		
Insert of 12.7 mm diameter		

(b)

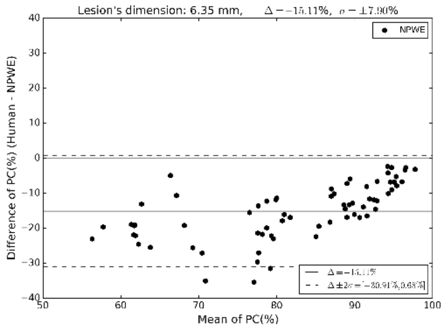
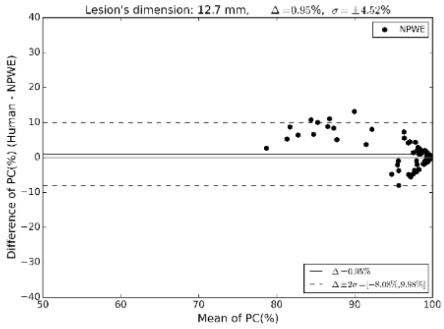
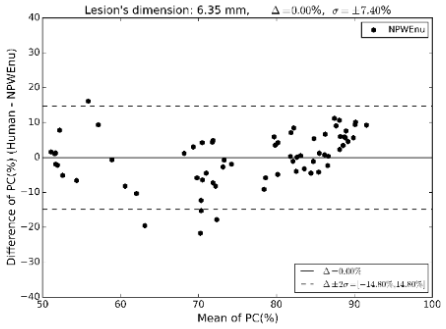
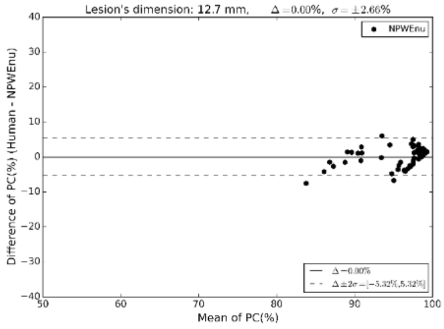
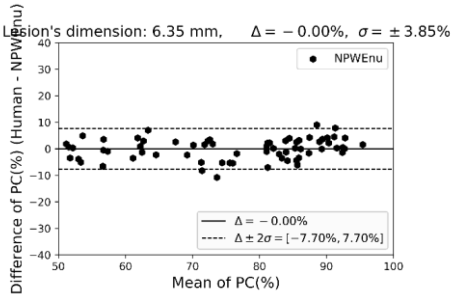
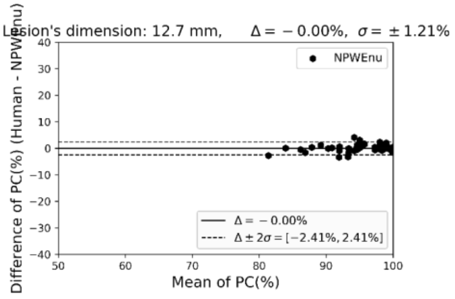
Figure 6: Comparison of percent correct (PC) calculated by the model observer before rescaling according to the CTDIvol for the different tube voltages, using 30 % and 70 % ASIR settings. (a) Detection task, inserts of 3.5 mm, 5 mm and 7 mm diameter (epoxy resin), (b) discrimination task, inserts of 6.35 mm and 12.7 mm size (Teflon).

3.3 Comparison between the humans and the model observer

Bland-Altman method was used to evaluate the degree of agreement between the humans and the model observer, before and after rescaling using both methods. We remind that "Case 1" corresponds to the rescaling by insert size and "Case 2" is for the rescaling by insert size and kVp, pitch and slice thickness conditions. The corresponding Bland-Altman plots are presented Figure 7(a) for the detection task and Figure 7(b) for the discrimination task, with the means of the differences Δ and the standard deviations of the differences σ. Initial NPWE model showed high biases for the smallest inserts, with values of Δ equal to 10.57 % and 15.11 % respectively for detection and discrimination tasks, decreasing to zero after rescaling with both methods. The difference between Case 1-rescaling and Case 2-rescaling essentially appeared in standard deviations σ, about 5 % lower only using Case 1 (except for the 12.7 mm insert, for which σ decreased from 4.52 % to 2.66 %), but improved by around 50 % for Case 2 compared to the initial NPWE model. The σ parameter was thus equal to 1.5 for the largest inserts and about 4 for the smallest ones.

	Insert diameter: 3,5 mm	Insert diameter: 5 mm	Insert diameter: 7 mm
NPWE before rescaling	 $\Delta = -10.57\%, \quad \sigma = \pm 5.63\%$	 $\Delta = -0.38\%, \quad \sigma = \pm 3.95\%$	 $\Delta = 0.10\%, \quad \sigma = \pm 3.54\%$
rNPWE$_1$	 $\Delta = 0.00\%, \quad \sigma = \pm 5.31\%$	 $\Delta = 0.00\%, \quad \sigma = \pm 3.82\%$	 $\Delta = 0.00\%, \quad \sigma = \pm 3.41\%$
rNPWE$_2$	 $\Delta = 0.00\%, \quad \sigma = \pm 4.10\%$	 $\Delta = 0.00\%, \quad \sigma = \pm 2.35\%$	 $\Delta = 0.00\%, \quad \sigma = \pm 1.78\%$

(a)

	Insert size:6.35 mm	Insert size:12.7 mm
NPWE before rescaling	 $\Delta = -15.11\%, \quad \sigma = \pm 7.90\%$	 $\Delta = -0.95\%, \quad \sigma = \pm 4.52\%$
rNPWE₁	 $\Delta = 0.00\%, \quad \sigma = \pm 7.40\%$	 $\Delta = 0.00\%, \quad \sigma = \pm 2.66\%$
rNPWE₂	 $\Delta = 0.00\%, \quad \sigma = \pm 3.85\%$	 $\Delta = 0.00\%, \quad \sigma = \pm 1.21\%$

(b)

Figure 7: Bland-Altman plots corresponding to the NPWE model before rescaling and after rescaling using both cases. Means of differences Δ and standard deviations of the differences σ are given for each case. (a) Detection task, inserts of 3.5 mm, 5 mm and 7 mm diameter (epoxy resin), (b) discrimination task, inserts of 6.35 mm and 12.7 mm size (Teflon).

Concerning the values of the slope η of Eq. 8, they vary between 0.60 and 1.63 for the detection task (average: 1.07, median: 1.05, standard-deviation: 0.32) with a median correlation coefficient of 0.94 (standard-deviation: 0.09). Slope and correlation coefficient are generally lower for configurations with smaller sizes and lower kVp. For the discrimination task, the values are more spread out: the slope η ranges from 0.08 to 1.74 (average: 0.83, median: 0.77, standard-deviation: 0.53) with a median correlation coefficient of 0.92 (standard-deviation: 0.27). The lowest slopes are obtained for the largest insert with correlation coefficients higher than 0.91.

The effect of both methods of rescaling is particularly visible in Figure 8, showing an example of the variations of the PC according to CTDIvol for the NPWE model after rescaling in the discrimination task, for the different values of tube voltages. Compared to Figure 5 (b), it is clear that the NPWE model rescaled with Case 2 is much more representative of the human observers, especially for PC results at 80 kVp. It is the reason why the following results will use the NPWE model rescaled with the second method (noted rNPWE2).

