Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

Correlation between human detection accuracy and observer model-based image quality metrics in computed tomography

Justin Solomon
Ehsan Samei

SPIE.

Justin Solomon, Ehsan Samei, "Correlation between human detection accuracy and observer model-based image quality metrics in computed tomography," *J. Med. Imag.* **3**(3), 035506 (2016), doi: 10.1117/1.JMI.3.3.035506.

Correlation between human detection accuracy and observer model-based image quality metrics in computed tomography

Justin Solomon[a,*] **and Ehsan Samei**[a,b,c]
[a]Duke University Health System, Department of Radiology, Carl E. Ravin Advanced Imaging Laboratories, 2424 Erwin Road, Suite 302, Durham, North Carolina 27705, United States
[b]Duke University Medical Center, Department of Radiology, Clinical Imaging Physics Group, 2424 Erwin Road, Suite 302, Durham, North Carolina 27705, United States
[c]Duke University, Pratt School of Engineering, Departments of Biomedical Engineering and Electrical and Computer Engineering, 2424 Erwin Road, Suite 302, Durham, North Carolina 27705, United States

Abstract. The purpose of this study was to compare computed tomography (CT) low-contrast detectability from human readers with observer model-based surrogates of image quality. A phantom with a range of low-contrast signals (five contrasts, three sizes) was imaged on a state-of-the-art CT scanner (Siemens' force). Images were reconstructed using filtered back projection and advanced modeled iterative reconstruction and were assessed by 11 readers using a two alternative forced choice method. Concurrently, contrast-to-noise ratio (CNR), area-weighted CNR (CNRA), and observer model-based metrics were estimated, including nonprewhitening (NPW) matched filter, NPW with eye filter (NPWE), NPW with internal noise, NPW with an eye filter and internal noise (NPWEi), channelized Hotelling observer (CHO), and CHO with internal noise (CHOi). The correlation coefficients (Pearson and Spearman), linear discriminator error, E, and magnitude of confidence intervals, $|\mathrm{CI}_{95\%}|$, were used to determine correlation, proper characterization of the reconstruction algorithms, and model precision, respectively. Pearson (Spearman) correlation was 0.36 (0.33), 0.83 (0.84), 0.84 (0.86), 0.86 (0.88), 0.86 (0.91), 0.88 (0.90), 0.85 (0.89), and 0.87 (0.84), E was 0.25, 0.15, 0.2, 0.25, 0.3, 0.25, 0.4, and 0.45, and $|\mathrm{CI}_{95\%}|$ was 2.84×10^{-3}, 5.29×10^{-3}, 4.91×10^{-3}, 4.55×10^{-3}, 2.16×10^{-3}, 1.24×10^{-3}, 4.58×10^{-2}, and 7.95×10^{-2} for CNR, CNRA, NPW, NPWE, NPWi, NPWEi, CHO, and CHOi, respectively. [DOI: 10.1117/1.JMI.3.3.035506]

Keywords: image quality; computed tomography; observer models; detectability.

Paper 16065PRR received Apr. 18, 2016; accepted for publication Sep. 8, 2016; published online Sep. 22, 2016.

1 Introduction

X-ray computed tomography (CT) systems contain sophisticated and complex technology with an exceptionally large number of operational settings/modes. New technologies are being introduced on modern CT systems that could allow for very low-dose imaging while maintaining the diagnostic value of the examination.[1] However, their clinical implementation requires a robust and objective evaluation to ensure patient safety and optimal use. Therefore, image quality assessment plays a critical role in the design, optimization, and implementation of new CT technologies.

The most defensible definition of image quality follows the guidelines outlined in ICRU report 54, which defines image quality as the effectiveness by which an image can be used for its intended task.[2] Under this general definition, physical characteristics of the image, such as noise or resolution, may influence image quality but are not metrics of image quality themselves. In other words, resolution is not necessarily a metric of image quality, but it is likely that any proper task-based image quality metric would be sensitive to changing resolution properties of the imaging system, especially if the visualization or measurement of fine details is important to the clinical task.

Because clinical images are evaluated by radiologists, the gold standard of image quality is to assess how well radiologists can perform a clinical task on a set of images. For example, to assess the impact of a new reconstruction algorithm on liver lesion detection, it would be necessary to (1) scan a large number of patients with suspected liver lesions, (2) reconstruct the images with the standard and new algorithm, (3) perform a series of blinded reading sessions, and (4) quantitatively assess the detection rates for both datasets. This type of clinical trial is clearly resource intensive and logistically challenging. Further, optimizing a CT protocol often requires the investigation of many different scan and reconstruction settings, such as dose, tube voltage, bow-tie filtration, pitch, automatic exposure control, convolution kernel, and iterative strength. This sizeable parameter space makes protocol optimization via clinical trials practically impossible. Therefore, surrogate metrics of image quality that can be measured in phantoms (or patient images if possible) offer a sensible alternative. For such metrics to be useful, they must (1) be representative of clinical image quality (i.e., highly correlated with radiologist performance for a specific clinical task), (2) be generalizable such that images with varying noise and resolution properties can be compared (e.g., scanner A versus scanner B, or reconstruction A versus reconstruction B), and (3) be practical to measure with a reasonable number of images.

*Address all correspondence to: Justin Solomon, E-mail: justin.solomon@duke.edu

2329-4302/2016/$25.00 © 2016 SPIE

As described in great detail by Barrett et al.,[3] image quality metrics based on observer models constitute a rich and active field of research and have been proposed to meet the above criteria. In this work, we consider the task signal detection. Mathematically, an observer model can be described as an operator that transforms the input image data, **g**, into a single scalar test statistic, λ. Here, **g** is the vector lexicographically ordered to contain all the image pixels. The observer makes a decision by comparing λ with some threshold, λ_t (if $\lambda > \lambda_t$, then the decision is "signal-present," otherwise it decides "signal-absent"). As a metric of image quality, it is useful to look at the distributions of λ under ensembles of signal-present and signal-absent cases (acquired under identical conditions). A commonly used metric to summarize those distributions is the detectability index, d' (pronounced d-prime), defined as

$$d'^2 \equiv \frac{[\bar{\lambda}_0 - \bar{\lambda}_1]^2}{1/2 \cdot [\sigma_0^2 + \sigma_1^2]}, \tag{1}$$

where $\bar{\lambda}_0$, $\bar{\lambda}_1$, σ_0^2, and σ_1^2 are the means and variances of the test statistic under signal-present and signal-absent conditions, respectively.[4] As defined above, d' is essentially the signal-to-noise ratio of the observer model in discriminating signal-present and signal-absent images. Images of high quality are those that lead to a greater separation (i.e., less overlap) in the signal-present and signal-absent distributions of λ and thus higher d'. Note that d' serves as a scalar summary of the separation between two probability distributions. In the case that λ is normally distributed (in both signal-present and signal-absent cases), then the separation is fully parameterized by d' and d' is an appropriate description. If λ is not normally distributed, then d' may not be appropriate. Fortunately, for medical image data, λ can be assumed to be normal in most cases with appeals to the central limit theorem.[5]

Different observer models process image data differently and also vary in how much prior information they have with respect to the detection task. Recent literature in using observer models for CT image quality assessment contains instances of several different paradigms. For example, the signal known exactly (SKE) paradigm is commonly used.[6,7] In this paradigm, the observer model is assumed to have knowledge about all signal characteristics including size, shape, contrast, and location and uses this information when processing the image data to output a test statistic. Another common paradigm is to allow for the location of the signal to vary and have the observer perform a search throughout the image.[8–10] The test statistic for search-capable models is typically not normally distributed, thus, d' is not a good summary statistic for their detection performance. In another paradigm, the signal's characteristics are known only statistically thus allowing for variability of the signal to be incorporated.[11] For all these paradigms, it is also possible to consider anatomical background variability.[12–15] Readers are further pointed to Barrett et al.[3] for a comprehensive review of observer model theory and Abbey et al.[16,17] for foundational work on the practical considerations of assessing image quality with observer models. Due to the nature of the human detection data used as the basis of this study, we consider only the SKE paradigm with a uniform background in the remainder of this paper, with full recognition that other paradigms merit future comparison.

