Objective assessment of image quality and dose reduction in CT iterative reconstruction

J. Y. Vaishnav[a)] and W. C. Jung
Diagnostic X-Ray Systems Branch, Office of In Vitro Diagnostic Devices and Radiological Health,
Center for Devices and Radiological Health, United States Food and Drug Administration,
10903 New Hampshire Avenue, Silver Spring, Maryland 20993

L. M. Popescu, R. Zeng, and K. J. Myers
Division of Imaging and Applied Mathematics, Office of Science and Engineering Laboratories,
Center for Devices and Radiological Health, United States Food and Drug Administration,
10903 New Hampshire Avenue, Silver Spring, Maryland 20993

(Received 2 November 2013; revised 11 May 2014; accepted for publication 15 May 2014; published 10 June 2014)

Purpose: Iterative reconstruction (IR) algorithms have the potential to reduce radiation dose in CT diagnostic imaging. As these algorithms become available on the market, a standardizable method of quantifying the dose reduction that a particular IR method can achieve would be valuable. Such a method would assist manufacturers in making promotional claims about dose reduction, buyers in comparing different devices, physicists in independently validating the claims, and the United States Food and Drug Administration in regulating the labeling of CT devices. However, the nonlinear nature of commercially available IR algorithms poses challenges to objectively assessing image quality, a necessary step in establishing the amount of dose reduction that a given IR algorithm can achieve without compromising that image quality. This review paper seeks to consolidate information relevant to objectively assessing the quality of CT IR images, and thereby measuring the level of dose reduction that a given IR algorithm can achieve.
Methods: The authors discuss task-based methods for assessing the quality of CT IR images and evaluating dose reduction.
Results: The authors explain and review recent literature on signal detection and localization tasks in CT IR image quality assessment, the design of an appropriate phantom for these tasks, possible choices of observers (including human and model observers), and methods of evaluating observer performance.
Conclusions: Standardizing the measurement of dose reduction is a problem of broad interest to the CT community and to public health. A necessary step in the process is the objective assessment of CT image quality, for which various task-based methods may be suitable. This paper attempts to consolidate recent literature that is relevant to the development and implementation of task-based methods for the assessment of CT IR image quality. © *2014 American Association of Physicists in Medicine*. [http://dx.doi.org/10.1118/1.4881148]

Key words: iterative reconstruction, dose reduction, computed tomography

1. INTRODUCTION

With the increasing clinical utilization of diagnostic CT imaging, individual and population doses of ionizing radiation have become a public health concern.[1] Academia, industry, and government have responded with efforts to reduce the radiation dose required to obtain diagnostic-quality images. One step has been CT manufacturers implementing iterative reconstruction (IR) methods that for certain clinical tasks can improve dose efficiency over the conventional reconstruction method, filtered back projection (FBP).

In contrast to FBP which reconstructs an image from projections of an object in a single step, IR algorithms estimate the image via multiple passes over the projection data. Algebraic IR algorithms update the image repeatedly until its projections match the actual data as closely as possible while satisfying other constraints, e.g., piecewise image smoothness. Statistical iterative algorithms make use of models of the image acquisition process, including details about the imaging hardware and measurement noise. IR algorithms generally involve operations resulting in nonlinear processing, like imposition of a positivity constraint or adaptive, signal-dependent smoothing. Relative to FBP images, IR images thus have noise that is highly spatially dependent, and image resolution that depends on contrast. These dependences complicate the assessment of image quality for images obtained by IR, since many traditional metrics of image quality involve noise and resolution. An assessment of image quality is necessary to measure dose reduction.

A standardizable method of measuring the dose reduction achievable by a particular IR algorithm would enable buyers to compare commercial algorithms and users to independently validate manufacturer claims. It would also provide additional information that industry could use to market their CT devices, motivating them to continue optimizing their IR

0094-2405/2014/41(7)/071904/12/$30.00

TABLE I. IR algorithms with associated dose reduction claims cleared by USFDA as of March 2014. The Siemens SAFIRE (Ref. 6), Philips IMR (Ref. 5), and GE ASiR-V (Ref. 3) each claim specific percentages of dose reduction, though the claims have disclaimers attached.

Manufacturer	Iterative reconstruction	Labeling claims cleared by USFDA
General Electric	ASiR, MBIR (Veo), ASiR-V	For ASiR-V, dose reduction by 50%–82%, 59%–135% low-contrast detectability improvement, and noise reduction up to 91% (Ref. 3)
Toshiba	AIDR, AIDR+, AIDR3D	Dose reduction (Ref. 4)
Philips	IRT, IMR	For IMR, 60%–80% lower dose along with 43%–80% low-contrast detectability improvement and 70%–83% less image noise (Ref. 5)
Siemens	SAFIRE, iTRIM (cardiac)	For SAFIRE, 54%–60% dose reduction (Ref. 6)

algorithms and protocols. Developing such a method is therefore of interest not only to the entire CT community, but to public health in general.

Table I lists some IR algorithms commercially available in the United States, as well as any dose reduction claims that the United States Food and Drug Administration (USFDA) has cleared. Two of the cleared claims are qualitative and do not claim a specific percentage by which they reduce dose; the other two include quantitative claims.

Dose reduction always comes at the expense of increased quantum noise. This paper lays out approaches to assessing the quality of quantum-limited CT images. Note that many other factors, including artifacts and anatomical structure, can adversely affect image quality, and developing image quality metrics that account for such factors is a subject of ongoing research.

A method of validating dose reduction claims must take into account clinical considerations, minimize possible sources of bias, and be least burdensome to manufacturers. A practical method is task-based, requiring (1) a task to be performed, (2) an observer to perform the task, and (3) a way of measuring observer performance on the task.[2] This paper outlines a range of possible tasks, observers, and measures of observer performance on the task. To be representative of clinical scenarios, task-based assessments should be performed under a variety of conditions, e.g., different reconstruction kernels, dose levels, etc. As a result, these studies can result in a large number of images and image evaluations, making model observers a valuable tool for CT image quality assessment. While human observers are certainly a possibility, in addition to being more cost-effective, model observers are immune to effects like reader learning, fatigue, and inter- and intraobserver variability that can be issues for humans.

This paper is organized as follows: Sec. 1 provides background information on the challenges in evaluating IR images. Section 2 discusses task-based image quality assessment as a way of dealing with these challenges. Section 3 lays out model observers commonly used for these tasks, and Sec. 4 discusses the calculation of observer performance using different reconstruction algorithms, which provides a metric of image quality and dose reduction.

1.A. Limitations of conventional and Fourier-based image quality metrics for assessing IR images

Below we discuss some commonly used metrics of image quality, and why they are not always sufficient to assess IR images.

1.A.1. Pixel noise and noise power spectrum (NPS)

Pixel noise is the standard deviation of a pixel over an ensemble of images. It is commonly estimated over a region of interest (ROI) in a single image, employing assumptions of ergodicity and stationarity that may not be valid. Additionally, due to correlations introduced by the reconstruction process, CT noise is textured and highly nonstationary (see Fig. 1). Since image texture can impact diagnostic quality, and pixel noise does not contain information about these correlations, it is not an acceptable metric of image quality.[7–9]

The NPS, which contains the spatial frequency content of the noise in an image, characterizes the noise correlations in Fourier space. The definition of NPS relies on the assumption of wide-sense stationary noise (see p. 400 of Ref. 2). This assumption is generally violated by CT IR images. The NPS is therefore of limited utility in describing the noise or assessing the quality of CT IR images, though it does contain more information than the pixel noise and can be useful for FBP images.