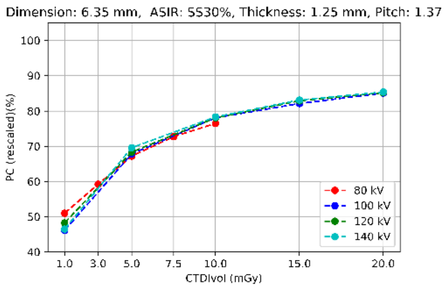
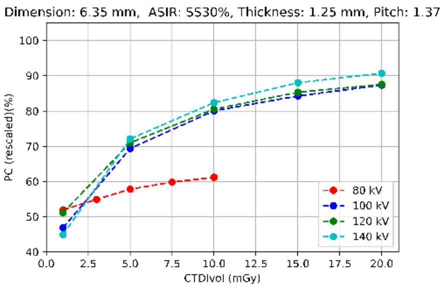
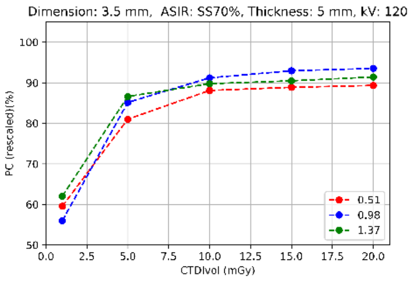
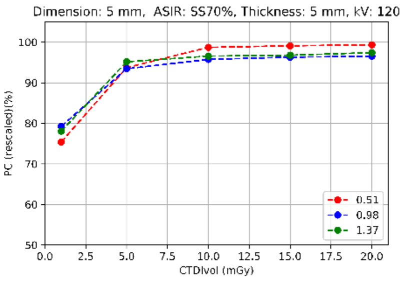
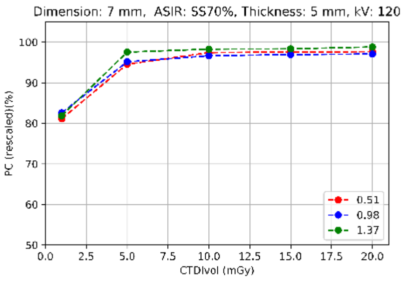
	rNPWE₁ (rescaling Case 1)	*rNPWE₂* (rescaling Case 2)
Discrimination task, insert 6.35 mm		

Figure 8: Comparison of percent correct (PC) calculated by the model observer after rescaling using both methods according to the CTDIvol for the different tube voltages, for the insert of size 6.35 mm in discrimination task.

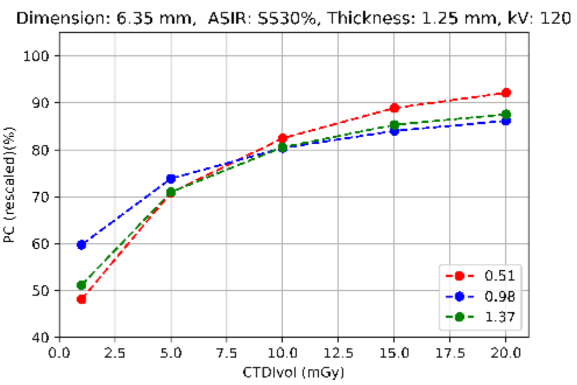
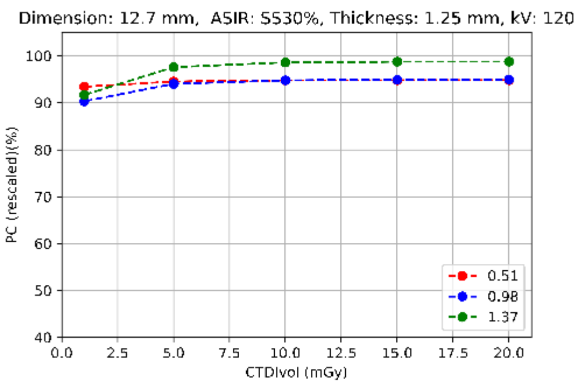
3.4 Impact of acquisition parameters at CTDIvol constant

The model observer *rNPWE₂* was then used to evaluate the impact of acquisition parameters on image quality at CTDIvol constant, following the conditions of **Table 1** and **Table 2** (except for the results on the impact of kVp and ASIR, reported in Appendix B).

The variations of the PC results according to CTDIvol are given for the three available pitch values in Figure 9(a) and Figure 9(b), respectively for the detection task and the discrimination task. They show that the PC results at CTDIvol constant do not seem to depend largely on the pitch, whatever the task and the insert size. This result can be imputable to the cylindrical form of the inserts. Spherical inserts were planned for the detection task but specifications had to be modified due to fabrication problems encountered by the manufacturer.

Insert diameter: 3.5 mm	Insert diameter: 5 mm	Insert diameter: 7 mm
		

(a)

Insert size: 6.35 mm	Insert size: 12.7 mm
	

(b)

Figure 9: Comparison between the percent correct (PC) calculated by the model observer after rescaling *rNPWE₂* according to the CTDIvol for the three values of pitch 0.516, 0.984 and 1.375. (a) Detection task, inserts of 3.5 mm, 5 mm and 7 mm diameter (epoxy resin), (b) discrimination task, inserts of 6.35 mm and 12.7 mm size (Teflon).

The impact of the slice thickness was evaluated on the inserts of diameters 5 mm and 7 mm of the detection task. Figure 10 represents the comparison of the PC according to CTDIvol between the 1.25 mm thickness and the 5 mm thickness, for the 5 mm insert (left) and the 7 mm insert (right). It is generally established that it is necessary to multiply the tube charge per rotation (or mAs) by a factor of 4 (and in consequence the associated CTDIvol and the patient dose) to compensate a slice thickness divided by the same factor, in order to preserve an equivalent image quality. We can observe that this rule is fairly consistent at low dose in the increasing part of the curves. However, as the curves present an asymptotic form, as soon as the required PC threshold is reached for an insert size and a slice thickness, it is not necessary to increase the mAs beyond. Nevertheless, as for the pitch, this assertion could be modified for inserts of different shapes and sizes (typically with a spherical shape).

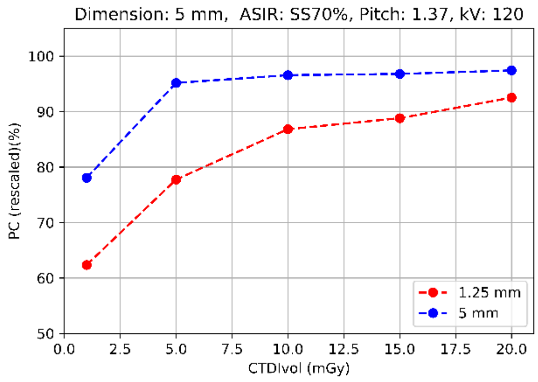
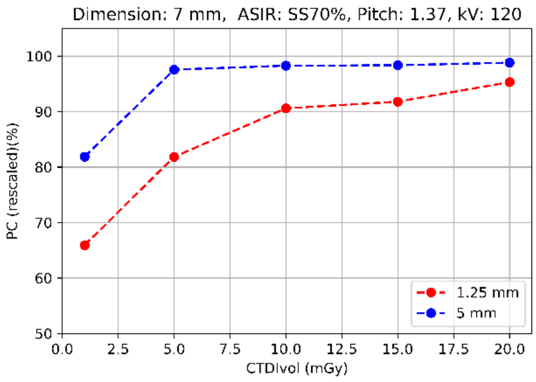
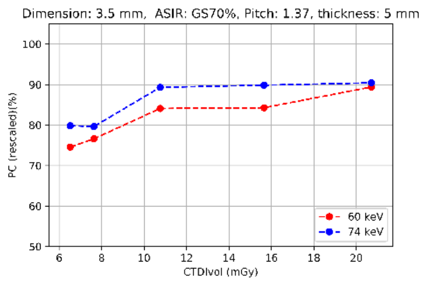
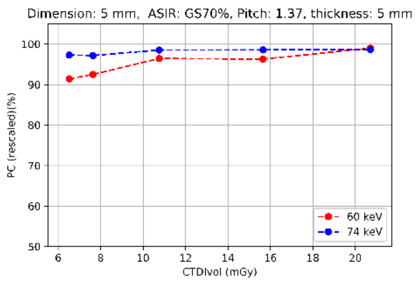
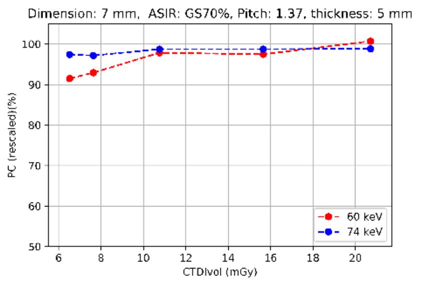
Insert diameter: 5 mm	Insert diameter: 7 mm
	