Within the SKE paradigm, for the same ensemble of images, different observer models will have a different d', and there are many potential models to choose from. Broadly speaking, three general approaches are commonly used to assess tomographic images under this paradigm. The first approach is to use simple first-order image statistics, such as contrast-to-noise ratio (CNR) as direct surrogates of low-contrast detectability.[18,19] Although this approach is not based on mathematical observer models, it is nevertheless considered in this paper due to its prevalence (particularly in clinical papers). The second approach is to measure physical aspects of the imaging system, such as the modulation transfer function (MTF) and noise power spectrum (NPS) and then compute d' based on an analytical relationship between d' and the system MTF/NPS for a given observer model.[6,20–23] This approach most often uses a nonprewhitening (NPW) matched filter observer model whose d' can be computed in the Fourier domain. The third approach utilizes large ensembles of image data to estimate the observer model's d' directly from the signal-present and signal-absent images. The most common model used with this approach is a channelized Hotelling observer (CHO).[9,24–26]

The correlation between detection accuracy of the models and humans has been explored individually for each of these approaches. However, to our knowledge, a comparison between these approaches utilizing the same image data (along with corresponding human reader data) has not been reported in the context of multi-row detector computed tomography images. Therefore, the goal of this work was to compare CT low-contrast detectability as measured by a human perception experiment with several observer model-based estimates of detectability. The objectives were to (1) ascertain the strength of the correlation between the humans and models, (2) assess if the models can be used to characterize how different reconstruction algorithms affect detectability, and (3) evaluate the practicality of measuring the model's performance with a finite number of image realizations.

2 Materials and Methods

The image data and human detection data for this study are drawn from a subset of data acquired as part of a previously published paper designed to assess the impact of iterative reconstruction on low-contrast detectability.[27] The phantom design, image acquisition protocol, and human perception experiment are described in detail in the aforementioned paper and a brief description is given below. This project was financially supported by Siemens Healthcare but the authors maintained full control over all data and had absolute autonomy over inclusion/exclusion of any results or information that may present a conflict of interest to the supporting party.

2.1 *Phantom Design and Image Acquisition*

A custom phantom was designed as a cylindrical disk (diameter 165 mm; axial length 30 mm) containing 45 low-contrast inserts of five contrast levels (5, 9, 12, 15, and 20 HU at 120 kVp) and three sizes (6, 4, and 2 mm diameter) with repeats of each insert located at three radial distances (33, 48.75, and 64.5 mm). The phantom was fabricated using a multimaterial three-dimensional (3-D) printer (Objet Connex, Stratasys Ltd.) and imaged 20 repeated times on a third-generation dual-source CT system (SOMATOM Force, Siemens Healthcare). Images were captured at 2.9 mGy $CTDI_{Vol}$ and reconstructed at 0.6 mm slice thickness using filtered back projection (FBP) and advanced modeled iterative reconstruction (ADMIRE, strength-3) with the BF44 kernel (Fig. 1). Each image series contained 15 slices

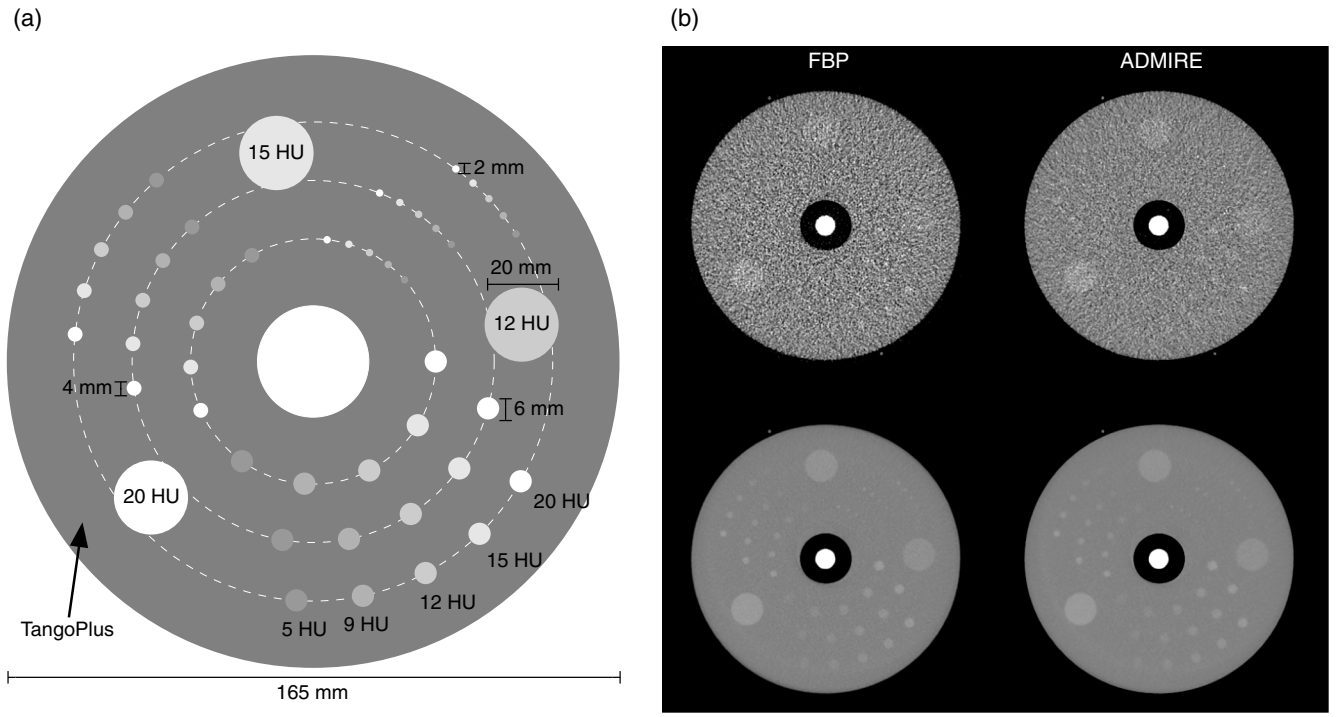

Fig. 1 (a) Diagram of the contrast-detail phantom and (b) example CT images of phantom for FBP (left) and ADMIRE (right). The top row shows a single-image realization, while the bottom row is the expected (i.e., average) image over all 300 realizations.

of interest resulting in a total of 300 image realizations (15 slices × 20 repeats) for each reconstruction condition.

2.2 *2AFC Human Detection Experiment*

A two alternative forced choice (2AFC) multireader diagnostic performance experiment was carried out in order to assess the dependence of lesion detectability on signal contrast, signal size, and reconstruction algorithm. Under the signal known exactly/background known exactly (SKE/BKE) paradigm, a custom–user interface was designed in which two images were shown to a human observer, one containing the signal and one containing only noise. The observer was asked to choose which of the two images was most likely to contain the signal. The signal-present images represented circular regions of interest (ROIs), 15 mm in diameter, drawn about the phantom's cylindrical inserts (signal is always centered in the ROI). The signal-absent images of the same size were taken from independent uniform regions of the phantom at the same phantom radius. For a given combination of the reconstruction algorithm, insert size, and insert contrast, 15 image trials were shown to each observer and the observer scores were taken as the accuracy across those 15 trials. A total of 11 observers participated in the study (6 physicists, 1 physics resident, 3 doctoral students, and 1 radiologist). For each condition, the average score across observers was computed. Also 95% confidence intervals were computed as $\mathrm{CI}_{95\%} = 1.96 \cdot \mathrm{STD}(A)/\sqrt{\#\mathrm{Readers}}$, where A is the detection accuracy for a given observer. All images were viewed in clinical reading-room ambient lighting conditions using a display calibrated to the DICOM standard and at a window width of 250 HU and window level of 75 HU. A total of 20 contrast/size/reconstruction conditions were tested as described in Table 1. As mentioned above, these human reader data represent a subset of data from a previous study that spans a wider range of conditions (3 versus 1 dose levels) compared with this current study. The raw binary 2AFC responses from that full dataset were analyzed with a generalized linear mixed effects statistical model (binomial distribution with a probit link function) to confirm that ADMIRE had a significant effect on detection accuracy.

2.3 *Model Detection Accuracy*

A total of eight image quality figures of merit were considered for this study, including both traditional pixel-value-based metrics and observer model-based metrics. The metrics can be divided into three main groups: (1) traditional pixel-value metrics, (2) NPW matched filter metrics, and (3) CHO metrics. These metrics were extracted from the image data as described below for each reconstruction, size, and contrast condition shown in Table 1.