1.A.2. Contrast to noise ratio (CNR)

The CNR is a useful metric for describing the signal amplitude relative to the ambient noise for simple and largely homogeneous objects. However, the CNR depends only on contrast and noise. Actual signal detectability also depends on factors including signal size, shape, and density distribution; background level, variability, and correlation; the variance and covariance of measurement noise; spatial resolution; and the observer and detection strategy used. The CNR can be useful in some simple situations, e.g., determining thresholds of contrast agents at which signals on a test phantom become

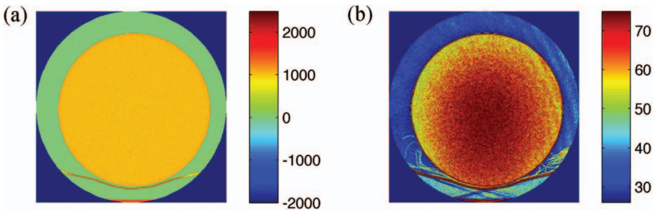

FIG. 1. As CT noise is not stationary, the properties of the noise depend strongly on the phantom radius. (a) CT number uniformity assessment module of the ACR CT Accreditation GAMMEX 464 phantom. Reconstruction via FBP. (b) Pixel standard deviation, calculated over a set of 56 images of the phantom. Images are in HU. The noise in (b) varies both radially and angularly. The centroid location was constant over the image set, suggesting that the angular variation results from beam hardening due to the phantom bed [visible at the bottom of (a)].

visible. However, the CNR is in general not a complete description of an observer's ability to detect lesions, and this is even more true for IR images which are more likely to be nonstationary and textured.

1.A.3. Modulation transfer function (MTF)

Another common metric of image quality is the MTF, which USFDA accepts as one component of performance validation for solid state digital detectors.[10] The MTF, related to image resolution, is the Fourier transform of the point spread function (PSF), the system's response to a point object. The MTF and PSF are useful quantities for linear, shift-invariant systems, where the imaging system's response to an arbitrary object can be determined by convolving the true object and the point spread function. This is not true for nonlinear IR algorithms or even for FBP when nonlinear filters are used. For nonlinear systems, the MTF—whose definition assumes independence of location in the image, dose, contrast, and algorithm parameters—acquires dependence on these quantities.[11] As the MTF is not well defined for images reconstructed by IR, it is of limited utility in assessing the quality of these images.

The above Fourier-based metrics of image quality assessment are convenient and useful in many contexts; e.g., in monitoring changes as part of routine quality assurance checks on instruments. However, assessment of IR images requires more sophisticated methods which do not rely on assumptions of system linearity and noise stationarity.

1.A.4. Modified Fourier metrics

One approach to assessing CT IR images has involved modifying the traditional Fourier metrics. Table II provides references to recent work that expands the definitions of the MTF, NPS, and other traditional metrics. Reconstruction filters often improve resolution at the expense of increasing noise, and vice versa. Moreover, IR typically yields images with different noise textures (correlations) than FBP. A remaining challenge to Fourier-based approaches is developing a single image quality metric that predicts the joint impact of noise, resolution, and texture on task performance.

1.B. Task-based assessment of CT images

An alternative to the Fourier methods described in Sec. 1.A is task-based image quality assessment. Task-based methods involve (1) defining a task suitably representative of a clinical task, (2) selecting an observer to perform that task (either a computational "model" observer or an actual human reader), and (3) evaluating the observer's task performance, which provides an objective metric of image quality. For a task-based method to be useful in practice, it should allow determination of uncertainties, require a practical number of images, and be readily standardized.

While CT data are generally 2D or 3D, for mathematical simplicity, the data set is typically flattened into a 1D vector $\mathbf{g}$ with M elements. Generally, $\mathbf{g}$ represents the raw projection data that a reconstruction algorithm operates on to create a viewable image. For a detailed discussion of image formation, see Barrett and Myers.[2]

Let us define the reconstructed image obtained by applying a particular reconstruction operator $\mathscr{R}$ as

$$\mathbf{g}^{(\mathscr{R})} = \mathscr{R}\mathbf{g}. \tag{1}$$

$\mathscr{R}$ can represent FBP or an IR technique. Our goal is to compare the quality of the images obtained via different $\mathscr{R}$.

An "observer" uses the reconstructed data $\mathbf{g}^{(\mathscr{R})}$ to generate a scalar quantity $\lambda(\mathbf{g}^{(\mathscr{R})})$ called a test statistic, which is a function of the reconstruction algorithm. An observer can be either a human or an algorithm ("model observer"). The use of human observers is costly and sometimes impractical; humans are also subject to effects like fatigue or reader learning that introduce bias. These factors make the use of model observers attractive. A large and growing literature demonstrates the use of model observers for evaluating systems and predicting human performance.

TABLE II. Papers extending Fourier metrics, or examining situations where they break down.

Ref.	Summary
12	Pineda *et al.* introduce a local NPS. Additionally, they use the off-diagonal elements of the covariance matrix of Σ_{DFT}, the discrete Fourier transform of noise-only images, as a measure of stationarity. The authors calculate the local NPS for 3D CBCT images, and demonstrate its spatial variation. They also analyze the off-diagonal elements of Σ_{DFT}, to check the validity of the NPS (which assumes that those elements are zero). Although this paper deals specifically with images reconstructed via the Feldkamp algorithm, the results and metrics are of general interest.
13	Evans *et al.* use a simulation of "clock phantom" data to measure the MTF via the edge spread function of a disk. The paper presents noise-resolution curves for the alternative minimization (AM) IR algorithm, and compares AM to FBP. See also the actual experiments performed in Ref. 11, which provide related results on resolution. The authors demonstrate the dependence of the edge spread function and MTF on contrast and dose.
11	Richard *et al.* develop an alternative MTF (denoted MTF_T) based on a task transfer function. The measurement of MTF_T uses the ACR (Model 464) phantom to determine the edge spread function of a disk. The authors evaluate the MTF_T for different IR algorithms, finding that it is a function of dose and contrast.
14	Brunner *et al.* present another alternative to the MTF, based on the object transfer matrix and the covariance matrix. The authors use their methodology to assess a lab cone-beam CT system. This paper deals with the Feldkamp algorithm. Note that for nonlinear iterative algorithms, the MTF can depend on contrast, dose, or reconstruction parameters.
15	Gang *et al.* extend Fourier metrics as well as detectability measures to tomosynthesis and CBCT, using a 3D cascaded systems model to generate a generalized detectability index. This paper deals with the Feldkamp algorithm, and not with iterative algorithms.

TABLE III. Recent results relevant to objective assessment of image quality for IR algorithms in CT.

Ref.	Summary
16	Hara *et al.* compare ASiR to FBP. In addition to a rating scale experiment on clinical images performed by two human observers, they present results on pixel noise, low-contrast resolution, and spatial resolution. Due to the limitations of these performance metrics for evaluating IR images, USFDA recommends other methods for validation of quantitative dose reduction claims; however, the paper provides a useful general study on dose reduction.
17	Miéville *et al.* comparatively evaluate ASiR, Veo, and iDose^4 via an analysis of physical metrics like CT number accuracy, pixel noise, NPS, and MTF on *Catphan 600* phantom images. They also perform a four-alternative forced-choice (4AFC) experiment using pediatric phantoms with five readers.
18	Boedeker and McNitt-Gray use a phantom and nonprewhitening (NPW) model observer based on frequency-based metrics (NPS and MTF) to test the influence of reconstruction kernel and dose on signal to noise ratio (SNR). The task was a simple detection task with a simulated signal. They find that while NPS and NEQ can be implemented in modern CT, NEQ is not a good metric for certain nontraditional reconstruction filters or in helical mode.
19	Hernandez-Girón *et al.* assess low-contrast sensitivity using the low contrast module of the *Catphan* 600 phantom. The purpose of this study was to develop and validate software for automated objective low contrast detectability studies based on a model observer. The study compares the model observers with human observers with favorable results. The authors propose assessment of low-contrast sensitivity via the *Catphan* 600 phantom as a method for image quality assessment. In Sec. 2.A, we discuss shortcomings of commercial phantoms for the validation of quantitative dose reduction claims.
20	Wunderlich and Noo outline methods for receiver-operator characteristic (ROC) analysis, applicable to (for example) linear model observers applied to binary classification low-contrast lesion detection tasks. They test their methods on CT images. The authors introduce new confidence interval estimators for ROC curves and summary measures.
21	Yu *et al.* predict human observer performance on two-alternative forced choice (2AFC) lesion-detection tasks at various dose levels, using FBP and Siemens SAFIRE. The authors design a phantom for image quality assessment, and perform a 2AFC study using model and human observers. This study lays out methodologies that are potentially useful in designing a task-based study for validating quantitative dose claims.
22	Popescu and Myers lay out a custom phantom design that addresses issues with the *Catphan*. See Fig. 2. The authors suggest a low-contrast detectability study with unknown signal location for image quality assessment.