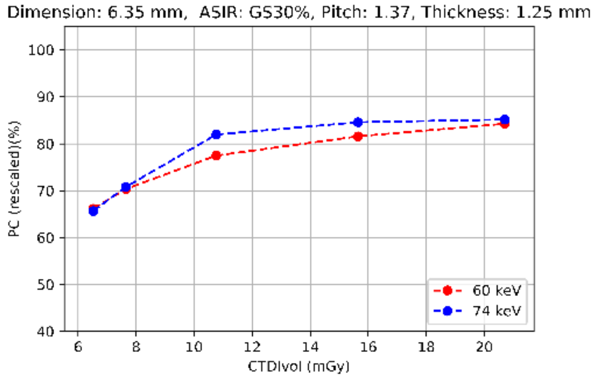
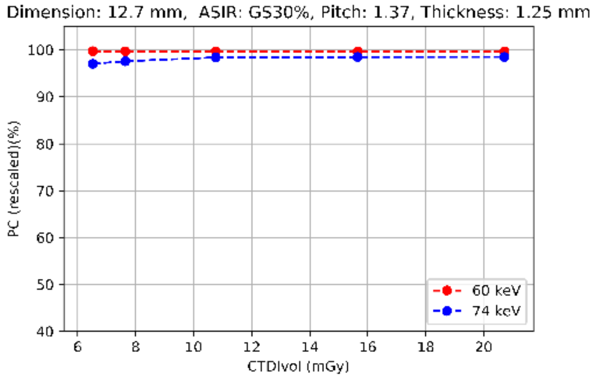
Figure 10: Comparison between the percent correct (PC) calculated by the model observer after rescaling *rNPWE₂* according to the CTDIvol for two values of slice thicknesses 1.25 mm and 5 mm, for the detection task (inserts of 5 mm and 7 mm diameters).

Concerning the impact of the GSI energy level, compared to the default value of 60 keV, the second used energy level (74 keV) globally gives better results of PC, especially on the small inserts or the low CTDIvol for the detection task (see Figure 11(a)), with a gain of PC of about 7 points at the maximum. The non-regular form of curves is due to the fact that the

CTDIvol range (6 mGy – 20 mGy) for the GSI experiments begins towards the inflection point of the curves. For the discrimination task (Figure 11(b)), slightly better results of PC for 74 keV appear for the mean values of CTDIvol (10 mGy-16 mGy) for the smallest insert, due to the form of the response curve and the high uncertainty of the human observers that were used as a reference for the rescaling. At the ends of the range (6-8 mGy and 20 mGy), the answer rates are similar for both energies. For the 12.7 mm insert, the PC results are close to 100 % for both energy levels on the CTDIvol range. Because of the high human uncertainty associated to the reference points, the rNPWE2 correct answer rates for 60 keV appearing a little better than the correct answer rates at 74 keV have no real base.

Insert diameter: 3,5 mm	Insert diameter: 5 mm	Insert diameter: 7 mm
		

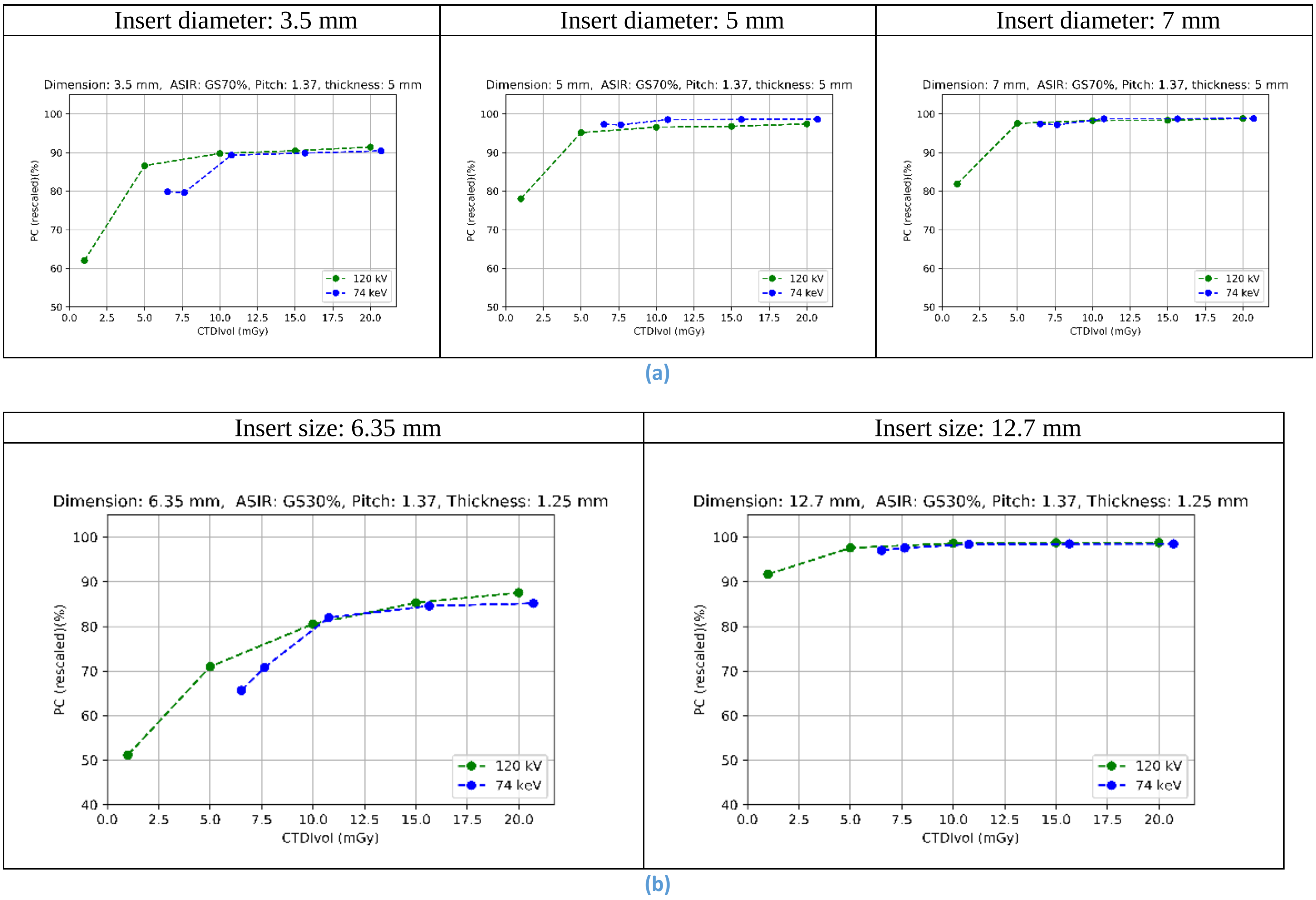
(a)

Insert size: 6.35 mm	Insert size: 12.7 mm
	

(b)

Figure 11: Percent correct (PC) calculated by the model observer after rescaling *rNPWE$_2$* according to the CTDIvol, compared for GSI energy levels 60 keV and 74 keV. (a) Detection task, inserts of 3.5 mm, 5 mm and 7 mm diameter (epoxy resin), (b) discrimination task, inserts of 6.35 mm and 12.7 mm size (Teflon).

