Validation of algorithmic CT image quality metrics with preferences of radiologists

Yuan Cheng
Clinical Imaging Physics Group, Medical Physics Graduate Program, Carl E. Ravin Advanced Imaging Laboratories, Duke University, 2424 Erwin Rd, Suite 302, Durham, NC 27705, USA

Ehsan Abadi
Carl E. Ravin Advanced Imaging Labs and Clinical Imaging Physics Group, Duke University Health System, 2424 Erwin Road, Suite 302, Durham, NC 27710, USA
Department of Radiology, Duke University Health System, Box 3808,Room 1531, Erwin Rd, Durham, NC 27710, USA

Taylor Brunton Smith
Clinical Imaging Physics Group, Medical Physics Graduate Program, Carl E. Ravin Advanced Imaging Laboratories, Duke University, 2424 Erwin Rd, Suite 302, Durham, NC 27705, USA

Francesco Ria
Carl E. Ravin Advanced Imaging Labs and Clinical Imaging Physics Group, Duke University Health System, 2424 Erwin Road, Suite 302, Durham, NC 27710, USA

Mathias Meyer, and Daniele Marin
Department of Radiology, Duke University Health System, Box 3808,Room 1531, Erwin Rd, Durham, NC 27710, USA

Ehsan Samei[a)]
Clinical Imaging Physics Group, Medical Physics Graduate Program, Carl E. Ravin Advanced Imaging Laboratories, Departments of Radiology, Physics, Biomedical Engineering, and Electrical and Computer Engineering, Duke University, 2424 Erwin Rd, Suite 302, Durham, NC 27705, USA

(Received 18 April 2019; revised 13 August 2019; accepted for publication 13 August 2019; published 20 September 2019)

Purpose: Automated assessment of perceptual image quality on clinical Computed Tomography (CT) data by computer algorithms has the potential to greatly facilitate data-driven monitoring and optimization of CT image acquisition protocols. The application of these techniques in clinical operation requires the knowledge of how the output of the computer algorithms corresponds to clinical expectations. This study addressed the need to validate algorithmic image quality measurements on clinical CT images with preferences of radiologists and determine the clinically acceptable range of algorithmic measurements for abdominal CT examinations.

Materials and methods: Algorithmic measurements of image quality metrics (organ HU, noise magnitude, and clarity) were performed on a clinical CT image dataset with supplemental measures of noise power spectrum from phantom images using techniques developed previously. The algorithmic measurements were compared to clinical expectations of image quality in an observer study with seven radiologists. Sets of CT liver images were selected from the dataset where images in the same set varied in terms of one metric at a time. These sets of images were shown via a web interface to one observer at a time. First, the observer rank ordered the CT images in a set according to his/her preference for the varying metric. The observer then selected his/her preferred acceptable range of the metric within the ranked images. The agreement between algorithmic and observer rankings of image quality were investigated and the clinically acceptable image quality in terms of algorithmic measurements were determined.

Results: The overall rank-order agreements between algorithmic and observer assessments were 0.90, 0.98, and 1.00 for noise magnitude, liver parenchyma HU, and clarity, respectively. The results indicate a strong agreement between the algorithmic and observer assessments of image quality. Clinically acceptable thresholds (median) of algorithmic metric values were (17.8, 32.6) HU for noise magnitude, (92.1, 131.9) for liver parenchyma HU, and (0.47, 0.52) for clarity.

Conclusions: The observer study results indicated that these algorithms can robustly assess the perceptual quality of clinical CT images in an automated fashion. Clinically acceptable ranges of algorithmic measurements were determined. The correspondence of these image quality assessment algorithms to clinical expectations paves the way toward establishing diagnostic reference levels in terms of clinically acceptable perceptual image quality and data-driven optimization of CT image acquisition protocols. [https://doi.org/10.1002/mp.13795]

Key words: automated assessment, clinical images, CT image quality, diagnostic reference level, observer study, validation of algorithmic metrics

1. INTRODUCTION

Image quality reflects the quality of the information content that the imaging system delivers for its intended function. The key information content a medical imaging system presents includes anatomic and physiologic information of a patient. Effective rendering of such information in clinical images is indispensable for assuring quality of clinical decision making and mitigating clinical risks.[1–4] In particular, Computed tomography (CT) is well known for its capability to produce high quality clinical images of a patient quickly in three dimensions (3D).[5] While CT is an essential diagnostic imaging tool for a variety of clinical scenarios, assuring quality of clinical CT images is fundamental for their effective use in patient care.[6]

In the current clinical practice, the quality of clinical CT images is controlled through mandatory image quality evaluation of CT scanners.[7] A significant part of the evaluation is the analysis of images of specific test objects, that is, phantoms. Phantom-based tests have been successful in identifying malfunctioning CT scanners, including those with incorrectly configured image acquisition protocols. Although standardized phantom-based tests are valuable and no doubt necessary, they are more reflective of the technical capabilities of a CT scanner than the actual quality of clinical CT images; diagnostic CT scanners that pass the phantom tests can produce clinical CT images of poor quality. Anthropomorphic phantom-based tests tend to provide better insights into image quality, but the phantom still does not fully represent the complexity of a patient. Furthermore, modern CT scanners that employ dose reduction techniques[8,9] such as Automatic Tube Current Modulation (ATCM), nonlinear iterative reconstruction, and/or postprocessing techniques, produce images with variable quality depending on the object-specific properties.[10,11] Thus, the framework of monitoring the quality of CT imaging *primarily* through phantom-based measurements is not optimal for assuring the quality of clinical CT images. Image quality measurements made on clinical CT images are indispensable for assuring their quality.