Finally, a figure of merit should be defined that evaluates the observer's performance at the task. We discuss this further in Sec. 4.A.

A choice of the task, observer, and figure of merit specify the design of a validation method for dose reduction claims. Table III discusses some recent task-based studies on image quality assessment in CT. Although none of these studies provides a direct prescription for validating quantitative dose reduction claims, each provides information that may be relevant in choosing the preferred study design for a given CT system and IR algorithm.

2. TASKS

The first part of task-based image quality assessment is defining a task that is suitably representative of a clinical task. Since the results of the task-based assessment should be standardizable, practical studies are phantom-based and require selecting a phantom. This paper focuses on signal detection and localization tasks, which are the most commonly used for CT image quality assessment; the signals represent pathological features in clinical images. However, many other possibilities exist for validating dose reduction claims, including estimation of a quantity (e.g., signal size, location, or intensity), discrimination of objects of different sizes and/or shapes, and other possibilities entirely.

2.A. Selecting a phantom

Detection and localization tasks generally require an observer to distinguish a ROI or image containing a signal from the remaining ROIs or images which do not contain signals. An appropriate phantom for CT image quality assessment minimizes extraneous clues to signal presence or location that might bias the observer and result. Potential issues could include the presence of stitching artifacts in ROIs, or differences in background levels or noise texture between ROIs being compared. For instance, if signal-absent ROIs are consistently selected from a different part of the phantom than signal-present ROIs, the correlation between the background noise texture and the presence and absence of a signal might result in a human reader being more easily able to distinguish the presence of a signal. Depending on the particular task, means of mitigating such issues may include rotating the phantom between acquisitions, defining ROIs randomly, or randomly rotating or flipping ROIs. If completely eliminating a particular source of extraneous information is impractical, the test methodology should be validated by demonstrating that the effect does not bias the reader.

Commercial CT image quality phantoms, such as the ACR accreditation phantom (Gammex, Middleton, WI) or *Catphan* 600 (Phantom Laboratories, Salem, NY) are useful in subjective evaluations of image quality, in which a human reader identifies size/contrast thresholds beneath which s/he is unable to locate signals. However, in such evaluations, the known location of signals in these phantoms can lead to subjective bias. Moreover, the tight spacing of the signals can necessitate that the multiple ROIs being compared in a task be drawn from different radial distances from the center of the phantom. Since CT noise is generally nonstationary (see Fig. 1), the ROIs being compared might have different noise distributions. The differing appearance of the ROIs, combined with the known signal locations, could provide clues to signal detection or localization, introducing bias. Commercial phantoms are not optimal for model observer studies, either. Within each set of signals on the phantom, the *Catphan* and ACR phantom each contain only one signal on the threshold of detectability. Hence, many images are required for training and testing a model observer, resulting in inefficiency. In view of the limitations of commercially available phantoms, the potential for using a phantom with fewer signals has been examined.[22] The additional space enables selection of large ROIs, in which the location of the signal in each ROI is random. See Fig. 2. Different phantom designs[21,23] and related methods have been used to evaluate IR image quality using human and model observers.

2.B. Designing an experiment

In order to compare different reconstruction algorithms' performance at different doses, whatever experiment is chosen will need to be carried out at a variety of dose levels and for each reconstruction algorithm being assessed. The figure of merit should be calculated for a given algorithm, at a given dose. Below we discuss forced choice and rating-scale experiments.

2.B.1. Forced-choice and rating-scale experiments

Forced-choice experiments are useful for human observers. The simplest forced-choice experiment is a 2AFC experiment. A reader is presented two images, one with a signal present, the other with no signal (see, e.g., Fig. 3). The reader must identify the image containing the signal, which can be at either a fixed or an unknown location. The percentage correct in a 2AFC experiment is the figure of merit.[24]

A generalization of 2AFC is the multialternative forced choice (MAFC) experiment. An observer is presented groups of $M > 2$ images (or ROIs selected from a single image) and must determine which contains the signal. As in 2AFC experiments, the figure of merit is the percentage correct.

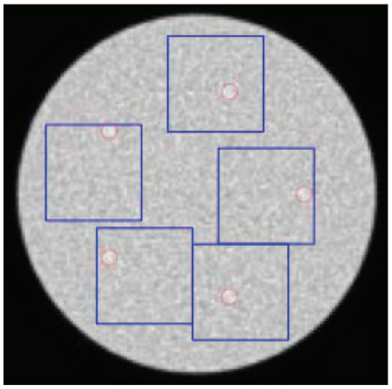

FIG. 2. Phantom design in Ref. 22. This design allows the selection of ROIs in a manner providing signal location uncertainty. The signals are also sufficiently widely spaced to allow for the selection of signal-absent ROIs with similar noise distributions to the signal-present ones.

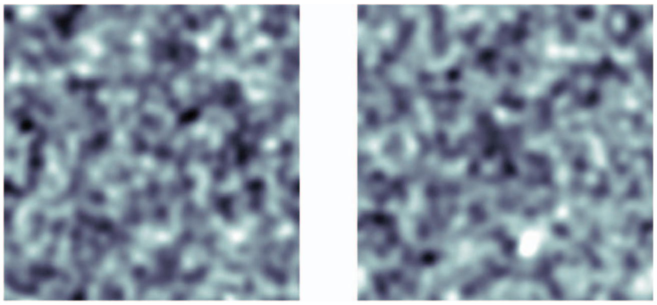

FIG. 3. A 2AFC task involves determining which of two images contains a signal. In this case, the observer must also indicate the signal location.

An alternative to forced-choice experiments are rating scale experiments, suitable for human or model observers. In rating scale experiments, the observer scores each image based on their confidence that a signal is present. For example, radiologists are often asked to rank images on a five-point scale such as 1 = disease definitely absent, 2 = disease probably absent, 3 = disease possibly present, 4 = disease probably present, 5 = disease definitely present.

2.B.2. The role of signal search

The location of the signal in the above experimental designs can be known or unknown. While most studies to date have used fixed signal locations, recent work[22] has proposed signal search methodologies.

2.B.2.a. Signal known exactly (SKE) tasks. In SKE experiments, the observer knows the signal location. The task is only to determine whether or not a given image contains the signal at that location. SKE studies are relatively simple to implement and analyze. As the observer knows exactly where to look, however, very low contrast signals are required to challenge the observer sufficiently to yield statistically significant performance information. Moreover, SKE detection-only experiments yield limited information regarding the frequency at which random occurrences of noise clumps throughout the image mimic signals. In clinical practice, where the location of a pathology is usually unknown, such noise clumps could result in false positives.