At last, the single-energy mode and the dual-energy mode were compared. The PC calculated by rNPWE2 is displayed for the detection task (Figure 12(a)) and the discrimination task (Figure 12(b)) for 120 kVp, which is a standard tube voltage for abdomen examinations, and the GSI energy level 74 keV, which seemed to be the best of both tested energies. For the largest inserts (5 mm, 7 mm, 12.7 mm), there were no particular differences between 120 kVp and 74 keV on the common CTDIvol range (CTDIvol higher than 6.5 mGy), with PC results in the asymptotic part of the curves. The same effect was noticed for the smallest inserts (3.5 mm and 6.35 mm), for the CTDIvol higher than 10 mGy. For the lowest CTDIvol (between 6.5 mGy and 10 mGy), the correct answer rates at 120 kVp were better than the PC results at 74 keV, by 5-10 %.



(a)

(b)

Figure 12: Comparison between the percent correct (PC) calculated by the model observer after rescaling *$rNPWE_2$* according to the CTDIvol in both 120 kVp single-energy and 74 keV GSI energy level. (a) Detection task, inserts of 3.5 mm, 5 mm and 7 mm diameter (epoxy resin), (b) discrimination task, inserts of 6.35 mm and 12.7 mm size (Teflon).

For all the applied variations, it can be observed that the graphs exhibit little difference in results between 5 mm insert and 7 mm insert for the detection task; for the 12.7 mm insert of the discrimination task, it can also noticed that PC is systematically higher than 90 %.

4. Discussion

In this study, our objective was to assess the capability of the NPWE model observer to be a robust and accurate estimator of IQ with the aim, in a second step, to compare protocols, including dual-energy mode, from a double point of view of IQ and dose. In this context, it was necessary to check that the model observer represents the human observers as closely as possible when the scanning and reconstruction parameters are modified.

We observed that the first results of our model observer, the NPWE before rescaling, presented the same trends as the human observers, with generally better results, in agreement with all authors comments. To rescale the model, we compared two alternatives: the first one was similar to [6][20], including all images of a given insert. In the second one, we fitted the human observers' PC results separately for each set of scanning parameters (insert size, kVp, pitch, slice thickness). Using the usual Bland-Altman method, we observed that the first method failed to rescale the model observer to agree with humans in all cases, with PC deviations of up to 20 % between the model and humans (for instance, the point 80 kVp-10 mGy, discrimination task, Figure 8 left). That assessment can reflect a limit of the model, because it pinpoints its difference with human observer behavior when scanning parameters vary. Nevertheless, rescaled by scanning configuration (second method), a good accuracy is obtained, with differences with humans smaller than 4.10 % in 68 % of points in the worst case (the smallest insert of the detection task), and we can conclude that the model will allow to compare our CT scanner protocols in an objective way.

A second objective of the study was to evaluate the impact of several scanning and reconstruction parameters with this IQ metrics. To our knowledge, such a large multiparametric study using a *NPWE* model observer rescaled on human observers has never been undertaken. Studies are generally limited to a single parameter, generally the reconstruction algorithm type. Authors in [20] also studied the influence on kVp, but in the single-energy mode exclusively, and only on a low contrast detection task. Pitch and slice thickness seem to be well characterized parameters, but they have never been analyzed from the point of view of clinical tasks. *A fortiori*, the more recent dual-energy mode and the single-energy mode have never been contrasted in this way. All of these scanning parameters could however provide an assistance in reducing the dose and have therefore to be studied from an IQ point of view.

We could mainly establish that

(1) The pitch did not seem to have a great influence on PC at CTDIvol constant, for both tasks. It can be due to the cylindrical form of the inserts.

(2) The dual-energy mode at 74 keV seemed to be equivalent to the single mode at 120 kVp for the easiest cases in the common range of CTDIvol, but provided lower results for the smallest inserts when the dose decreased, for both tasks. However, the choice of 74 keV as energy level could be still questioned. For low contrast detection, energy levels close to 74 keV are effectively usual, but for high contrasts, other energies could be explored, in particular 40 keV, usual for protocols with injection of iodinated contrast medium, which has a contrast close to Teflon.

From the human observer study, no particular difference was noticed between ASIR 30 % and ASIR 70 %, which was coherent with [43] for the detection task (same CT scanner, 120 kVp). On the discrimination task, to our knowledge, no study has ever been conducted with such a high contrast (Teflon CT number of about 900). With 70 HU and 90 HU contrast lesions, Zhang [26] found that PCs corresponding to the humans were better for the iterative algorithm (SAPHIRE) than for the FBP, but by a few percent only, depending on lesion size, contrast and dose.

Concerning the impact of the tube voltage on PC results, we saw that, for constant CTDIvol, the 80 kVp voltage seemed to be globally less favorable than the other three voltages for both tasks, especially for the discrimination task. On the discrimination task, few studies have been led in particular for this type of material (Teflon). The high value of the Teflon CT number and the associated high contrast, combined with the high diameter of our water-filled phantom, could match an examination of prosthesis material. Our results reveal that, for a standard adult abdomen size, all the tube voltages between 100 kVp and 140 kVp are suitable, and would be then compatible with the high tube voltages recommended by the constructor (120 kVp, 140 kVp). However, for the detection task, this result was unexpected as it is generally admitted that low kVps favour low contrasts. This statement was confirmed in a similar study led in [20], indicating that, for the smallest objects and for both tested contrasts (1 % and 0.5 %), the best tube voltage was 80 kVp at dose constant. Several reasons can explain the difference in the results:

(a) Our phantom, with dimensions of 26 cm x 35 cm, is larger than the Catphan phantom (20 cm diameter) used in [20].

(b) The nature of the used materials: epoxy resin and Plastic Water do not seem to have the same behavior as human tissues or the Catphan materials. Indeed, we measured a slightly lower contrast at 80 kVp (2.5 %) than at 120 kVp (3 %) between the inserts and the background around them, not only due to the epoxy resin, which gives higher HU values at 80 kVp, but essentially due to Plastic Water, whose CT number increases more at 80 kVp than epoxy resin (about 22 HU at 80 kVp, 11 HU at 120 kVp).

(c) The noise level in the images: the measured NPS at 80 kVp is particularly higher than at the other kVps for a same value of CTDIvol, which contributes to make the inserts more difficult to distinguish. The NPWE observer being unable to represent completely this trend, an improvement would be then to increase the NPS contribution in the model formula. The improved formulation of NPWE introduced by [6] should also be evaluated.