2.3.1 *Traditional pixel-value metrics*

The traditional pixel-value metrics considered were CNR and area-weighted CNR (CNRA). CNR was defined as

$$\mathrm{CNR} = \frac{\mu_s - \mu_b}{\sigma_b} = \frac{C}{\sigma_b}, \tag{2}$$

where μ_s is the attenuation (in Hounsfield units) of the signal, μ_b is the attenuation of the background, C is the contrast, and σ_b is the standard deviation of pixel values in the background ROI. The nominal contrast of each signal (see Fig. 1) was used for this calculation and the noise was taken as the standard deviation of pixel values in a background annulus-shaped ROI surrounding the signal having an outer radius of 7 mm (~17 pixels) and an

Table 1 Different conditions considered in the human detection accuracy experiment. Here, the size represents the diameter of the signal and the listed nominal contrast is for 120 kVp. Each combination of size and contrast was tested for with reconstruction algorithm, resulting in 20 total conditions.

Condition #	Reconstruction algorithm	Nominal size (mm)	Nominal contrast (HU)
1, 2	FBP/ADMIRE	2	12
3, 4	FBP/ADMIRE	2	15
5, 6	FBP/ADMIRE	2	20
7, 8	FBP/ADMIRE	4	9
9, 10	FBP/ADMIRE	4	12
11, 12	FBP/ADMIRE	4	15
13, 14	FBP/ADMIRE	6	5
15, 16	FBP/ADMIRE	6	9
17, 18	FBP/ADMIRE	6	12
19, 20	FBP/ADMIRE	6	15

inner radius of 1.5 mm (~4 pixels), 3 mm (~7 pixels), and 4.5 mm (~11 pixels) for the 2, 4, and 6 mm nominally sized signal, respectively. Noise was measured for each slice using the ensemble of 20 repeated scans, and the final CNR value was averaged across all 15 slices and 95% confidence intervals were computed as $\mathrm{CI}_{95\%} = 1.96 \cdot \mathrm{STD}(\mathrm{CNR})/\sqrt{\#\mathrm{Slices}}$. Note that this formulation of confidence intervals assumes that each slice is an independent sample. This assumption is based on the fact that the images were acquired in axial mode and there is no overlap between slices. Thus, any correlations across slices would be due only to detector cross talk, which is likely a minimal effect.

The CNRA was computed as

$$\mathrm{CNRA} = \sqrt{A} \cdot \mathrm{CNR}, \tag{3}$$

where A is the nominal area of the signal. As shown in Fig. 1, the phantom contains three sized signals: 6, 4, and 2 mm in diameter. The 95% confidence intervals were calculated in the same way as with CNR.

2.3.2 *Nonprewhitening matched filter*

The NPW matched filter is a linear observer model whose template, $\boldsymbol{\omega}_{\mathrm{NPW}}$, is the difference between the expected signal, $\bar{\mathbf{g}}_{\mathbf{s}}$, and the expected background, $\bar{\mathbf{g}}_{\mathbf{b}}$ (i.e., $\boldsymbol{\omega}_{\mathrm{NPW}} = \bar{\mathbf{g}}_{\mathbf{s}} - \bar{\mathbf{g}}_{\mathbf{b}}$). Thus, the NPW observer forms its test statistic, λ_{NPW}, as

$$\lambda_{\mathrm{NPW}} = \boldsymbol{\omega}_{\mathrm{NPW}}^{t}\mathbf{g} = (\bar{\mathbf{g}}_s - \bar{\mathbf{g}}_b)^t\mathbf{g} = \sum_{\mathbf{i=1}}^{\mathbf{N}}(\bar{g}_{si} - \bar{g}_{bi}) \cdot g_i, \tag{4}$$

where N is the number of pixels in the image (or ROI).[2,4,28] Under the assumption that the noise is wide-sense stationary (at least locally within an small ROI), and that the system behaves in a quasi-linear fashion, the detectability index for the NPW observer, d'_{NPW}, was computed in the Fourier domain as[29]

$$d'^{2}_{\mathrm{NPW}} = \frac{\left[\iint |W(u,v)|^2 \cdot \mathrm{TTF}^2(u,v)\mathrm{d}u\,\mathrm{d}v\right]^2}{\iint |W(u,v)|^2 \cdot \mathrm{TTF}^2(u,v) \cdot \mathrm{NPS}(u,v)\mathrm{d}u\,\mathrm{d}v}, \tag{5}$$

where u and v are the spatial frequencies in the x and y directions, respectively, $W(u,v)$ is the task function (i.e., the Fourier transform of the signal to be detected), $\mathrm{TTF}(u,v)$ is the task transfer function (i.e., the contrast-dependent MTF),[30] and $\mathrm{NPS}(u,v)$ is the noise power spectrum. For this phantom, the signals to be detected were uniform circular disks and thus $W(u,v)$ was the Fourier transform of a disk, given as

$$W(u,v) = \frac{\sqrt{3r}}{4f} J_1(2\pi \cdot f \cdot r), \tag{6}$$

where r is the radius of the disk, f is the radial spatial frequency ($f = \sqrt{u^2 + v^2}$), and J_1 is a Bessel function of the first kind. The TTF was measured from the large rods in the low-contrast detectability phantom with a circular ROI (radius of 20 mm, ~49 pixels) using the method described by Richard et al.[30] and refined by Chen et al.[31] For each image slice, the NPS was measured by first subtracting the ensemble averaged image from each realization to remove the mean signal and achieve noise-only images. Using all pixels within a radius of 10 mm (~15 pixels) from the center of the signal, the autocorrelation of the noise, $R_N(\Delta x, \Delta y)$, was estimated where Δx and Δy are the distance between two pixels in the x and y directions, respectively. The NPS was computed by taking the two-dimensional (2-D) Fourier transform of $R_N(\Delta x, \Delta y)$ as

$$\mathrm{NPS}(u,v) = p^2 \cdot |\mathcal{F}[R_N(\Delta x, \Delta y)]|, \tag{7}$$

where p is the pixel size and $\mathcal{F}[]$ denotes the discrete 2-D Fourier transform.[32] The d'_{NPW} was estimated for each slice and averaged. The d'_{NPW} was measured for each slice using the ensemble of 20 repeated scans and the final value was averaged across all 15 slices and 95% confidence intervals were computed as $\mathrm{CI}_{95\%} = 1.96 \cdot \mathrm{STD}(d'_{\mathrm{NPW}})/\sqrt{\#\mathrm{Slices}}$.

The NPW observer was extended to include an eye filter, $E(\rho)$, that models the human visual system's sensitivity to different spatial frequencies.[28,33–35] This model was called the NPW matched filter with eye filter (NPWE). The eye filter was defined as

$$E(\rho) = |\eta\rho^{a_1} \cdot e^{-a_2\rho^{a_3}}|^2, \tag{8}$$

where ρ is the angular spatial frequency in cycles/degrees, a_1, a_2, and a_3 are the constant parameters with values of 1.5, 3.22, and 0.68, respectively, and η normalizes the function to have a maximum of one.[36] The image domain radial spatial frequency, r ($r = \sqrt{u^2 + v^2}$), was converted to angular spatial frequency, ρ, with

$$\rho = r \cdot \frac{\mathrm{FOV} \cdot R \cdot \pi}{D \cdot 180}, \tag{9}$$

where FOV is the reconstructed field of view of the image (209 mm), R is the viewing distance (assumed to be 475 mm), and D is the display size (350 mm). Thus, $E(\rho)$ was defined as a function of u and v. When the eye filter is applied, the

detectability index for the NPWE model observer, d'_{NPWE}, was calculated as[22]

$$d'^{2}_{\mathrm{NPWE}} = \frac{\left[\iint |W(u,v)|^2 \cdot \mathrm{TTF}^2(u,v) \cdot E^2(u,v)\,\mathrm{d}u\,\mathrm{d}v\right]^2}{\iint |W(u,v)|^2 \cdot \mathrm{TTF}^2(u,v) \cdot \mathrm{NPS}(u,v) \cdot E^4(u,v)\,\mathrm{d}u\,\mathrm{d}v}. \tag{10}$$

The d'_{NPWE} and corresponding confidence intervals were measured using the same methods as with the NPW model.