2.B.2.b. Search and free-response tasks. Due to the limitations of SKE tasks, recent studies examine the possibility of allowing for unknown signal location.[22,25–30] For forced-choice experiments, the figure of merit remains the percentage correct. More generally, the probability of correct joint detection and localization can be plotted vs the probability of false positives; the figure of merit is the area under this curve.

Another variant of search is free-response analysis. Here, one or more lesions are present in an image and the observer is asked to localize all lesions to ROIs, without prior knowledge of how many lesions are present. These results can be analyzed using either free response ROC (FROC) or exponential transformation of FROC (EFROC) methods.[31]

Experiments involving signal location uncertainty offer a more realistic simulation for many clinical applications than studies where signal location is known, and also allow for the use of higher signal contrast levels. Despite these advantages, the use of search tasks in CT image quality assessment has been limited, as the results are difficult to model analytically, though approximate solutions and tractable approaches for signal search and data analysis are available.

2.B.3. Comparison of forced choice and rating scale experiments

2AFC studies simply require comparing two images, and are simpler and faster for humans to perform than rating scale experiments. However, 2AFC experiments tend to require more images than rating scale experiments.[32,33] MAFC has advantages and disadvantages as compared to 2AFC.[33] In particular, the difficulty of MAFC experiments increases with M. However, the MAFC technique allows the measurement of observer performance at higher signal contrast levels. Burgess has discussed the tradeoffs involved in choosing M for a forced-choice experiment.[32] Forced-choice experiments are useful mainly for human observers, and can be most efficient in situations where images are easily obtained, but observer time is at a premium. A shortcoming is that they generate only a summary statistic for the figure of merit, without information regarding the distribution of the observer's underlying test statistics.

Rating-scale experiments, which provide data along the full ROC curve, provide a more complete description of observer performance than forced-choice experiments. For a given variance in the figure of merit, a rating scale study requires half as many image pairs as a 2AFC study.[32] In situations where images are difficult to obtain, but observer time is inexpensive, the rating scale can be the most efficient use of a limited data set.

3. COMMON MODEL OBSERVERS FOR ASSESSING CT IMAGE QUALITY

In Sec. 2, we discussed the use of human observers for forced-choice and rating scale tasks; however, the use of model observers is an equally valid possibility. As the majority of studies to date use SKE experiments, we focus on model observers for SKE tasks. Here, we review the observers most commonly used in CT image quality assessment: the channelized Hotelling observer (CHO) and the NPW matched filter observer. Literature also provides useful general reviews on the design of task-based model observer studies.[34,35]

We briefly discuss the use of model observers for search tasks.

3.A. Formalism

The set of available images is generally finite, so that quantities like the mean and the covariance are based on sample statistics. Important quantities that model observers use are the mean vector of a sample of N images,

$$\overline{\mathbf{g}}^{(\mathscr{R})} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{g}_n^{(\mathscr{R})} \tag{2}$$

and the covariance matrix $\mathbf{C}$, defined for $N > 1$ as

$$\mathbf{C} = \frac{1}{N-1}\sum_{n=1}^{N} \left(\mathbf{g}_n^{(\mathscr{R})} - \overline{\mathbf{g}}^{(\mathscr{R})}\right)\left(\mathbf{g}_n^{(\mathscr{R})} - \overline{\mathbf{g}}^{(\mathscr{R})}\right)^T. \tag{3}$$

As $\mathbf{C}$ is a matrix with dimensions $M \times M$, the limiting factor in the calculation of the test statistic is often the dimensionality of this matrix.

3.B. The channelized Hotelling observer

3.B.1. The Hotelling observer

The CHO, commonly used in CT image quality assessment, is based on the Hotelling observer.[2] The Hotelling observer is a linear observer whose test statistic is

$$\lambda(\mathbf{g}^{(\mathscr{R})}) = \mathbf{w} \cdot \mathbf{g}^{(\mathscr{R})} \tag{4}$$

with the Hotelling template $\mathbf{w}$ defined by

$$\mathbf{w} = \mathbf{C}^{-1}\mathbf{s}, \tag{5}$$

the "difference signal," $\mathbf{s}$, by

$$\mathbf{s} = \overline{\mathbf{g}}_1 - \overline{\mathbf{g}}_0. \tag{6}$$

For detection tasks, $\overline{\mathbf{g}}_j$ ($j = 0, 1$) is the mean of an ensemble of signal absent ($j = 0$) or present ($j = 1$) images; for tasks involving discrimination between shapes, $\overline{\mathbf{g}}_j$ is the ensemble mean of images containing signals with Shape j. This choice of $\mathbf{w}$ maximizes the task SNR, and the Hotelling observer is thus the optimal linear observer.[2]

The Hotelling observer includes information about spatial correlations in the template: $\mathbf{C}^{-1}$ decorrelates (whitens) the noise in the image. This observer thus matches the mean signal profile to one that has passed through the whitening process. As Eq. (5) shows, the Hotelling observer requires only knowledge of the mean data vectors and $\mathbf{C}$.

Computing Hotelling observer performance can be difficult, largely due to the dimensionality of the covariance matrix requiring inversion in Eq. (5). For even a simple set of thirty-two 128 × 128 images, $\mathbf{s}$ has 1.3×10^6 elements, so $\mathbf{C}$ is a 10^6 dimensional matrix. Additionally, $\mathbf{C}$ can be ill-conditioned.

3.B.2. The channelized Hotelling observer

One approach to the problem of reducing the dimensionality of the data with minimal information loss is the introduction of channels. The CHO was first introduced in Refs. 36 and 37; further details are available in Ref. 38. "Channels" can be thought of as processors that selectively respond to different features, spatial or temporal frequency bands, or spatial orientations. The human visual system is generally understood to consist of multiple channels, with each channel sensitive to a different narrow range of spatial frequencies. Channelized model observers use similar channels as a preprocessing filter applied to the data, reducing an image to a set of channel response variables before making a decision. The channels reduce the dimension of an image to the number of channels, which is usually a large reduction.

A channelized image is described by $\mathbf{v}^{(\mathscr{R})} = \mathbf{T}\mathbf{g}^{(\mathscr{R})}$ where the dimension of the image $\mathbf{g}^{(\mathscr{R})}$ is M, the dimension of the channelized image $\mathbf{v}^{(\mathscr{R})}$ is N_c, and the operator $\mathbf{T}$ is an $N_c \times M$ matrix. In general $N_c \ll M$. For example, for a 256 × 256 image, $M = 65\,536$, whereas a satisfactory channelized representation can generally be obtained with $N_c < 10$.

The new effective covariance matrix $\mathbf{C}_v = \mathbf{T}^T\mathbf{C}\mathbf{T}$ is the $N_c \times N_c$ dimensional covariance matrix for $\mathbf{v}$. Reducing the dimensionality of the covariance matrix by orders of magnitude removes the computational roadblock posed by its inversion.

3.B.3. Types of channels

The first implementation of the CHO used radially symmetric channels, since the signal was radially symmetric and the noise was isotropic.[36] Subsequent research has expanded to a greater variety of channels. Types of channels commonly encountered in recent model observer studies include the following:

- **Laguerre-Gauss** (LG) channels are a traditional choice of channel. They are rotationally symmetric, smooth channels that are products of Laguerre polynomials and Gaussians. The LG channels form a basis for the space of rotationally symmetric, square-integrable functions. They are useful for signal-detection tasks with radially symmetric, smooth, signals where the background correlations have no preferred orientation (e.g., lumpy backgrounds).[39] While six channels are generally considered adequate for characterizing isotropic signals in uncomplicated backgrounds,[40] realistic clinical CT images do not generally meet these criteria.
- **Sparse and dense difference of Gaussians** (SDOG, DDOG) channels model the human visual system's spatial frequency selectivity. They have been validated against human performance and shown to correlate well (following noise regularization) (see, e.g., Ref. 41).
- **Gabor** channels have a psychophysical basis: the neuronal response to a small spot of light as a function of position.[42] These channels are not isotropic, not independent, and not based on any assumptions of signal isotropy. Each channel is parametrized by an octave bandwidth and an angular orientation. Reference 42 provides an overview of Gabor channels, noting that the number of Gabor channels has varied between 16 and 80.