Prior work has attempted to incorporate image quality measurements on patient CT images in addition to in-phantom measurements.[12–14] Several automated algorithms assessing essential image quality metrics on patient CT images and phantom CT images have been demonstrated recently with validated accuracies. These algorithmic image quality metrics include organ Hounsfield Unit (HU), noise magnitude, spatial resolution, noise texture, and detectability index.[15–18] Automated assessment of image quality on clinical CT images by computer algorithms has the potential to greatly facilitate data-driven monitoring and optimization of CT image acquisition protocols. However, the application of these techniques in clinical operation requires the knowledge of how the output of the computer algorithms corresponds to clinical expectations.

This study aimed to address this question by investigating the correspondence of algorithmic image quality measurements on clinical CT images with preferences of radiologists, and further determining the clinically acceptable range of algorithmic measurements for abdominal CT examinations.

2. MATERIALS AND METHODS

In summary, algorithmic measurements of image quality metrics (organ HU, noise magnitude, and clarity – a new metric of normalized detectability) were performed on a clinical CT image dataset using techniques which were developed previously.[15–18] The algorithmic measurements were compared to clinical expectations of image quality by radiologists in an observer study. First, the observers rank ordered the CT images according to their preference for the varying metric (noise magnitude, liver parenchyma HU, and clarity). The observers further selected a range within which they judged the quality of the images acceptable. The agreement between algorithmic (quantitative) and observer (qualitative visual) rankings of image quality was investigated and the clinically acceptable image quality in terms of algorithmic (quantitative) measurements were determined. The materials and methods of this study are described in more detail in the following sections.

2.A. Clinical CT image dataset

The clinical CT image dataset was obtained from two CT imaging studies: (a) a diagnostic imaging quality improvement study at Duke, and (b) a previously investigated multicenter study.[19] These studies were compliant with Health Insurance Portability and Accountability Act of 1996 (HIPAA) and approved by their associated Institutional Review Board (IRB). The obtained CT image dataset included 472 adult CT series acquired with scanner models from two manufacturers (Discovery CT750 HD, GE Healthcare, Waukesha, WI; and SOMATOM Definition Flash, Siemens Healthineers, Erlangen, Germany). The CT series were acquired at 120 kV and ATCM with a variety of contrast-enhanced abdominal imaging protocols that were available at three institutions. They included images of patient liver and had relatively thin slice reconstruction (0.6 mm for Siemens scanner and 0.625 mm for GE scanner). The reconstruction algorithms of the CT series included Filtered Back Projection (FBP), Sinogram-AFfirmed Iterative REconstruction (SAFIRE), Adaptive Statistical Iterative Reconstruction (ASIR), and Model-Based Iterative Reconstruction (MBIR).[19]

2.B. Algorithmic image quality assessments on clinical CT series

2.B.1. Organ Hounsfield Units

The automated algorithm for measuring organ Hounsfield Units (HU) values[15] was adapted in this study for measuring liver parenchyma HU and aorta HU. Algorithmic aorta HU measurements can help to identify those CT series that are likely to be acquired during undesired phase of contrast-enhanced CT.

2.B.2. Noise Magnitude

The automated algorithm for measuring noise magnitude[17] was applied to measure noise magnitude of individual CT images. The method works by measuring the standard deviation of many small (9 × 9) Region Of Interests (ROIs) within the homogenous areas of soft tissue (HU ∈ [−300, 100]) of the patient, and creating a histogram of these standard deviation values. The mode (most common) value of the standard deviation histogram is taken to be the noise magnitude in that image.

2.B.3. Clarity

"Clarity" is introduced in this study to quantify the perceived image quality in terms of the signal-to-noise ratio for humans to detect a lesion (with certain size and shape) which has a unit value of contrast-to-noise ratio (CNR) in an image. When there are lesions with same CNR appearing in images with different clarity levels, images with higher clarity have better perceived image quality for detecting such lesions. Clarity combines spatial resolution weighted by noise texture and a lesion model via a Non-PreWhitening (NPW) matched filter observer model into a composite image quality metric:

$$\text{Clarity} = \sqrt{\frac{\left[\int\int |nW(u,v)|^2 \cdot MTF^2(u,v) dudv\right]^2}{\int\int |nW(u,v)|^2 \cdot MTF^2(u,v) \cdot nNPS(u,v) dudv}} \quad (1)$$

here, *nW(u,v)* is the Fourier-representation of the task function which describes a hypothetical lesion (uniform disk, 2 mm in diameter) in frequency space and is normalized to have a unit integral. The *nNPS(u,v)* is the noise power spectrum which has been normalized to have a unit integral. The NPW matched filter observer model, which does not decorrelate noise, has been shown to explain human observer's responses for low-contrast detection tasks[20] such as the one investigated in this study. We used a 2-mm disk signal detection task to generically represent small noncalcified hepatic lesions of a typical shape.