The NPW model was alternatively modified to include a component of internal noise (i.e., human visual system noise). This model was called the NPW matched filter with internal noise (NPWi). The internal noise, $N(u,v)$, was assumed to have a constant (i.e., white) power spectrum whose magnitude is proportional to the pixel variance:

$$N(u,v) = \alpha_{\mathrm{NPW}} \left(\frac{R}{1000}\right)^2 \cdot \sigma^2, \tag{11}$$

where α_{NPW} is a proportionality constant, R is the viewing distance [same as in Eq. (9)], and σ^2 is the pixel variance, computed as the 2-D integral of $\mathrm{NPS}(u,v)$.[22,37] With this internal noise factor added, the detectability index for the NPWi model observer, d'_{NPWi}, was calculated as[22]

$$d'^{2}_{\mathrm{NPWi}} = \frac{\left[\iint |W(u,v)|^2 \cdot \mathrm{TTF}^2(u,v)\,\mathrm{d}u\,\mathrm{d}v\right]^2}{\iint [|W(u,v)|^2 \cdot \mathrm{TTF}^2(u,v) \cdot \mathrm{NPS}(u,v) + N(u,v)]\,\mathrm{d}u\,\mathrm{d}v}. \tag{12}$$

The d'_{NPWi} and corresponding confidence intervals were measured using the same methods as with the NPW model. The proportionality constant, α, was empirically chosen such that it maximized the correlation between the model and human data when performing linear regression analysis (see Sec. 2.5). This was done by computing the average correlation (over 100 random resamples, each containing 75% of the data) for many potential values of α.

Finally, the NPW model was extended to include both the eye filter and internal noise simultaneously. This model was called the NPW matched filter with eye filter and internal noise (NPWEi), and its detectability index, d'_{NPWEi}, was calculated as[22]

$$d'_{\mathrm{NPWEi}} = \frac{\left[\int |W(u,v)|^2 \cdot \mathrm{TTF}^2(u,v) \cdot E^2(u,v)\,\mathrm{d}u\,\mathrm{d}v\right]^2}{\int [|W(u,v)|^2 \cdot \mathrm{TTF}^2(u,v) \cdot \mathrm{NPS}(u,v) \cdot E^4(u,v) + N(u,v)]\,\mathrm{d}u\,\mathrm{d}v}. \tag{13}$$

The d'_{NPWEi} and corresponding confidence intervals were measured using the same methods as with the NPW model.

2.3.3 *Channelized Hotelling observer*

The CHO is a linear model that operates on channelized image data, meaning image data that have been passed through M number of filters.[4,38,39] Each filter produces a single scalar output and thus the collection of filter outputs is an $M \times 1$ vector called "channel outputs," denoted $\mathbf{g}_c$. The transformation of image data, $\mathbf{g}$, to the channelized data, $\mathbf{g}_c$, can be described as a matrix multiplication:

$$\mathbf{g}_c = \mathbf{U}^t \mathbf{g}, \tag{14}$$

where $\mathbf{U}$ is an $N \times M$ matrix where each column corresponds to a separate channelizing filter (N is the number of pixels in the image or ROI). For a given image, the test statistic, λ_{CHO}, is computed by taking the inner product of a template, $\boldsymbol{\omega}_{\mathrm{CHO}}$, with $\mathbf{g}_c$ as

$$\lambda_{\mathrm{CHO}} = \boldsymbol{\omega}^t_{\mathrm{CHO}} \mathbf{g}_c. \tag{15}$$

The CHO template is formed by taking the difference between the expected channel output when the signal is present and the expected channel output when the signal is absent, multiplied by the inverse of the intraclass channel scatter matrix as

$$\boldsymbol{\omega}_{\mathrm{CHO}} = \mathbf{S}_c^{-1}[\bar{\mathbf{g}}_{sc} - \bar{\mathbf{g}}_{bc}], \tag{16}$$

where $\bar{\mathbf{g}}_{sc}$ is the expected (i.e., average) channelized output when the signal is present, $\bar{\mathbf{g}}_{bc}$ is the expected channelized output when the signal is absent, and $\mathbf{S}_c$ is the intraclass channel scatter matrix, defined as the average of the channel output covariance matrices, under the signal-present and signal-absent conditions.

The detectability index for the CHO model, d'_{CHO}, is given as[4]

$$d'^{2}_{\mathrm{CHO}} = [\bar{\mathbf{g}}_{sc} - \bar{\mathbf{g}}_{bc}]^{\mathbf{t}} \mathbf{S}_c^{-1} [\bar{\mathbf{g}}_{sc} - \bar{\mathbf{g}}_{bc}]. \tag{17}$$

Following a recent publication by Yu et al.,[38] Gabor filter channels were utilized in this study. Such filters have been designed to emulate the human visual system.[34,39] A Gabor filter, $G(x,y)$, is an exponential function with a given center location, (x_0, y_0), and channel width, w_s, that is modulated by a sinusoid with a given central frequency, f_c, orientation, θ, and phase, β, expressed as[40]

$$G(x,y) = \cos\{2\pi f_c[(x-x_0)\cos\theta + (y-y_0)\sin\theta] + \beta\} \cdot e^{-4\ln(2)[(x-x_0)^2 - (y-y_0)^2]/w_s^2}. \tag{18}$$

The center location was chosen to be the center of the signal. A total of 60 Gabor filters (i.e., $M = 60$) were used corresponding to six channel passbands: $[1/128, 1/64]$, $[1/64, 1/32]$, $[1/32, 1/16]$, $[1/16, 1/8]$, $[1/8, 1/4]$, and $[1/4, 1/2]$ cycles/degrees, four orientations: 0, $2\pi/5$, $4\pi/5$, and $6\pi/5$, rad, and two phases: 0 and $\pi/2$ rad. w_s and f_c were taken as the widths and centers of the passbands, respectively. The passbands were converted to cycles/mm using the inverse of Eq. (9) and the same viewing conditions described in Sec. 2.4.2 (image FOV of 209 mm, viewing distance of 475 mm, and image display size of 350 mm). The filters were designed to be the same size (in pixels) as the images that were shown to the human observers. Figure 2 shows the images of the 60 filters.

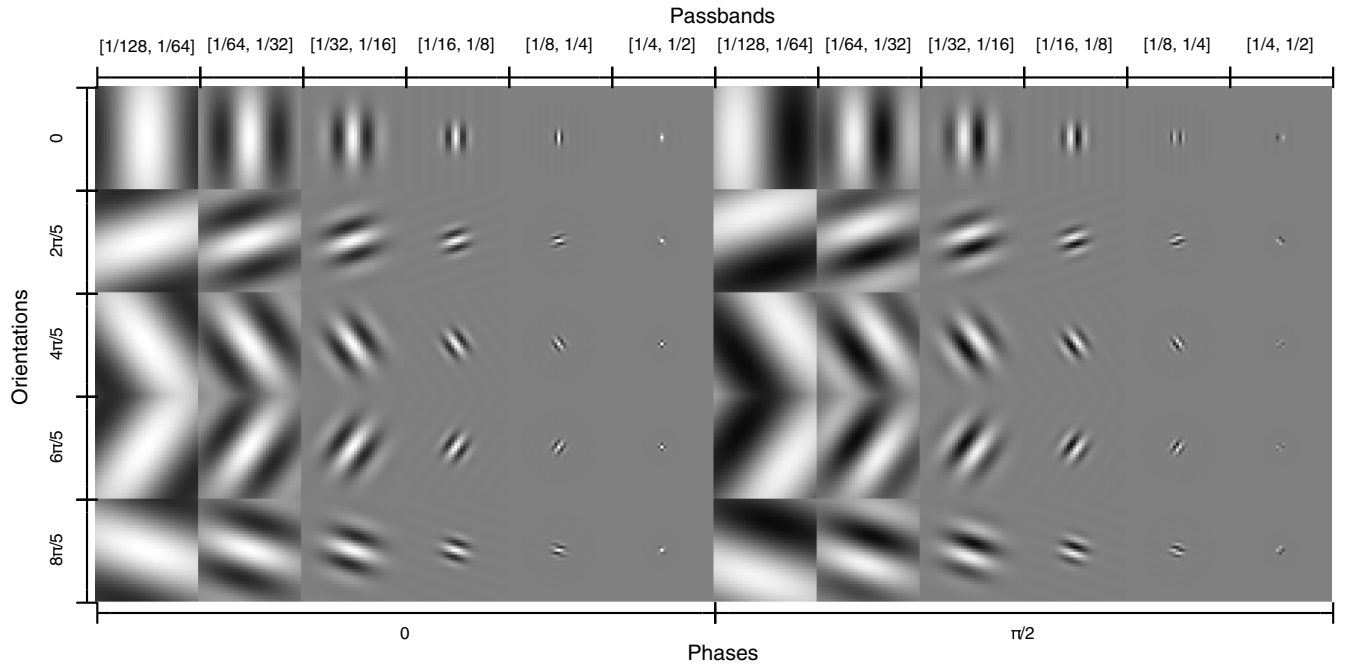

Fig. 2 Montage of the 60 Gabor filters used to channelize the image data.