In Table IV, we list some current research on the use of the CHO for CT image quality assessment.

3.C. Nonprewhitening matched filter observers

The NPW matched filter observer, like the Hotelling observer (Sec. 3.B.1) is a linear observer. As in Eq. (4), the test statistic is a scalar product

$$t(\mathbf{g}^{(\mathscr{R})}) = \mathbf{w} \cdot \mathbf{g}^{(\mathscr{R})}, \tag{7}$$

where $\mathbf{w}$ is the difference signal

$$\mathbf{w}(\mathbf{g}^{(\mathscr{R})}) = \mathbf{s} \tag{8}$$

as in Eq. (6). The matched filter observer is relatively easy to implement, since the template does not depend on the image covariance matrix. Unlike the Hotelling observer, the NPW observer does not decorrelate ("prewhiten") noise. In fact, the NPW observer includes no information at all about noise correlations, since the template in Eq. (8) is independent of $\mathbf{C}$. This observer is therefore suboptimal when the noise is correlated.

Reference 8 provides a good overview of the NPW filter as applied to CT. Research on the NPW observer's ability to match human performance in CT includes Ref. 47, which

TABLE IV. Recent developments in CHO-related methodology for image quality assessment in CT.

Reference	Summary
43	Yendiki and Fessler work with the type of IR algorithms that can be approximated as linear operators. They demonstrate that the CHO, among other linear and channelized linear observer models, can exactly achieve optimal performance for the given task and internal noise model, without needing regularization. They do however also show that known-location tasks are of limited use in designing regularized reconstruction methods that optimize lesion detectability. The task involved detection of a statistically varying signal with known location on a statistically varying background. Analytic results were provided for general channels; data was taken using square, SDOG, and DDOG channels.
44	Yendiki and Fessler provide results relevant to optimizing iterative reconstruction algorithms (for which a necessary step is assessing the quality of images generated by such algorithms.) The authors present and examine variants of the Hotelling observer. The results demonstrate that the observer's prewhitening capabilities are important in determining whether optimized regularization of IR algorithms is useful. The task is detection of a spatially localized target signal with unknown location (cf. Ref. 43) in an image reconstructed from noisy data. The observer used four circularly symmetric, frequency-selective channels.
21	Yu *et al.* design a phantom with three signals for use in a 2AFC experiment. They demonstrate that the CHO with Gabor channels is successful at predicting human observer performance for both FBP and the Siemens SAFIRE iterative reconstruction algorithm.
45	Sidky *et al.* devise a relatively computationally tractable calculation of Hotelling observer performance, which does not require storage of the full covariance matrix in computer memory. This paper uses signal and background known exactly tasks on reconstructed cone-beam images, and the observer uses 2, 4, 10, and 20 channels.
46	Wunderlich and Noo consider fan-beam CT images reconstructed using FBP and present a new method for calculating the covariance. The authors apply a CHO (with 40 Gabor channels) to a lesion detection task. Their methodology may be relevant to evaluating CT dose reduction claims, although the paper does not treat IR methods directly.

validates the NPW observer's results against human observers for many different scenarios.

3.C.1. Eye filters

The spectrum of random noise in simulated images, mammograms, and some nuclear medicine images can be concentrated at low power, and fall off rapidly. In this case, the simple NPW model may not predict human performance well.

Adding an eye filter, **E**, can improve the model's ability to predict human performance; the eye filter was introduced by Burgess.[48] The eye filter represents the effect of the human visual system's contrast sensitivity function, which is suppressed at low spatial frequencies because of retinal processing. This NPW model with eye filter is called the NPWE. Rather than the signal **s**, the model uses as a filter **Es** where **E** is the eye filter. Because the NPWE observer applies this template to noisy images to obtain its test statistic, the observer's figure of merit includes a factor of $\mathbf{E}^T\mathbf{E}$ times the noise power in the denominator (see Table VI). As CT noise is not low-pass, the utility of including an eye filter is not clear, although the presence of an eye filter is unlikely to have a large adverse impact on an observer's ability to predict human performance.

In Table V, we discuss recent implementations of the NPW observer with and without eye filters that may be relevant to dose reduction claims in CT IR. We discuss the results of Ref. 19 further in Sec. 2.A.

3.D. Observers for localization tasks

The observer models we discussed thus far have been for SKE tasks only. Observers for search tasks scan over all possible signal locations, and compute a test statistic as a function of position. Khurd and Gindi[49] developed the analog of an ideal observer for localization tasks that maximized the area under the LROC curve. Whitaker *et al.*[50] extended this work to generate the scanning linear estimator, and scanning Hotelling observers have also been developed. Swensson[28] suggested a semiparametric "binormal" estimation strategy for the LROC curve and the area under it; however, this work made strict assumptions about the observer's search process. Work by Popescu[26] relaxed some of these assumptions, and also provided estimators for the variance of the figure of merit. Wunderlich and Noo[51] provided a fully nonparametric approach for fully crossed multireader multicase (MRMC) analysis of areas under LROC curves.

While search tasks have been applied more in nuclear medicine than in CT, some progress has been made for CT as well. Popescu and Myers[22] have laid out a method for CT image quality assessment using a scanning observer; see also related work by Yendiki and Fessler.[44] Leng *et al.* have performed validation studies comparing the performance of a CHO with Gabor channels to human performance for search tasks on CT phantom images[29] and IR images specifically.[30]

4. ASSESSING AND COMPARING THE PERFORMANCE OF ITERATIVE RECONSTRUCTION ALGORITHMS

4.A. Figures of merit

In order to summarize how well an observer can detect signals when images are reconstructed using different IR algorithms, we need a figure of merit that will enable quantitative comparison of the performance of different algorithms for a given task. In this section, we discuss various choices for task-based figures of merit.

A commonly used figure of merit is the task SNR. This figure of merit is based on the means and variances of the observer's test statistics λ for each of the underlying hypotheses. Let $\overline{\lambda}_1$ and $\overline{\lambda}_2$ denote the mean values of λ under the two hypotheses (e.g., signal present and absent cases). Similarly, let σ_1^2 and σ_2^2 denote the variances of λ under the two different hypotheses. The task SNR is

$$\mathrm{SNR} = \frac{\overline{\lambda}_1 - \overline{\lambda}_2}{\sqrt{\frac{1}{2}\left(\sigma_1^2 + \sigma_2^2\right)}}. \tag{9}$$

The expression in Eq. (9) is a measure of the separability of the test statistics under the two hypotheses.

When the variance of the test statistic is the same under the two hypotheses, so that $\sigma_1 \approx \sigma_2 = \sigma$, the task SNR is commonly referred to as the detectability, written

$$d' = \frac{\overline{\lambda}_1 - \overline{\lambda}_2}{\sigma}. \tag{10}$$

The SNR and detectability are meaningful figures of merit as long as the test statistic has approximately Gaussian statistics. Because the test statistic for a linear observer is the result of a summation operation over an area of the image, for example, in Eq. (4), the Central Limit Theorem implies that

TABLE V. Two recent implementations of NPW observers (with and without eye filters) that contain results relevant to the objective assessment of IR image quality. These references discuss the use of the NPW observer as a tool (Ref. 8) and implement it in a detectability study (Ref. 19).