The model observer approach has the advantage to obtain more regular trends than humans' ones. Indeed, despite the large number of human observers (between 4 and 13 observers according to the series) and the high number of similar images of the same case (48), the evolution of PC as a function of the CTDIvol appeared to be rather chaotic, due to the high variability of the human responses, as observed in [42]. It can be noted that the effect of the different number of human observers on the uncertainty bars was rather limited: on the one hand, this number was well taken into account in the calculation of $CI_{95\%}$, and on the other hand, the lowest number of human observers (4) was associated with the smallest raw standard deviations of the responses (140 kVp). In consequence, this did not negatively affect the accuracy on the estimation of PCs of one kVp compared to another kVp.

In contrast, a study limitation might have been the window settings for the human observers: as the discrimination slab initially included both polystyrene and Teflon inserts, the same display window of the abdomen protocol was used for the discrimination task despite the high contrast level of Teflon objects, which thus appeared completely saturated. The inappropriate window settings may have reduced the discrimination possibilities of human observers and this part of the protocol should be improved.

A limitation of our work may lie in the use of a model incorporating elements sensitive to nonlinearities, in a study implying IR algorithm. We adapted the resolution calculation to the object contrast, with the use of the TTF on low contrast objects, like [6][7], and the use of the MTF on high contrast inserts. However, despite the use of MTF_{task} adapted to both contrasts, we did not obtain a universal observer fitting the human observers in all experimental conditions. It would be consistent with the fact that some authors ([21]) seemed to have used the MTF in the NPWE expression in low contrast detection task without impact on their results. This tends to show that the MTF_{task} does not seem to be a factor of importance for the NPWE observer.

Most importantly, this work would gain in robustness by an estimation of the uncertainties on the NPWE results, which cannot be naturally expressed from the model. For that, a feasible approach could rely on a bootstrap resampling, in its simple form [6], i.e. directly taking into account the uncertainties of the repeated acquisitions used by the model [6], or in the form of cross-validation [20]. A more involved statistical analysis based on scans replicates could also be finally considered in the medium term.

Finally, it seems essential to evaluate PC on other shapes of inserts, spherical for instance, more difficult to detect in some configurations, and, more generally, to design phantoms (including pediatric size) closer to actual clinical problematics, in shapes, sizes and natures of materials.

5. Conclusions

In conclusion, we showed the capacity of the NPWE model observer to quantify the image quality in a larger context that aims to compare and optimize CT protocols, especially the ones using new irradiating modes such as the dual-energy mode. We based our work on CT images, acquired in a large set of scanning conditions, of a phantom that we designed for two types of tasks: a low contrast detection task and a discrimination task. We demonstrated that a rescaling method using specific parameters according to the scanning conditions (in particular kVp, slice thickness) and the insert size was the most efficient so that the NPWE model can replace humans as equivalent objective metrics. Using the rescaled model observer, the impact of some scanning parameters (especially the comparison between the dual mode and the single mode) was estimated, for both tasks and each insert size. This was carried out using the CTDIvol as metrics of dose, which is the first step of our approach. The next stage is to develop a Monte Carlo model of the scanner in order to obtain simulated patient dose, which will represent a new estimation of dose in replacement of the CTDIvol. The last step is then to compare and optimize protocols using these two new estimations of image quality and dose. First results of the complete process were presented in an oral session of the International Conference on Monte Carlo Techniques for Medical Applications in Montréal (Canada, June 2019, [44]).

Acknowledgements

This work was funded by the French ANSM (Agence Nationale de la Sécurité du Médicament et des dispositifs médicaux) through a grant to the QuaDos project.

References

[1] Institut de Radioprotection et de Sûreté Nucléaire, "Exposition de la population française aux rayonnements ionisants liée aux actes de diagnostic médical en 2012",RapportPRP-HOM N°2014-6, 2014.

[2] Directorate-General for Energy (European Commission),"Medical radiation exposure of the European population", EU Publication, 2015. (URL: https://publications.europa.eu/en/publication-detail/-/publication/d2c4b535-1d96-4d8c-b715-2d03fc927fc9/language-en/format-PDF/source-65412720# (accessed on Feb. 15th 2018))

[3] F. A. Miéville, F. Gudinchet, F. Brunelle, F. O. Bochud, and F. R. Verdun, "Iterative reconstruction methods in two different MDCT scanners: Physical metrics and 4-alternative forced-choice detectability experiments - A phantom approach," *Phys. Medica*, vol. 29, no. 1, pp. 99–110, 2013.

[4] J. Y. Vaishnav, W. C. Jung, L. M. Popescu, R. Zeng, and K. J. Myers, "Objective assessment of image quality and dose reduction in CT iterative reconstruction," *Med. Phys.*, vol. 41, no. 7, 2014.

[5] A. C. Silva, H. J. Lawder, A. Hara, J. Kujak, and W. Pavlicek, "Innovations in CT dose reduction strategy: Application of the adaptive statistical iterative reconstruction algorithm," *Am. J. Roentgenol.*, vol. 194, no. 1, pp. 191–199, 2010.

[6] Christianson, O., Chen, J.J.S., Yang, Z., Saiprasad, G., Dima, A., Filliben, J.J., Peskin, A., Trimble, C., Siegel, E.L. and Samei, E. (2015). An Improved Index of Image Quality for Task-based Performance of CT Iterative Reconstruction across Three Commercial Implementations. Radiology, 275(3), pp.725–734.

[7] J. Solomon and E. Samei, "Correlation between human detection accuracy and observer model-based image quality metrics in computed tomography.," *J. Med. imaging (Bellingham, Wash.)*, vol. 3, no. 3, p. 035506, 2016.

[8] J. D. Evans, D. G. Politte, B. R. Whiting, J. A. O'Sullivan, and J. F. Williamson, "Noise-resolution tradeoffs in x-ray CT imaging: A comparison of penalized alternating minimization and filtered backprojection algorithms," *Med. Phys.*, vol. 38, no. 3, pp. 1444–1458, 2011.

[9] S. Richard, D. B. Husarik, G.Yadava, S. N. Murphy, and E. Samei, "Towards task-based assessment of CT performance: System and object MTF across different reconstruction algorithms," *Med. Phys.* Vol. 39, pp.4115–4122, 2012.

[10] E. Samei and S. Richard, "Assessment of the dose reduction potential of a model-based iterative reconstruction algorithm using a task-based performance metrology," *Med. Phys.*, vol. 42, no. 1, pp. 314–323, 2015.

[11] B. Chen, O. Christianson, J. M. Wilson, and E. Samei, "Assessment of volumetric noise and resolution performance for linear and nonlinear CT reconstruction methods," *Med. Phys.*, vol. 41, no. 7, 2014.

[12] Barrett HH and Myers KJ, "Foundations of image science," *Comput. Med. Imaging Graph.*, vol. 31, no. 2, pp. 114–115, 2004.

[13] H. H. Barrett, J. Yao, J. P, and K. J. Myers, "Model Observers for Assessment of Image Quality," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 90, no. 21, pp. 9758–9765, 1993.

[14] G. J. Gang, J. Lee, J. W. Stayman, D. J. Tward, W. Zbijewski, J. L. Prince, and J. H. Siewerdsen, "Analysis of Fourier-domain task-based detectability index in tomosynthesis and cone-beam CT in relation to human observer performance," *Med. Phys.*, vol. 38, no. 4, pp. 1754–1768, 2011.

[15] W. Vennart, "ICRU Report 54: Medical imaging—the assessment of image quality.," *Radiography*, vol. 3, pp. 243–244, 1996.