Using the 300 image realizations, $\bar{\mathbf{g}}_{sc}$, $\bar{\mathbf{g}}_{bc}$, and $\mathbf{S}_c$ were estimated for each reconstruction, signal size, and signal contrast condition. Based on those estimates, the CHO model's detectability index, d'_{CHO}, along with 95% confidence intervals were estimated using a method described by Wunderlich et al.[41]

The CHO model was further extended to include internal noise, denoted CHOi. The covariance matrix of the internal noise, $\mathbf{N}_c$, was assumed to be diagonal (i.e., noise between channels was uncorrelated) with values proportional to the variance of the channel outputs (i.e., proportional to the diagonal elements of $\mathbf{S}_c$) as

$$N_{ci,j} = \begin{cases} \alpha_{\mathrm{CHO}} \cdot S_{ci,j}, & i = j \\ 0, & \text{otherwise} \end{cases}, \tag{19}$$

where $N_{c_{i,j}}$ and $S_{c_{i,j}}$ are the i'th/j'th elements of $\mathbf{N}_c$ and $\mathbf{S}_c$, respectively, and α_{CHO} is a proportionality constant. The detectability index for the CHOi model, d'_{CHOi}, is given as[4]

$$d'^{2}_{\mathrm{CHOi}} = [\bar{\mathbf{g}}_{sc} - \bar{\mathbf{g}}_{bc}]^{\mathbf{t}}[\mathbf{S}_c + \mathbf{N}_c]^{-1}[\bar{\mathbf{g}}_{sc} - \bar{\mathbf{g}}_{bc}]. \tag{20}$$

The method described by Wunderlich et al. to estimate CHO performance and confidence intervals does not include the internal noise component, and a closed-form expression for obtaining unbiased estimates of d'_{CHOi} and corresponding exact confidence intervals based on sample estimates of $\bar{\mathbf{g}}_{sc}$, $\bar{\mathbf{g}}_{bc}$, and $\mathbf{S}_c$ is currently an open problem. As such, a simulation approach was used in which zero-mean normally distributed internal noise was added to the channel outputs. After adding the noise, d'_{CHOi} and its corresponding 95% confidence interval were obtained using the Wunderlich method on the updated (i.e., after adding internal noise) estimates of $\bar{\mathbf{g}}_{sc}$, $\bar{\mathbf{g}}_{bc}$, and $\mathbf{S}_c$. This process was repeated 1000 times for each condition and the final estimate of d'_{CHOi} and its confidence interval was taken as the average across the 1000 repeated trials. As with the NPWi model, the internal noise proportionality constant, α_{CHO}, was empirically chosen such that it maximized the coefficient of determination (R^2) between the model and human data when performing linear regression analysis (see Sec. 2.5).

2.4 *Statistical Analysis*

The model-based image quality metrics (CNR, CNRA, d'_{NPW}, d'_{NPWE}, d'_{NPWi}, d'_{NPWEi}, d'_{CHO}, and d'_{CHOi}) were transformed (along with their corresponding confidence intervals) to detection accuracy, A_x, using

$$A_x = \Phi\left(\frac{x}{\sqrt{2}}\right), \tag{21}$$

where x is the image quality metric of interest and $\Phi()$ is the standard normal cumulative distribution function. The results from the 2AFC perception experiment were compared with the detection accuracy as predicted by each observer model using linear regression analysis. The goal of the analysis was to assess the models with respect to the criteria given in Sec. 1 (i.e., highly correlated with human performance, correctly characterizes images with varying noise and resolution properties, and reasonable to measure with a finite number of images). The Pearson and Spearman correlation coefficients (r_p and r_s, respectively) were used as a goodness of fit metrics. The slope of each regression line was normalized to unity for easier visual comparison.

In our experience, image quality metrics that do not consider changes in noise texture or resolution tend to predict a greater improvement in detection accuracy for iterative algorithms (compared with FBP) than manifest with human readers. The result is a situation where the model often predicts a pair of FBP and iterative image sets to have similar quality, but the humans performed better with the FBP images. This is an undesirable property for an observer model and is manifest by distinct separation of the FBP and iterative data on a human versus

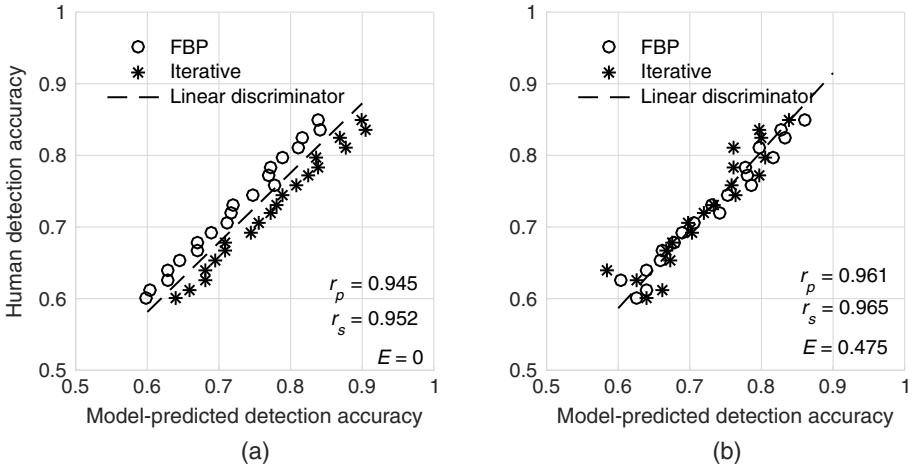

Fig. 3 Example of (a) a poor model and (b) a better model in terms of properly characterizing the effects of iterative reconstruction. FBP and iterative images with different detection accuracies as measured by human readers are predicted to have similar detection accuracy by the poor model. This results in data that are easily separable by a linear discriminator (i.e., low-error rate). In contrast, the data are not easily separable for the better model (i.e., high error rate). Note that both correlation coefficients were similar between the poor and better models in this example.

model regression plot. A demonstration of this scenario is shown in Fig. 3. For a good observer model, one would not expect the FBP and ADMIRE data to be easily separable by a linear discriminator. Thus, the error rate, E, of the linear classifier was used as a metric of how well each model observer can properly characterize different reconstruction algorithms (larger E implies a better model).

Finally, the average width of the 95% confidence intervals for each model's detection accuracy was computed to demonstrate how precisely the model performance can be estimated using a finite number of image realizations.[16] This value is important because it speaks to the minimum effect magnitude that could be detected using the observer model. For example, if the confidence interval is 10%, one would likely not be able to observe effects of less than 10% using that observer model with that corresponding number of images. In practice, we have found that many factors of interest (e.g., dose or iterative algorithms) affect detection accuracy in a relatively subtle manner. All computations were done in MATLAB (Mathworks).

3 Results

From the 2AFC experiment, the human accuracy ranged from about 50% (i.e., guessing) to 87%. On average, $CI_{95\%}$ (representing interobserver variability) was $\pm 7\%$. In general, performance increased with increasing contrast, dose, and signal size (Fig. 4). Based on the linear mixed effects model, ADMIRE increased detection accuracy compared with FBP ($P < 0.001$). The internal noise proportionality constants were found to be approximately 30 and 4.5 for α_{NPW} and α_{CHO}, respectively. From the linear regression analysis (Fig. 5), Pearson (Spearman) correlation was 0.36 (0.33), 0.83 (0.84), 0.84 (0.86), 0.86 (0.88), 0.86 (0.91), 0.88 (0.90), 0.85 (0.89), and 0.87 (0.84) for CNR, CNRA, NPW, NPWE, NPWi, NPWEi, CHO, and CHOi, respectively. The linear discriminator error was 0.25, 0.15, 0.2, 0.25, 0.3, 0.25, 0.4, and 0.45, and the magnitude of the 95% confidence intervals of the model's detection accuracies was 2.84×10^{-3}, 5.29×10^{-3}, 4.91×10^{-3}, 4.55×10^{-3}, 2.16×10^{-3}, 1.24×10^{-3}, 4.58×10^{-2}, and 7.95×10^{-2} for CNR, CNRA, NPW, NPWE, NPWi, NPWEi, CHO, and CHOi, respectively (Fig. 6). The correlations were statistically significant (95% significance level) for all models except CNR ($P = 0.1$ for CNR and $P < 0.001$ for all other models).

4 Discussion

The data show how CNR and CNRA are inadequate metrics of image quality. In fact, CNR had no statistically significant correlation with human performance, probably due to the fact that it is not task specific. Although CNRA had a relatively high correlation, it was found to be inadequate for properly assessing reconstruction algorithms that produce images with varying noise and resolution properties. As can be seen in Fig. 4, for the same CNRA, the human observers performed better with the FBP images compared with the ADMIRE images. Thus, the linear discriminator was able to separate those cases with few errors and CNRA should not be used to assess the impact of iterative reconstruction on low-contrast detectability.