Reference	Summary
8	Boedeker and McNitt-Gray use a NPW observer (without eye filter) calculated from the MTF and NPS obtained via phantom scans to calculate the SNR under a variety of conditions. They thus examine the influence of different reconstruction kernels and radiation doses on SNR. The results provide useful information regarding the tradeoffs between dose and image quality. The paper also provides a thorough background on NPW matched filters in the evaluation of CT images, and a description of the methodology and its limitations.
19	Hernandez-Girón *et al.* use a NPW matched filter with an eye filter to perform a 2AFC experiment using model and human observers. The authors validate their model observer against human observer performance with favorable results. The paper uses the *Catphan* 600 phantom, and lesion-absent images are drawn from a ROI at 1.2 times the radius of the lesions. The inability to extract lesion-absent images from the same radius as signal-present images is a shortcoming of commercial phantoms.

this requirement is very often satisfied regardless of the statistics of the image pixels themselves.

For a linear observer, where $\lambda = \mathbf{w} \cdot \mathbf{g}^{(\mathscr{R})}$, after some algebra

$$d' = \frac{\mathbf{w} \cdot \mathbf{s}}{\sqrt{\mathbf{w}^T \mathbf{C} \mathbf{w}}}, \tag{11}$$

where $\mathbf{s}$ is the signal [Eq. (6)].

When observer training (calculation of the template) uses image samples from the data set, the SNR must be obtained using test statistics resulting from applying that template to an *independent* set of images (the test set) to avoid overestimation of the SNR.[52] Calculation of Eq. (11) using all available images is equivalent to training and testing with the same image set, which is sometimes called "resubstitution." The resulting figure of merit will be optimistically biased with respect to the true detectability one would obtain with an infinite number of image samples. Alternatively, if one used a set of images to train the model observer, and a separate set to test the observer, the test statistic will underestimate the true detectability one would obtain on an infinite number of images. There is reasonable consensus in the imaging assessment community that the use of a conservative estimate of task performance is preferred over ones that are biased high. Of course, the larger the sample size, the lower the bias. Petrick *et al.*[53] provide a helpful discussion of issues related to the training and testing of machine classifiers and the determination of sample sizes which can be applied to the use of model observers.

Table VI contains expressions for the templates and detectability for each of the linear observers we have discussed in Sec. 3.

The task-based SNR and the detectability index are functions of the first order (means) and second order (variance) statistics of the observer's test statistic λ under each hypothesis. A more complete understanding of the observer's performance can be obtained through an analysis of the observer's full ROC curve, that is, a plot of the observer's true-positive fraction (TPF) vs the false-positive fraction (FPF) for all possible thresholds or cut-off values applied to the distributions of the test statistics.[54] This plot can be obtained for experiments using model or human observers. For the latter, a rating scale experiment involves human observers providing a score for each image (or ROI) on a reporting scale that may be a five- or seven-point scale or quasicontinuous (a 100-point scale, for example). Machine observers readily output a continuous test statistic for each image without need for binning and the resultant loss of information. Binned or quantized rating scales are relevant for studies involving human observers, who are more comfortable and able to provide more repeatable data using such reporting scales. In either case, once the distributions of the observer test statistics are obtained for the test image set, the experimenter can plot the observer's TPF vs their FPF for all thresholds or decision cutoffs. The AUC is a widely used summary measure of observer performance.[52] The plot of the full ROC curve can provide the experimenter with additional information regarding the curve shapes, and especially whether two ROC curves representing two observers or modalities (different reconstruction algorithms, for example) cross.

TABLE VI. Detectability of various linear observers, derived from Eq. (11).

Observer	Template $\mathbf{w}$	Detectability d'
Hotelling	$\mathbf{C}^{-1}\mathbf{s}$	$\sqrt{\mathbf{s}^T \mathbf{C}^{-1} \mathbf{s}}$
Channelized Hotelling	$\mathbf{C}_v^{-1} \mathbf{T}^T \mathbf{s}$	$\sqrt{\mathbf{s}^T \mathbf{T} \mathbf{C}_v^{-1} \mathbf{T}^T \mathbf{s}}$
NPW matched filter	$\mathbf{s}$	$\frac{\mathbf{s}^T \mathbf{s}}{\sqrt{\mathbf{s}^T \mathbf{C} \mathbf{s}}}$
NPW matched filter with eye filter	$\mathbf{E}^T \mathbf{E} \mathbf{s}$	$\frac{\mathbf{s}^T \mathbf{E}^T \mathbf{E} \mathbf{s}}{\sqrt{\mathbf{s}^T \mathbf{E}^T \mathbf{E} \mathbf{C} \mathbf{E}^T E \mathbf{s}}}$

It is interesting and important to note that the AUC can be shown to be equivalent to the percent correct in a 2AFC experiment (see p. 823 of Ref. 2).

A number of software packages are publicly available for use in plotting ROC curves and analyzing the results in terms of the AUC and its uncertainty; see, for example, the University of Chicago[57] or University of Iowa[58] websites. There is general consensus in the imaging assessment community on the importance of reporting confidence intervals for the figure or merit, in addition to the figure of merit itself. Methods for computing such error bars are obtainable using the aforementioned software packages. For tasks involving random locations in which the observer must localize the signal, a number of variants to standard ROC analysis exist.[52]

Experiments involving human readers must contend with the issue of reader variability. The most common approach to this challenge is to utilize a study design involving multiple readers, where each reader reads all cases/images from each modality (reconstruction algorithm, for our purposes here). This design is known as a fully-crossed MRMC study. Tools and software are available in the literature and online that design and analyze data from MRMC studies.[55]

4.B. Comparing the performance of different reconstruction algorithms at different doses

To determine the dose level at which two different reconstruction algorithms perform equivalently on a task, the task should be performed at different dose levels for each of the reconstruction algorithms. The typical data obtained will be a set of performance values for a finite set of dose values for the algorithms under comparison.

See Fig. 4, where we plot a figure of merit summarizing observer performance for each of the dose values that were used to create images.[22] CT data were collected at seven doses and reconstructed using two algorithms: FBP (black curve) and a representative iterative algorithm (blue curve). The error bars should account for the number of images used to measure performance as well as reader variability (due to the range of human observer skill and threshold in a reader study, or the training variability encountered in a study using a model observer.) A recent example of the application of nonparametric analysis[26,51] of model and human performance data for comparing reconstruction image quality is given by Leng *et al.*[56]

In order to demonstrate that an algorithm improves image quality for a given dose, a comparison is made between

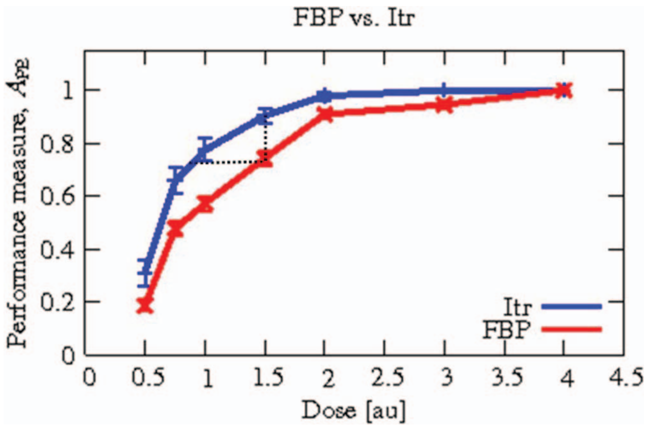

FIG. 4. Adapted from Ref. 22. CT data were simulated at seven doses and reconstructed using two algorithms: FBP (blue curve, "+" markers) and a representative IR algorithm (red curve, "X" markers). In order to demonstrate that an algorithm improves image quality for a given dose, the comparison is between two points that are vertically separated and showing that their difference is statistically significant (vertical dotted line). Alternately, one may wish to show that the iterative algorithm allows for a dose reduction while achieving image quality (horizontal dotted line). Note that this leads to a part of the performance curve for the iterative algorithm that was not experimentally sampled.

two points in a plot like that shown in Fig. 4 that are vertically separated (vertical dotted line, Fig. 4), with the goal of demonstrating that the figures of merit at that dose are statistically significantly different. Alternately, one may wish to demonstrate that the iterative algorithm allows for dose reduction while maintaining image quality (horizontal dotted line, Fig. 4). This latter goal may be complicated by the fact that the horizontal line leads to a part of the performance curve for the iterative algorithm that was not experimentally sampled. Methods for dealing with this potential problem include taking additional data or interpolation of the existing data in some justifiable manner.