[16] X. He and S. Park, "Model observers in medical imaging research," *Theranostics*, vol. 3, no. 10, pp. 774–786, 2013.

[17] K. J. Myers, "Handbook of Medical Imaging: Physics and Psychophysics.," *SPIE Opt. Eng. Press*, vol. 1, pp. 561–87, 2000.

[18] R. F. Wagner, D. G. Brown, and M. S. Pastel, "Application of information theory to the assessment of computed tomography," *Med. Phys.*, vol. 6, no. 2, pp. 83–94, 1979.

[19] A. E. Burgess, "Statistically defined backgrounds: performance of a Modified Nonprewhitening Observer Model," *J. Opt. Soc. Am. A*, vol. 11, no. 4, pp. 0–5, 1994.

[20] I. Hernandez-Giron, A. Calzado, J. Geleijns, R. M. S. Joemai, and W. J. H. Veldkamp, "Low contrast detectability performance of model observers based on CT phantom images: KVp influence," *Phys. Medica*, vol. 31, no. 7, pp. 798–807, 2015.

[21] I. Hernandez-Giron, A. Calzado, J. Geleijns, R. M. S. Joemai, and W. J. H. Veldkamp, "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.*, vol. 87, no. 1039, 2014.

[22] M. Eckstein, J. Bartroff, C. Abbey, J. Whiting, and F. Bochud, "Automated computer evaluation and optimization of image compression of x-ray coronary angiograms for signal known exactly detection tasks," *Opt Express*, vol. 11, no. 5, pp. 460–475, 2003.

[23] K. Li, D. Gomez-Cardona, J. Hsieh, M. G. Lubner, P. J. Pickhardt, and G.-H. Chen, "Statistical model based iterative reconstruction in clinical CT systems. Part III. Task-based kV/mAs optimization for radiation dose reduction.," *Med. Phys.*, vol. 42, no. 9, pp. 5209–5221, 2015.

[24] D. Racine, A. H. Ba, J. G. Ott, F. O. Bochud, and F. R. Verdun, "Objective assessment of low contrast detectability in computed tomography with Channelized Hotelling Observer.," *Phys. Medica*, vol. 32, no. 1, pp. 76–83, 2016.

[25] J. G. Ott, A. Ba, D. Racine, N. Ryckx, F. O. Bochud, H. Alkadhi, and F. R. Verdun, "Patient Exposure Optimisation Through Task-Based Assessment of a New Model-Based Iterative Reconstruction Technique," *Radiat. Prot. Dosimetry*, vol. 169, no. 1–4, pp. 68–72, 2016.

[26] Y. Zhang, S. Leng, L. Yu, R. E. Carter, and C. H. McCollough, "Correlation between human and model observer performance for discrimination task in CT.," *Phys. Med. Biol.*, vol. 59, no. 13, pp. 3389–404, 2014.

[27] S. Leng, L. Yu, Y. Zhang, R. Carter, A. Y. Toledano, and C. H. McCollough, "Correlation between model observer and human observer performance in CT imaging when lesion location is uncertain," *Med. Phys*, vol. 40, no. 36, pp. 81908–5007, 2013.

[28] L. Yu, S. Leng, L. Chen, J. M. Kofler, R. E. Carter, and C. H. McCollough, "Prediction of human observer performance in a 2-alternative forced choice low-contrast detection task using channelized Hotelling observer: impact of radiation dose and reconstruction algorithms.," *Med. Phys.*, vol. 40, no. 4, p. 041908, 2013.

[29] J. Xu, M. K. Fuld, G. S. K. Fung, and B. M. W. Tsui, "Task-based image quality evaluation of iterative reconstruction methods for low dose CT using computer simulations.," *Phys. Med. Biol.*, vol. 60, no. 7, pp. 2881–901, 2015.

[30] S. K. N. Dilger, S. Leng, B. Chen, R. Carter, C. P. Favazza, J. G. Fletcher,C. H. McCollough, L. Yu, "Localization of liver lesions in abdominal CT imaging: II. Mathematical model observer performance correlates with human observer performance for localization of liver lesions in abdominal CT imaging", *Phys. Med. Biol.*, vol. 64, no. 10,105012, 10 pp, 2019

[31] E. Samei, S. Richard, and L. Lurwitz, "Model-based CT performance assessment and optimization for iodinated and noniodinated imaging tasks as a function of kVp and body size.," *Med. Phys.*, vol. 41, no. 8, p. 081910, 2014.

[32] S. Richard and J. H. Siewerdsen, "Comparison of model and human observer performance for detection and discrimination tasks using dual-energy x-ray images.," *Med. Phys.*, vol. 35, no. 11, pp. 5043–5053, 2008.

[33] A. K. Burgess, "Statistically defined backgrounds: Performance of a modified nonprewhitening matched filter model," *JOSA A*, vol. 11, pp. 1237–1242, 1994.

[34] I. Reiser and R. M. Nishikawa, "Identification of simulated microcalcifications in white noise and mammographic backgrounds," *Med. Phys.*, vol. 33, no. 8, pp. 2905–2911, 2006.

[35] J. L. Mannos and D. J. Sakrison, "The Effects of a Visual Fidelity Criterion on the Encoding of Images," *IEEE Trans. Inf. Theory*, vol. 20, no. 4, pp. 525–536, 1974.

[36] The Phantom Laboratory Inc., "Catphan 500 and 600 manual," *Phantom Lab.*, pp. 1–33, 2006.

[37] J. C. Dainty, Rodney Shaw, and L. J. Cutrona, "Image science: principles, analysis and evaluation of photographic-type imaging processes," *Phys. Today*, vol. 29, p. 71, 1976.

[38] International Commission on Radiation Units and Measurements, "ICRU Report No. 87: Radiation Dose and Image-Quality Assessment in Computed Tomography," vol. 12, no. 87, pp. 89–98, 2012.

[39] A. E. Burgess, "Visual perception studies and observer models in medical imaging," *Semin. Nucl. Med.*, vol. 41, no. 6, pp. 419–436, 2011.

[40] Y. Zhang, B. T. Pham, and M. P. Eckstein, "Evaluation of internal noise methods for Hotelling observer models," *Med. Phys.*, vol. 34, no. 8, pp. 3312–3322, 2007.

[41] J. M. Bland and D. G. Altman, "Agreement between methods of measurement with multiple observations per individual," *J. Biopharm. Stat.*, vol. 17, no. 4, pp. 571–582, 2007.

[42] J. G. Ott, A. Ba, D. Racine, A. Viry, F. O. Bochud, F. R. Verdun,"Assessment of low contrast detection in CT using model observers: Developing a clinically-relevant tool for characterising adaptive statistical and model-based iterative reconstruction",*Zeitschrift für Medizinische Physik*,Vol. 27, no 2,pp 86-97,2017.

[43] A. Viry, C. Aberle, D. Racine, J-F.Knebel, S. T. Schindera, S. Schmidt, F. Becce, F. R. Verdun, "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",Physica Medica, no 48, pp 111-118, 2018.

[44] A. C. Simon et al, "Development and association of new metrics of dose and image quality for optimizing protocols in CT imaging", oral presentation at *2nd International Conference on Monte Carlo Techniques for Medical Applications* (*Montréal, Canada, June 2019*).