As opposed to CNR, the NPW family of models (NPW, NPWE, NPWi, and NPWEi) had a higher correlation with humans and they were reasonable in characterizing FBP and ADMIRE images. Also, the error bars were small with the given number of image realizations. Further, the NPWEi model had the strongest observed correlation among all the models due to (1) the fact that it incorporated the eye filter, which attempts to weight the data by spatial frequencies for which the human observers are most sensitive, and (2) the fact that it incorporated internal noise, which attempts to model inconsistencies in the human perception of signals. Despite these encouraging results, the measurement of NPW model performance in the Fourier domain relies on some assumptions about the imaging system (quasi-linearity) and the noise statistics (stationarity). Under the conditions used in this study (uniform background with an SKE task), those assumptions appear to have been valid. However, for more complicated detection tasks, such as if the signal is only known statistically or if the background is inhomogeneous, those assumptions may not be valid and thus these Fourier-based metrics may not be

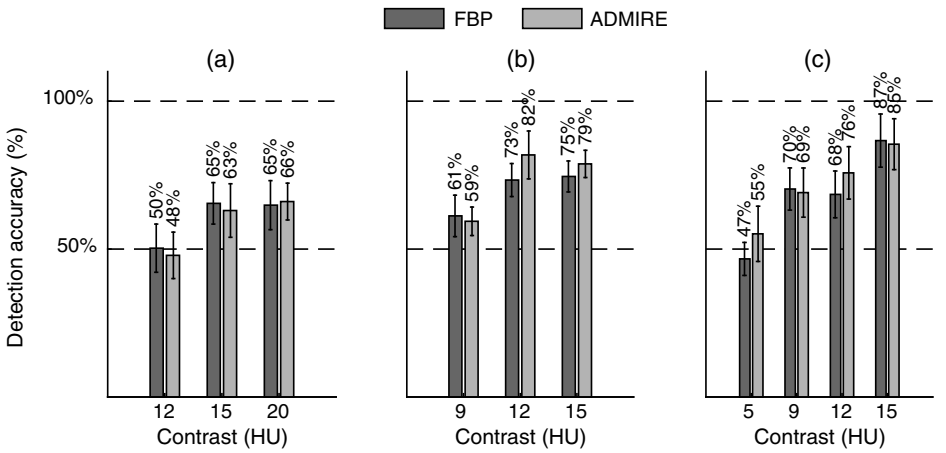

Fig. 4 Results of the 2AFC detection experiment showing detection accuracy versus contrast for the (a) 2 mm, (b) 4 mm, and (c) 6 mm insert sizes.

as strongly correlated with human performance. Also, for the CNR and NPW-based calculations, the nominal contrast was used as opposed to an estimated contrast measured from the reconstructed images. This was done because we have empirically found that the nominal contrast to be a good approximation of measured contrast (for large signals). For the smaller signals, the measured contrast can be reduced due to spatial blurring. CNR is typically measured in a large signal thus avoiding the effect of spatial blurring on contrast. Because of this, using the nominal contrast is more consistent with what CNR values typically represent. In defining the task function for the NPW-based calculations, it was actually necessary to use the nominal contrast because the task function by definition describes an idealized (i.e., preimaging) version of the object to be imaged. It should be noted that using the nominal contrast helps to narrow the confidence intervals of CNR, CNRA, and NPW-based measures. However, if the nominal contrast was not representative of the image contrast, one would expect this to be negatively impact the model's correlation with human performance.

The CHO model demonstrated a high correlation with human performance, and this correlation was further improved by incorporating internal noise in the CHOi model. Also, these models properly assessed FBP and ADMIRE images as demonstrated by large errors of the linear discriminators on the regression plots ($E = 0.4$ and 0.45 for CHO and CHOi, respectively). The downside of these models is the relatively large magnitude of the error bars for the number of images used (approximately 5% to 8% on average). This implies that for this number of images, it would be difficult to precisely assess two different conditions that had a difference in human performance of less than 5%. The CHO models have wider confidence intervals compared with the NPW models in this study due to the fact that the assumptions made in conjunction with the NPW models (nominal contrast, noise stationarity, and a quasi-linear shift-invariant system) allowed us to estimate the NPW model's performance precisely. In contrast, the large number of parameters in CHO model must be estimated (expected channel outputs and covariances) directly from the data, which propagates into relatively high uncertainty for estimating the model's performance. Thus, despite the overall strong correlation of the CHOi model with human observers, it may not be practical to acquire the relatively large ensemble of images needed to compute it, especially when attempting to optimize a clinical protocol over a large parameter space (e.g., dose, reconstruction algorithm, kernel, slice thickness, and so on). The major advantage of the CHO and CHOi models is that they do not assume a linear system or stationary noise statistics. Thus, they can confidently be used for more complicated detection tasks (i.e., inhomogeneous backgrounds) with highly nonlinear systems, provided that a large number of images for such evaluations are available.[3] Unfortunately, acquiring a large number of image realizations is not always feasible when assessing commercial CT systems, and simulation techniques, while extremely valuable, cannot always properly simulate proprietary components of a commercial system (e.g., tube current modulation, beam hardening corrections, reconstruction algorithms, and so on). It should be noted that recent work on search capable models has been shown to provide better statistical power with the same number of images compared with SKE CHO observers.[8–10] In this study, search-capable observers were not considered because the human reader results were conducted using the SKE paradigm.

Also, CHO performance for each imaging condition was computed using the method described by Wunderlich et al.[41] This method estimates the model's d' by first estimating the CHO parameters (i.e., expected difference in channel outputs and covariance matrices) using the entirety of available data and then computing d' based on an analytical function of those parameters. An alternative method would be to use a training/testing technique in which the model's template is estimated with a portion of the data and then applied to the remaining data [see Eq. (15)] to get distributions of the CHO test statistic for signal-present and signal-absent cases. These distributions are then used to estimate d' [see Eq. (1)].[42] The advantage of using this training/testing method would be that the estimated model performance would be less susceptible to over-fitting biases. The tradeoff is that the precision of the estimate would likely be degraded. With this being said, both methods to estimate CHO performance are used in recent literature.

In comparing the NPW and CHO models, it should also be noted that the NPW models were calculated based on measured CT system parameters (e.g., NPS/MTF) and their corresponding d' values were computed based on analytical equations relating the NPS/MTF to d' in the Fourier domain. In contrast, d' for the CHO models was estimated directly from the ensemble of