It is recommended that the entire performance plot as a function of dose be presented for a full understanding of each algorithm's performance characteristic as a function of dose as well as an understanding of the selection of the particular dose point (placement of vertical line, Fig. 4) or image quality value (horizontal line, Fig. 4) at which the comparison is made. It is also understood that a dose improvement claim will be made with respect to a modern FBP algorithm. To avoid strong dependence on the particular choice of image quality or dose level at baseline, a summary statistic for the difference in algorithm performance based on the difference in the areas under these curves might be a useful alternative to point-based measures. Developing such a statistic is a topic of future research investigation.

5. CONCLUSIONS

In this paper, we have reviewed approaches to the objective assessment of the quality of iteratively reconstructed CT images, with the intent of providing a resource for CT manufacturers or users who are designing studies to validate quantitative dose reduction claims for iterative reconstruction. As each manufacturer uses a different reconstruction algorithm, the noise properties of the resulting images are quite different, and different manufacturers may find different tasks and observers to be most convenient for validating the performance of their particular device. For this reason, rather than focusing on a single task or observer, we have provided a general review of various tasks and observers, noting that the majority of current CT image quality assessment research focuses on the Channelized Hotelling observer or the NPWE observer, and uses signal detection tasks.

In validating quantitative dose reduction claims, model observers may provide a partial or even a complete replacement for the use of human observers. Studies using human subjects must be designed in a way that reduces bias or reader learning effects on the results to acceptable levels.

The phantom-based validation paradigm is more simplistic than a true clinical scenario, and the estimates of dose reduction that it generates may not necessarily be fully achievable in true clinical practice. USFDA therefore has recommended that manufacturers attach a disclaimer acknowledging these limitations to dose reduction claims; for example, such a disclaimer might read "In clinical practice, the use of this algorithm may reduce CT patient dose depending on the clinical task, patient size, anatomical location, and clinical practice. A consultation with a radiologist and a physicist should be made to determine the appropriate dose to obtain diagnostic image quality for the particular clinical task." Manufacturers also should describe the study methodology used to validate the dose reduction claims, in sufficient detail that the results can be reproduced.

USFDA cleared the first quantitative dose reduction claim, for the Siemens SAFIRE algorithm, in November 2011.[6] This clearance was based on a model observer study. A second quantitative claim by Philips[5] received clearance in June 2013; this clearance used a human observer study. Most recently, in March 2014 GE received clearance on a quantitative dose reduction claim for the ASIR-V algorithm, also based on a model observer study.[3]

[a] Author to whom correspondence should be addressed. Electronic mail: jay.vaishnav@fda.hhs.gov

[1] D. J. Brenner and E. J. Hall, "Computed tomography — an increasing source of radiation exposure," New Engl. J. Med. **357**, 2277–2284 (2007).
[2] H. H. Barrett and K. J. Myers, *Foundations of Image Science* (Wiley, New York, 2004).
[3] GE Healthcare, *510(k) Summary for the Discovery CT590 RT and Optima CT580W, FDA 510(k) Premarket Notification Database* (GE Medical Systems, LLC, Waukesha, WI, 2014) (available URL: http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K093581).
[4] Toshiba Medical Systems Corporation, *510(k) Summary for the TSX-302A, Aquilion Prime, FDA 510(k) Premarket Notification Database* (Toshiba Medical Systems Corporation, Tustin, CA, 2011) (available URL: http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K110066).
[5] Philips Medical Systems, Inc., *510(k) Summary for the Philips IMR Software Application, FDA 510(k) Premarket Notification Database* (Philips Medical Systems, Inc., Cleveland, OH, 2013) (available URL: http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K123576).
[6] Siemens Medical Systems, Inc., *510(k) Summary for the Somatom Definition Flash, FDA 510(k) Premarket Notification Database, FDA 510(k) Premarket Notification Database* (Siemens Medical Systems, Inc., Malvern, PA, 2011) (available URL: http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K113342).
[7] A. E. Burgess, "The Rose model, revisited," J. Opt. Soc. Am. A **16**, 633–646 (1999).

[8] K. Boedeker and M. McNitt-Gray, "Application of the noise power spectrum in modern diagnostic MDCT: Part II. Noise power spectra and signal to noise," Phys. Med. Biol. **52**, 4047–4061 (2007).

[9] K. J. Myers, H. H. Barrett, M. C. Borgstrom, D. D. Patton, and G. W. Seeley, "Effect of noise correlation on detectability of disk signals in medical imaging," J. Opt. Soc. Am. A **2**, 1752–1759 (1985).

[10] *Guidance for Industry and/or for FDA Reviewers/Staff and/or Compliance: Guidance for the Submission of 510(k)'s for Solid State X-ray Imaging Devices* (available URL: http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm073780.htm, August 1999).

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

[12] A. Pineda, D. Tward, A. Gonzalez, and J. Siewerdsen, "Beyond noise power in 3D computed tomography: The local NPS and off-diagonal elements of the Fourier domain covariance matrix," Med. Phys. **39**, 3240–3252 (2012).

[13] J. Evans, D. Politte, B. Whiting, J. O'Sullivan, and J. Williamson, "Noise-resolution tradeoffs in x-ray CT imaging: A comparison of penalized alternating minimization and filtered backprojection algorithms," Med. Phys. **38**, 1444–1458 (2011).

[14] C. C. Brunner, S. F. Abboud, C. Hoeschen, and I. S. Kyprianou, "Signal detection and location-dependent noise in cone-beam computed tomography using the spatial definition of the Hotelling SNR," Med. Phys. **39**, 3214–3228 (2012).

[15] 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. **38**, 1754–1768 (2011).

[16] A. Hara, R. Paden, A. Silva, J. Kujak, H. Lawder, and W. Pavlicek, "Iterative reconstruction technique for reducing body radiation dose at CT: Feasibility study," Am. J. Roentgenol. **193**, 764–771 (2009).

[17] 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. Med. **29**, 99–110 (2013).

[18] K. Boedeker and M. McNitt-Gray, "Application of the noise power spectrum in modern diagnostic MDCT: Part I. Measurement of noise power spectra and noise equivalent quanta," Phys. Med. Biol. **52**, 4027–4046 (2007).

[19] I. Hernandez-Girón, J. Geleijns, A. Calzado, and W. J. H. Veldkamp, "Automated assessment of low contrast sensitivity for CT systems using a model observer," Med. Phys. **38**, S25–S35 (2011).

[20] A. Wunderlich and F. Noo, "Confidence intervals for performance assessment of linear observers," Med. Phys. **38**, S57–S68 (2011).

[21] 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. **40**, 041908 (9pp.) (2013).

[22] L. M. Popescu and K. J. Myers, "CT image assessment by low contrast signal detectability evaluation with unknown signal location," Med. Phys. **40**, 111908 (10pp.) (2013).