Appendix A: NPWE Model Observer Calculation

In this work, we used the non-prewhitening matched filter model with an eye (NPWE). It allows to compute in the Fourier domain a scalar d', usually called detectability index. This index d' reflects the prediction score of a signal's detection in some given conditions. The computation of d' for the NPWE model that we used is as follows:

$$d'_{NPWE}{}^{2} = \frac{\left[\iint MTF^{2}{}_{task}(u,v).W^{2}{}_{task}(u,v)E^{2}(u,v)dudv\right]^{2}}{\iint NPS(u,v)\cdot MTF^{2}{}_{task}(u,v)\cdot W^{2}{}_{task}(u,v)E^{4}(u,v)dudv} \quad (1)$$

where *u* and *v* are the spatial frequencies, MTF_{task} represents the task-based Modulation Transfer Function, *NPS* denotes the Noise Power Spectrum and *E(u,v)* is the eye filter. The entity W_{task} represents the task template, which is the representation of the imaging task under investigation in the spatial domain. Following ICRU report #54 [15], W_{task} is expressed as the Fourier Transform of the difference between two hypotheses *h1* and *h2* as follows:

$$W_{task}(u,v) = |FT[h1(x,y) - h2(x,y)]| \quad (2)$$

where *FT* denotes the Fourier Transform, *h1* and *h2* are the hypothesis functions on the signal description in the spatial domain for the two hypotheses. The different quantities in (1) are supposed to be spatially stationary.

In this study, two clinical tasks with different lesions sizes were investigated, leading to several W_{task}. In the case of the detection task, the considered hypotheses were the following: a uniform background (no signal, *h1(x,y)=0*) and a circular signal (*h2(x,y)*=2D projection of a circular cross-section cylinder, corresponding to the lesion profile in the axial plan). For the discrimination case, the hypotheses were as follows: *h1(x,y)*= 2D projection of an hexagonal cross-section cylinder and *h2(x,y)*= 2D projection of a circular cross-section cylinder, both corresponding to the lesion profiles under investigation in the axial plan. The projections were blurred with a Gaussian filter as done in [34] in order to add some noise induced by the imaging system. In total, three and two W_{task} were respectively generated for the detection and discrimination tasks (one template for each lesion size).

The eye filter E(u,v) represents the contrast sensitivity function (CSF) of the human eye in the spatial domain. The CSF that we used in this study was the same as proposed in [35]:

$$A(f) = 2.6\cdot(0.0192 + 0.114f)\cdot e^{-(0.114f)^{1.1}} \quad (3)$$

where *f* is the spatial frequency (cycle/deg). The shape of the CSF curve shows a peak of the eye sensitivity at medium frequency (at 8 cycles/deg) and then the sensitivity is greatly attenuated from this peak to the high frequencies (60 cycles/deg) such as the human eye acuity.

The physical measures, i.e. the MTF_{task} and the NPS, were computed for all the acquisition parameters under investigation. Following the conclusions of [9], the MTF_{task} was calculated differently depending on the object contrast. For the high contrast inserts (discrimination task), the MTF_{task} was taken equal to the MTF, which is used in the classical formula of the NPWE model. To assess the different MTFs, the CTP528 High-Resolution Module of the CATPHAN® 503 phantom (Phantom Laboratories, New York, USA) was used. The module contains spherical beads in tungsten, embedded in a homogeneous material. One of those beads was used to estimate the point source response function of the CT system: the Point Spread Function (PSF). After that, the Line Spread Functions (LSF) in x and y were evaluated by integrating the PSF along the axes of the image. The MTF is then obtained by averaging the two-dimensional Fourier Transform of these two LSFs as detailed in the CATPHAN® manual [36].

For the low contrast inserts (detection task), the MTF_{task} was considered equal to the TTF (Task Transfer Function) and computed using the method described by [9] and [11] on the acrylic target of the CTP404 module of the CATPHAN® 503 phantom. Indeed, the acrylic rod presents a similar contrast with the CTP404 module background as the epoxy resin insert in the detection slab of our phantom. First, 45 small images of the acrylic cylinder (15 consecutive slices x 3 scan repetitions), 30 pixels side, were averaged to improve the statistics at low dose levels. Next, the distance of each pixel to the center of the disk, in a circular ROI of 25-pixel radius, was computed to generate the raw ESF (Edge-Spread Function). The exact center of the disk was estimated by taking the minimum of the variance of the ESF on the edge for different values of center positions at a subpixel level. The new ESF was then rebinned and smoothed. After that, the smoothed ESF was differentiated to form the LSF. Finally, the TTF was obtained by the Fourier transform of the smoothed LSF, combined with a Hahn window.

To measure the noise in the images, the NPS, which provides a complete description about the noise level over the frequency range of the image [37], was evaluated for all the conditions. To this end, the previous images of the slab with no-insert of the torso-shaped phantom were used. We remind that the slab is made in a uniform material (Plastic Water®) with CT numbers near to the water ones. For each combination of the scanning and reconstruction parameters of the study, several slices coming from the same scan were used to extract ROIs of a fixed size in the center of each slice. Then, the NPS is estimated by taking the magnitude squared of the 2D Fourier Transform of the differences between the ROIs and their mean pixel value as referred in ICRU report #87 [38]:

$$NPS(u,v) = \frac{\Delta x}{N_x}\frac{\Delta y}{N_y}\frac{1}{N_{ROIs}}\sum_{i=1}^{N_{ROIs}} \left|FT\left[ROI_i(x,y) - R\hat{O}I_i\right]\right|^2 \tag{4}$$

where Δx and Δy denote the pixel size in axes *x* et *y* ($\Delta x = \Delta y = 0.488$), N_x and N_y are the number of pixels of the ROIs in *x* and *y* directions ($N_x = N_y = 50$), N_{ROIs} denotes the number of ROIs used in the average, and *FT* is the Fourier Transform. The term$R\hat{O}I_i$ is the mean pixel value of the i^{th} ROI.

For each set of the scanning and reconstruction parameters, the detectability index was computed according to equation (1) for all the lesions of different sizes, i.e. lesions of diameters 3.5, 5 and 7 mm for the detection task, and lesions of size 6.35 and 12.7 mm for the discrimination task.

In order to compare the model detectability to the human performance, d'_{NPWE} values were converted to the same metrics previously used in the 2-AFC experiments, which were the percentage of correct answers (PC) as follows [19][29] (assuming that d'_{NPWE} follows a Gaussian distribution):