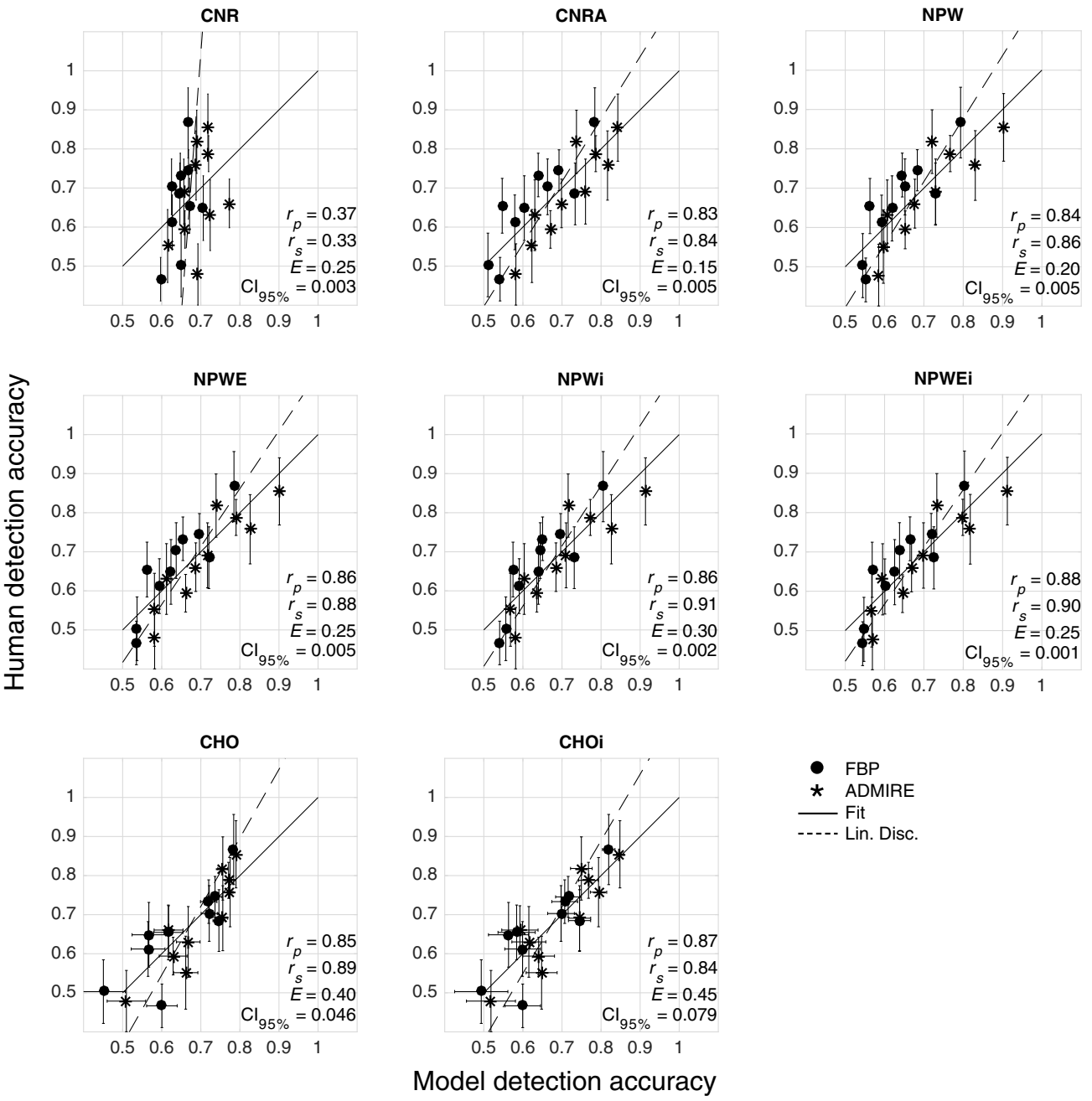

Fig. 5 Regression plots of human detection accuracy versus the accuracy as predicted by the model observers. The solid line is the linear regression fit and the dotted line is the linear discriminator that best separates the FBP (circles) and ADMIRE (stars) cases.

signal-present/signal-absent images. This is consistent with the approaches generally used for each of these models based on recent literature,[6,9,20–26] and thus, we attempted to be consistent with those practices. As a result, the data from this study represent a comparison between not only different observer models but also between different approaches and assumptions made to estimate the models' performances. It would be possible to estimate the NPW performance directly from the signal-present/signal-absent images in a similar fashion to the CHO models. One would not expect the results to change significantly but the size of the confidence intervals would likely increase.

This study used a linear discriminator error, E, to evaluate how well different models properly assess images from different reconstruction algorithms. This metric was devised to be sensitive to the situation illustrated in Fig. 3, which is often observed when simple image quality metrics, such as CNR, are used to compare image quality across images having known differences in noise–magnitude, noise–texture, and resolution. However, E alone should not be interpreted as a comprehensive validation of a model. In other words, a model with low E is likely a poor model, but a model with high E could be poor or good. The three metrics noted in this work (correlation, linear discriminator error, and magnitude of confidence intervals) should be

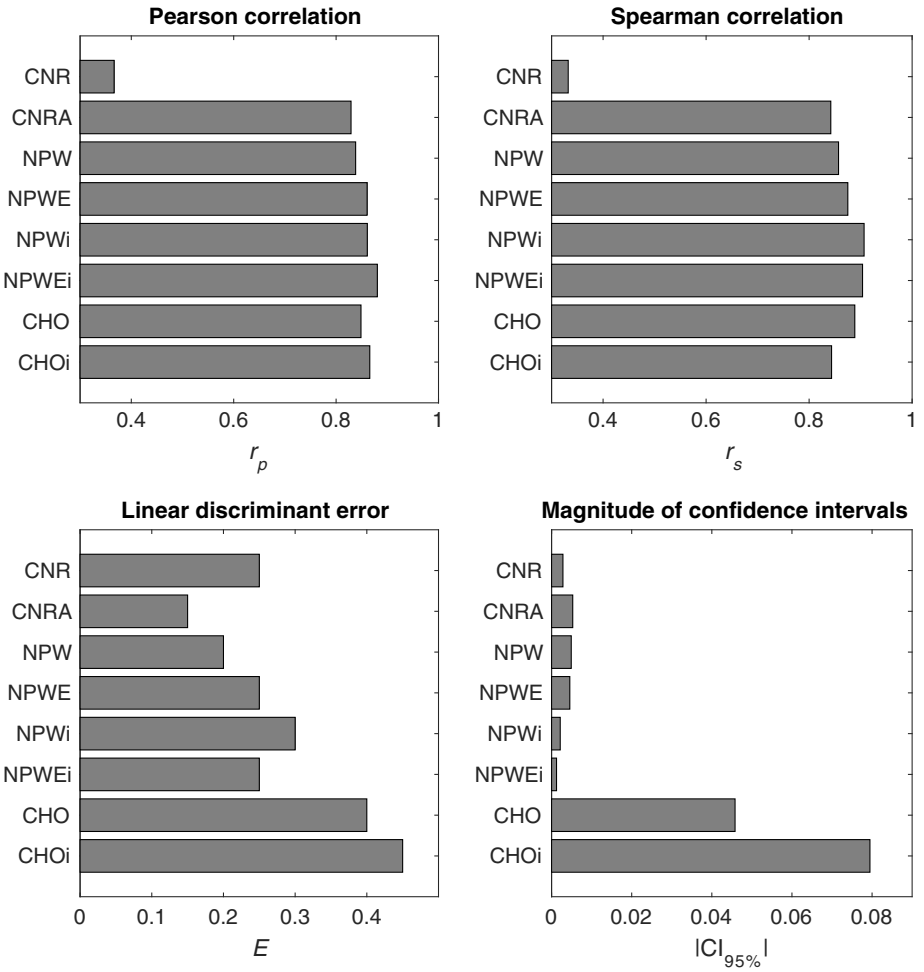

Fig. 6 Bar plots showing the summary statistics for each model.

considered jointly when assessing the validity of an observer model. To our knowledge, there is no standard metric available in the literature to capture the phenomenon illustrated in Fig. 3. Future work may expand on this metrology with the use of nonlinear classifiers or *k*-means clustering methods and further robustness test could be performed using Monte Carlo methods.[43,44]

Some limitations of this study should be acknowledged. First, the task examined was relatively simple with a uniform background and geometric 2-D signals under the SKE paradigm. Real clinical tasks deviate from this simplistic paradigm in several ways, including the fact that patients are not uniform, lesions can have complex 3-D shapes, and the radiologists need to perform a visual search when making a diagnosis. This oversimplified paradigm can sometimes lead to spurious optimizations where quantum noise is reduced as much as possible (often at the expense of resolution). In reality, a single image is used for multiple tasks and thus minimizing quantum noise may be good for a specific low-contrast detection task but might degrade performance for other tasks. Future work will be focused on updating this methodology to make it more clinically realistic. Second, only one dose level was considered and thus it was not possible to assess if each observer model properly characterized the effects of changing dose. Third, the human detection data were relatively noisy (i.e., large error bars). This is due to the fact that a limited number of trials [Eq. (15)] were performed for each condition. Ideally, many more trials would be used to minimize intraobserver variability. However, as stated in Sec. 2, the data used for this study represent a subset of data from a previous study.[27] In that previous study, a larger image parameter space was considered (compared with this current study). Because of this, the perception experiments done in that previous study were limited by the total number of images that could reasonably be shown to the observers. Finally, the internal noise parameters were chosen to maximize correlations between the models and humans, which means that the reported correlation coefficients are probably about as large as possible. It is possible that if other criteria were used for optimization (e.g., mean square error), the internal noise parameters may have been different.

5 Conclusion

The findings of this study imply that the NPW and CHO families of model observers provided strong correlation with human observer performance and correctly characterized the differences in image quality of FBP and iteratively reconstructed images. Thus, these models are good candidates to be used to help optimize CT scan protocols in terms of low-contrast detectability. Future work is needed to compare a broader range of

observer models including visual search and signal/background variability.

Acknowledgments

Dr. Samei reports grants from Siemens Medical Solutions during the conduct of the study.