[23] J. Fan, H.-W. Tseng, M. Kupinski, G. Cao, P. Sainath, and J. Hsieh, "Study of the radiation dose reduction capability of a CT reconstruction algorithm LCD performance assessment using mathematical model observers," in Medical Imaging 2013: Image Perception, Observer Performance, and Technology Assessment, edited by C. K. Abbey and C. R. Mello-Thoms [Proc. SPIE **8673**, 86731Q (2013)].

[24] D. M. Green and J. A. Swets, *Signal Detection Theory and Psychophysics* (Wiley, New York, 1966) [reprint (Krieger, New York, 1974)].

[25] L. M. Popescu and R. M. Lewitt, "Small nodule detectability evaluation using a generalized scan-statistic model," Phys. Med. Biol. **51**, 6225–6244 (2006).

[26] L. M. Popescu, "Nonparametric ROC and LROC analysis," Med. Phys. **34**, 1556–1564 (2007).

[27] L. M. Popescu, "Model for the detection of signals in images with multiple suspicious locations," Med. Phys. **35**, 5565–5574 (2008).

[28] R. G. Swensson, "Unified measurement of observer performance in detecting and localizing target objects on images," Med. Phys. **23**, 1709–1725 (1996).

[29] 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. **40**, 081908 (9pp.) (2013).

[30] S. Leng *et al.*, "Diagnostic performance assessment of an iterative reconstruction algorithm using a model observer: Correlation with human observers for a low contrast detection task with unknown lesion locations," Radiological Society of North America SSE23-04, 13019225, 2013.

[31] L. M. Popescu, "Nonparametric signal detectability evaluation using an exponential transformation of the FROC curve," Med. Phys. **38**, 5690–5702 (2011).

[32] A. E. Burgess, "Comparison of receiver operating characteristic and forced choice observer performance measurement methods," Med. Phys. **22**, 643–655 (1995).

[33] C. E. Metz, "Fundamental ROC analysis," in *Handbook of Medical Imaging. Vol 1: Physics and Psychophysics*, edited by J. Beutel, H. L. Kundel, and R. L. Van Metter (SPIE, Bellingham, WA, 2000), pp. 751–769.

[34] M. P. Eckstein, C. K. Abbey, and F. O. Bochud, "Practical guide to model observers for visual detection in synthetic and natural noisy images," in *Handbook of Medical Imaging. Volume 1. Physics and Psychophysics*, edited by J. Beutel, H. L. Kundel, and R. L. Van Metter (SPIE, Bellingham, WA, 2000), pp. 629–654.

[35] C. K. Abbey and F. Bochud, "Modeling visual detection tasks in correlated image noise with linear model observers," in *Handbook of Medical Imaging. Volume 1. Physics and Psychophysics* (SPIE, Bellingham, WA, 2000), pp. 629–654.

[36] K. J. Myers and H. H. Barrett, "Addition of a channel mechanism to the ideal-observer model," J. Opt. Soc. Am. A **4**, 2447–2457 (1987).

[37] H. H. Barrett, J. Yao, J. P. Rolland, and K. J. Myers, "Model observers for assessment of image quality," Proc. Natl. Acad. Sci. U.S.A. **90**, 9758–9765 (1993).

[38] C. K. Abbey and F. O. Bochud, "Modeling visual detection tasks in correlated noise with linear model observers," in *Handbook of Medical Imaging. Vol 1: Physics and Psychophysics*, edited by J. Beutel, H. L. Kundel, and R. L. Van Metter (SPIE, Bellingham, WA, 2000), pp. 629–654.

[39] B. D. Gallas and H. H. Barrett, "Validating the use of channels to estimate the ideal linear observer," J. Opt. Soc. Am. A **20**, 1725–1738 (2003).

[40] H. H. Barrett, C. K. Abbey, B. D. Gallas, and M. P. Eckstein, "Stabilized estimates of Hotelling-observer detection performance inpatient-structured noise," Proc. SPIE **3340**, 27–43 (1998).

[41] 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**, 473–488 (2001).

[42] A. Chawla, "Correlation imaging for improved cancer detection," Ph.D. thesis, Duke University, 2008.

[43] A. Yendiki and J. Fessler, "Analysis of observer performance in known-location tasks for tomographic image reconstruction," IEEE Trans. Med. Imaging **25**, 28–41 (2006).

[44] A. Yendiki and J. Fessler, "Analysis of observer performance in unknown-location tasks for tomographic image reconstruction," J. Opt. Soc. Am. A **24**, B99–B109 (2007).

[45] E. Sidky, S. LaRoque, and X. Pan, "Accurate computation of the Hotelling observer for the evaluation of image reconstruction algorithms in helical, cone-beam CT," Nucl. Sci. Symp. Conf. Rec. **5**, 3233–3236 (2007).

[46] A. Wunderlich and F. Noo, "Image covariance and lesion detectability in direct fan-beam x-ray computed tomography," Phys. Med. Biol. **53**, 2471–2493 (2008).

[47] P. F. Judy and R. G. Swensson, "Lesion detection and signal-to-noise ratio in CT images," Med. Phys. **8**, 13–23 (1981).

[48] A. E. Burgess, "Statistically defined backgrounds: Performance of a modified non-prewhitening matched filter model," J. Opt. Soc. Am. A **11**, 1237–1242 (1994).

[49] P. Khurd and G. Gindi, "Decision strategies that maximize the area under the LROC curve," IEEE Trans. Med. Imaging **24**, 1626–1636 (2005).

[50] M. Whitaker, E. Clarkson, and H. Barrett, "Estimating random signal parameters from noisy images with nuisance parameters: Linear and scanning-linear methods," Opt. Express **16**, 8150–8173 (2008).

[51] A. Wunderlich and F. Noo, "A nonparametric procedure for comparing the areas under correlated LROC curves," IEEE Trans. Med. Imaging **31**, 2050–2061 (2012).

[52] B. D. Gallas *et al.*, "Evaluating imaging and computer-aided detection and diagnosis devices at the FDA," Acad. Radiol. **19**, 463–477 (2012).

[53] N. Petrick, B. Sahiner, S. G. Armato III, A. Bert, L. Correale, S. Delsanto, M. T. Freedman, D. Fryd, D. Gur, L. Hadjiiski, Z. Huo, Y. Jiang, L. Morra, S. Paquerault, V. Raykar, F. Samuelson, R. M. Summers, G. Tourassi, H. Yoshida, B. Zheng, C. Zhou, and H.-P. Chan, "Evaluation of computer-aided detection and diagnosis systems," Med. Phys. **40**, 087001 (17pp.) (2013).

[54] C. E. Metz, "ROC analysis in medical imaging: A tutorial review of the literature," Radiol. Phys. Technol. **1**, 2–12 (2008).

[55] X. He and B. D. Gallas, "iMRMC, analyzing and sizing multi-reader multi-case ROC trials," 2012 (available URL: http://js.cx/~xin/index.html).

[56] S. Leng, L. Yu, L. Chen, J. C. R. Giraldo, and C. H. McCollough, "Correlation between model observer and human observer performance in CT imaging when lesion location is uncertain," Proc. SPIE **8313**, Medical Imaging 2012: Physics of Medical Imaging, 83131M (24 February 2012).

[57] Metz ROC Software at The University of Chicago (available URL: http://metz-roc.uchicago.edu/).

[58] OR-DBM MRMC 2.4 Software, Medical Image Perception Laboratory, University of Iowa, 2013 (available URL: http://perception.radiology.uiowa.edu/Software/ReceiverOperatingCharacteristicROC/MRMCAnalysis/tabid/116/Default.aspx).