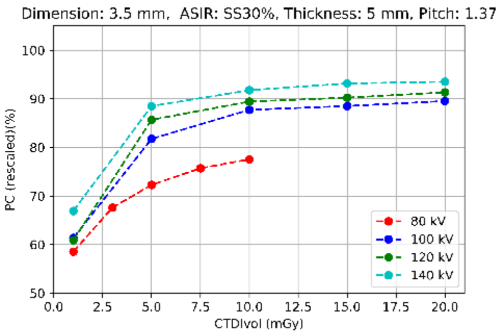
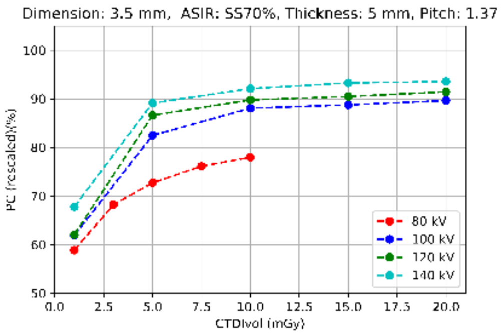
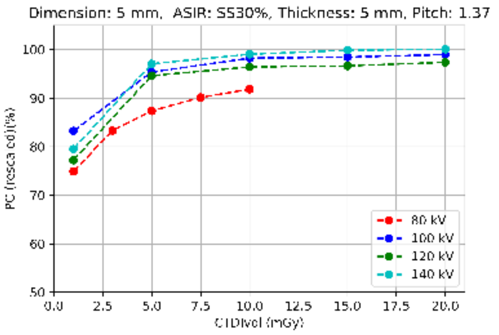
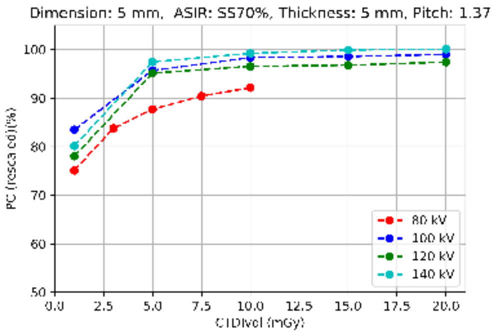
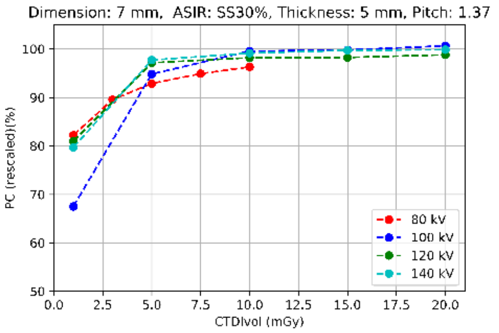
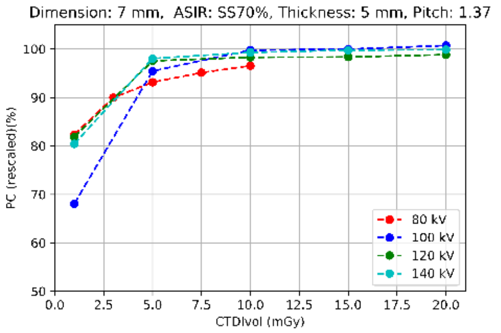
$$PC = \frac{1}{2} + \frac{1}{2} erf\left(\frac{d'_{NPWE}}{2}\right) \tag{5}$$

where *erf* is the Gaussian error function given by the following formula:

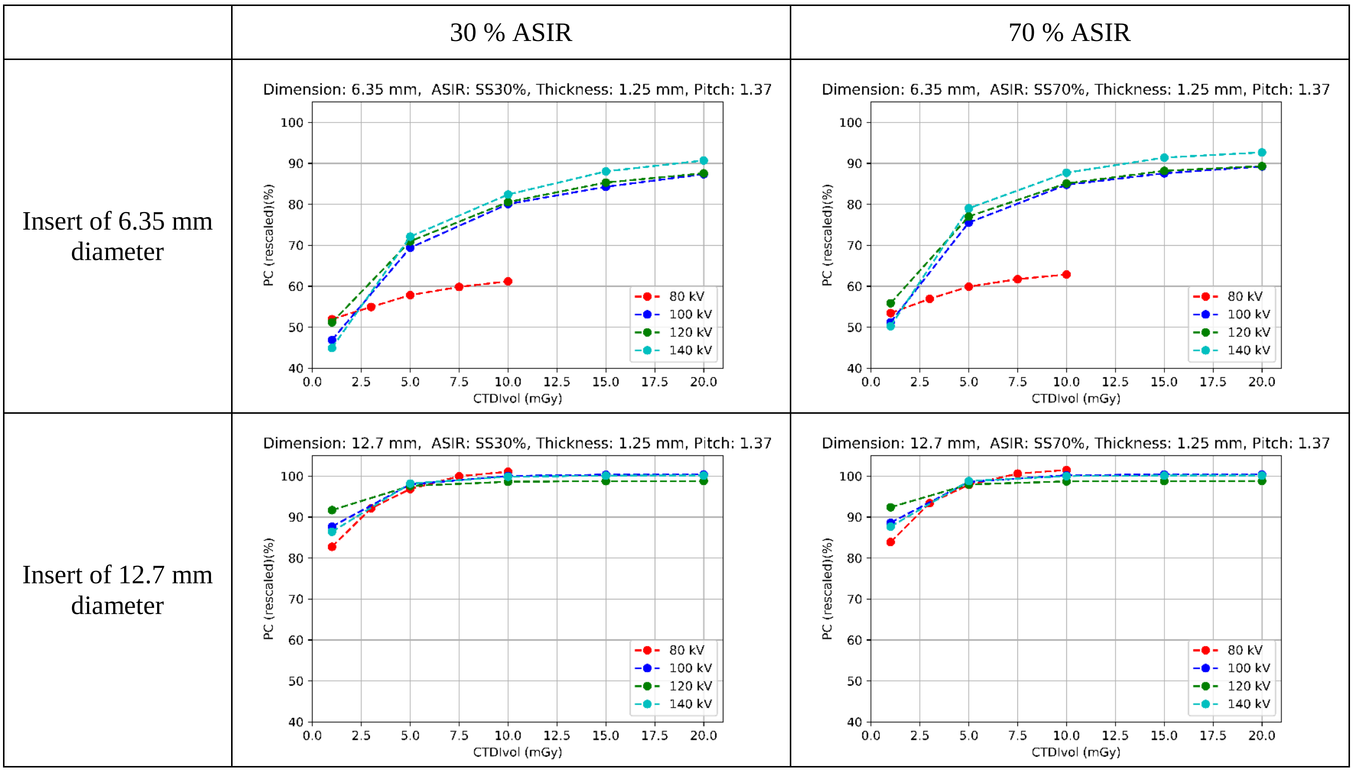
$$erf(x) = \frac{2}{\sqrt{\pi}}\int_0^{\infty} e^{-x^2} dx \tag{6}$$

Appendix B: Impact of kVp and ASIR on Image Quality using rescaled NPWE

The NPWE model observer rescaled by insert size and scanning conditions was used to evaluate the impact of kVp and ASIR on Image Quality. Figure 1a and 1b give the variations of the Percent Correct results according to CTDIvol for the four tube voltages (80 kVp, 100 kVp, 120 kVp and 140 kVp), respectively for the detection task and the discrimination task, for all insert sizes, and both used ASIR levels (30 % and 70 %).

	30 % ASIR	70 % ASIR
Insert of 3.5 mm diameter	 Dimension: 3.5 mm, ASIR: SS30%, Thickness: 5 mm, Pitch: 1.37 	 Dimension: 3.5 mm, ASIR: SS70%, Thickness: 5 mm, Pitch: 1.37
Insert of 5 mm diameter	 Dimension: 5 mm, ASIR: SS30%, Thickness: 5 mm, Pitch: 1.37 	 Dimension: 5 mm, ASIR: SS70%, Thickness: 5 mm, Pitch: 1.37
Insert of 7 mm diameter	 Dimension: 7 mm, ASIR: SS30%, Thickness: 5 mm, Pitch: 1.37 	 Dimension: 7 mm, ASIR: SS70%, Thickness: 5 mm, Pitch: 1.37

(a)



(b)

Figure 13: Comparison between the percent correct (PC) calculated by the model observer after rescaling *$rNPWE_2$* according to the CTDIvol for the four values of kVp and two values of ASIR levels. (a) Detection task, inserts of 3.5 mm, 5 mm and 7 mm diameter (epoxy resin), (b) discrimination task, inserts of 6.35 mm and 12.7 mm size (Teflon).