References

1. C. H. McCollough et al., "Achieving routine submillisievert CT scanning: report from the summit on management of radiation dose in CT," *Radiology* **264**, 567–580 (2012).
2. ICRU, *ICRU Report 54: Medical Imaging-The Assessment of Image Quality*, International Commission on Radiation Units and Measurements, Bethesda, Maryland (1995).
3. H. H. Barrett et al., "Task-based measures of image quality and their relation to radiation dose and patient risk," *Phys. Med. Biol.* **60**, R1 (2015).
4. C. Abbey and F. Bochud, "Modeling visual detection tasks in correlated image noise with linear model observers," in *Handbook of Medical Imaging*, R. L. Van Metter, J. Beutel, and H. L. Kundel, Eds., Vol. **1**, Physics and Psychophysics, SPIE Press, Bellingham, Washington (2000).
5. H. H. Barrett et al., "Model observers for assessment of image quality," *Proc. Natl. Acad. Sci. U. S. A.* **90**, 9758–9765 (1993).
6. J. Solomon, J. Wilson, and E. Samei, "Characteristic image quality of a third generation dual-source MDCT scanner: noise, resolution, and detectability," *Med. Phys.* **42**, 4941–4953 (2015).
7. L. Yu et al., "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.* **40**, 041908 (2013).
8. H. C. Gifford, Z. Liang, and M. Das, "Visual-search observers for assessing tomographic x-ray image quality," *Med. Phys.* **43**, 1563–1575 (2016).
9. S. Leng et al., "Correlation between model observer and human observer performance in CT imaging when lesion location is uncertain," *Med. Phys.* **40**, 081908 (2013).
10. L. M. Popescu and K. J. Myers, "CT image assessment by low contrast signal detectability evaluation with unknown signal location," *Med. Phys.* **40**, 111908 (2013).
11. M. P. Eckstein and C. K. Abbey, "Model observers for signal-known-statistically tasks (SKS)," *Proc. SPIE* **4324**, 91 (2001).
12. M. P. Eckstein, C. K. Abbey, and F. O. Bochud, "Visual signal detection in structured backgrounds. IV. Figures of merit for model performance in multiple-alternative forced-choice detection tasks with correlated responses," *J. Opt. Soc. Am. A* **17**, 206 (2000).
13. F. O. Bochud, C. K. Abbey, and M. P. Eckstein, "Visual signal detection in structured backgrounds. III. Calculation of figures of merit for model observers in statistically nonstationary backgrounds," *J. Opt. Soc. Am. A* **17**, 193 (2000).
14. M. P. Eckstein, A. J. Ahumada, Jr., and A. B. Watson, "Visual signal detection in structured backgrounds. II. Effects of contrast gain control, background variations, and white noise," *J. Opt. Soc. Am. A* **14**, 2406 (1997).
15. M. P. Eckstein and J. S. Whiting, "Visual signal detection in structured backgrounds. I. Effect of number of possible spatial locations and signal contrast," *J. Opt. Soc. Am. A* **13**, 1777 (1996).
16. C. K. Abbey, H. H. Barrett, and M. P. Eckstein, "Practical issues and methodology in assessment of image quality using model observers," *Proc. SPIE* **3032**, 182 (1997).
17. C. K. Abbey, M. P. Eckstein, and F. O. Bochud, "Estimation of human-observer templates in two-alternative forced-choice experiments," *Proc. SPIE* **3663**, 284 (1999).
18. M. E. Baker et al., "Contrast-to-noise ratio and low-contrast object resolution on full- and low-dose MDCT: SAFIRE versus filtered back projection in a low-contrast object phantom and in the liver," *Am. J. Roentgenol.* **199**, 8–18 (2012).
19. F. Holmquist et al., "Impact of iterative reconstructions on image noise and low-contrast object detection in low kVp simulated abdominal CT: a phantom study," *Acta Radiol.* **57**(9), 1079–1088 (2015).
20. B. Chen et al., "Evaluating iterative reconstruction performance in computed tomography," *Med. Phys.* **41**, 121913 (2014).
21. O. Christianson et al., "An improved index of image quality for task-based performance of CT iterative reconstruction across three commercial implementations," *Radiology* **275**, 725–734 (2015).
22. G. J. Gang et al., "Analysis of Fourier-domain task-based detectability index in tomosynthesis and cone-beam CT in relation to human observer performance," *Med. Phys.* **38**, 1754 (2011).
23. J. M. Wilson et al., "A methodology for image quality evaluation of advanced CT systems," *Med. Phys.* **40**, 031908 (2013).
24. C. K. Abbey and H. H. Barrett, "Human- and model-observer performance in ramp-spectrum noise: effects of regularization and object variability," *J. Opt. Soc. Am. A* **18**, 473 (2001).
25. B. L. Eck et al., "Computational and human observer image quality evaluation of low dose, knowledge-based CT iterative reconstruction," *Med. Phys.* **42**, 6098–6111 (2015).
26. Y. Zhang et al., "Correlation between human and model observer performance for discrimination task in CT," *Phys. Med. Biol.* **59**, 3389 (2014).
27. J. Solomon et al., "Diagnostic performance of an advanced modeled iterative reconstruction algorithm for low-contrast detectability with a third-generation dual-source multidetector CT Scanner: potential for radiation dose reduction in a multireader study," *Radiology* **275**, 735–745 (2015).
28. A. E. Burgess et al., "Efficiency of human visual signal discrimination," *Science* **214**, 93–94 (1981).
29. A. E. Burgess, "Visual perception studies and observer models in medical imaging," *Semin. Nucl. Med.* **41**, 419–436 (2011).
30. S. Richard et al., "Towards task-based assessment of CT performance: system and object MTF across different reconstruction algorithms," *Med. Phys.* **39**, 4115 (2012).
31. B. Chen et al., "Assessment of volumetric noise and resolution performance for linear and nonlinear CT reconstruction methods," *Med. Phys.* **41**, 071909 (2014).
32. J. Solomon and E. Samei, "Quantum noise properties of CT images with anatomical textured backgrounds across reconstruction algorithms: FBP and SAFIRE," *Med. Phys.* **41**, 091908 (2014).
33. A. E. Burgess, X. Li, and C. K. Abbey, "Visual signal detectability with two noise components: anomalous masking effects," *J. Opt. Soc. Am. A* **14**, 2420 (1997).
34. M. Eckstein et al., "Automated computer evaluation and optimization of image compression of x-ray coronary angiograms for signal known exactly detection tasks," *Opt. Express* **11**, 460 (2003).
35. M. Ishida et al., "Digital image processing: effect on detectability of simulated low-contrast radiographic patterns," *Radiology* **150**, 569–575 (1984).
36. R. S. Saunders, Jr. and E. Samei, "Resolution and noise measurements of five CRT and LCD medical displays," *Med. Phys.* **33**, 308 (2006).
37. 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.* **35**, 5043 (2008).
38. L. Yu et al., "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.* **40**, 041908 (2013).
39. Y. Zhang, B. T. Pham, and M. P. Eckstein, "The effect of nonlinear human visual system components on performance of a channelized Hotelling observer in structured backgrounds," *IEEE Trans. Med. Imaging* **25**, 1348–1362 (2006).
40. A. Wunderlich and F. Noo, "Image covariance and lesion detectability in direct fan-beam x-ray computed tomography," *Phys. Med. Biol.* **53**, 2471 (2008).
41. A. Wunderlich et al., "Exact confidence intervals for channelized Hotelling observer performance in image quality studies," *IEEE Trans. Med. Imaging* **34**, 453–464 (2015).
42. A. Ba et al., "Anthropomorphic model observer performance in three-dimensional detection task for low-contrast computed tomography," *J. Med. Imaging* **3**, 011009 (2016).
43. H. C. Gifford et al., "LROC analysis of detector-response compensation in SPECT," *IEEE Trans. Med. Imaging* **19**, 463–473 (2000).
44. H. C. Gifford et al., "Channelized Hotelling and human observer correlation for lesion detection in hepatic SPECT imaging," *J. Nucl. Med.* **41**, 514 (2000).

Justin Solomon received his doctoral degree in medical physics from Duke University in 2016 and is currently a medical physicist in the Clinical Imaging Physics Group (CIPG) at Duke University Medical Center's Radiology Department. His expertise is in x-ray computed tomography imaging and image quality assessment.

Ehsan Samei is a tenured professor at Duke University and the director of the Duke Medical Physics Graduate Program and the CIPG. His interests include clinically relevant metrology of imaging quality and safety for optimum interpretive and quantitative performance. He strives to bridge the gap between scientific scholarship and clinical practice by meaningful realization of translational research and the actualization of clinical processes that are informed by scientific evidence.