Clarity describes how clearly a lesion (with a certain shape) is rendered in an image whose signal and noise transfer properties are represented by MTF and *nNPS*, and is related to the detectability index d'_{NPW} of the image for a lesion with unit contrast and noise magnitude by Eq. (2):

$$\text{Clarity} = d'_{NPW}(\text{Lesion Contrast} = 1, \text{Noise} = 1) \quad (2)$$

In this way, the contribution to lesion detectability from the clearness of the rendering of a lesion as limited by spatial resolution weighted by noise texture (quantified by the clarity) can be decoupled from both (1) the contrast of that lesion to the background, and the (2) noise magnitude of the image. Clarity can be further combined with lesion contrast and noise magnitude to obtain detectability index d'_{NPW} by Eq. (3):

$$d'_{NPW} = \text{Clarity} * \frac{\text{Lesion Contrast}}{\text{Noise}}. \quad (3)$$

The automated algorithm for computing detectability index from a patient CT image[20] was adapted to assess clarity. To measure the clarity, the automated algorithm for measuring spatial resolution[18] was used to measure modulation transfer function (MTF) of a CT series.

In this study, we did not measure noise texture from clinical images, but rather from phantom images acquired at equivalent imaging protocols as clinical images. The previously investigated algorithm[16] was used to measure noise power spectrum (NPS) from a phantom CT series acquired with imaging protocols identical to those used in clinical CT series. In this estimation, we assume that the noise texture of phantom images represent that of clinical images (an assumption that requires a validation in the future)[21] with the provision of being able to use large homogenous regions of phantom images to provide low-noise NPS measurements. We acquired images of Mercury phantom at the same imaging protocols as clinical images. Mercury phantom had homogenous regions available in several diameters for measuring NPS. We took measurements in regions within a certain diameter of the phantom whose water equivalent diameter[22] matched that of an average adult. This allowed us to use the same set of ROIs for measuring NPS across different imaging protocols.

The algorithmic image quality metrics of organ HU, noise magnitude, and spatial resolution (MTF) were assessed on all the clinical CT series (cases). The MTF can vary depending on location and contrast level of measurements particularly for highly nonlinear reconstruction algorithms. To obtain comparable MTFs from CT series of different patients, we used MTF measurements at the same location (a reference distance-to-isocenter) and same contrast level (air/skin interface) across different patients. Although the MTF measured at the air–skin interface could be different from that associated with a lesion within a target organ, there was a correspondence between the two that was important for this study. When the measured MTFs of two CT series were similar, we anticipated those within the target organ in these CT series to also be similar. In order to choose the reference distance-to-isocenter which would be selected for the MTF measurements, we assessed the distance at which we can obtain MTF measurements for the greatest number of patients. The algorithmic MTF measurements were obtained at various distances-to-isocenter for each case where there were adequate number of MTF samples. Figure 1(a) shows a plot with the number of clinical cases with available MTF measurements at a certain distance-to-isocenter vs the distance-to-isocenter. A distance of 15.5 cm to the isocenter proved to have the largest number of cases with MTF measurements and thus was selected as the reference distance-to-isocenter. For computing clarity, the MTF measurements at this reference distance-to-isocenter across different patients were used without scaling.

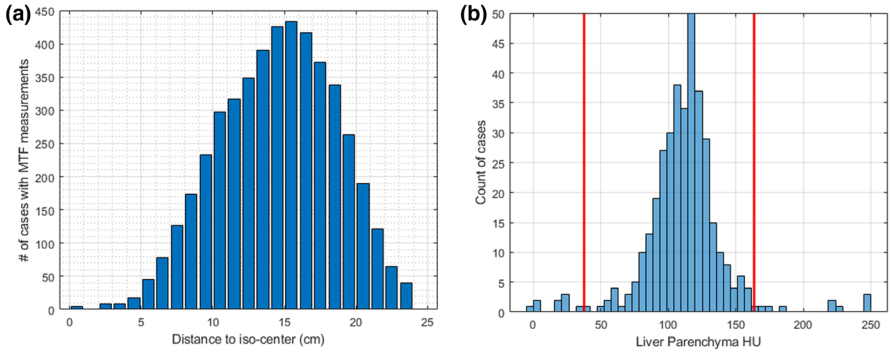

FIG. 1. Bar plot of number of cases with available MTF measurements vs distance of measurement points to isocenter. Max cases had MTF measurements at a distance of 15.5 cm to iso-center (a). Histogram plot of liver parenchyma HU values of valid liver CT images. The two vertical bars in red color mark the [2.5, 97.5] percentile range of the HU values (b). HU, Hounsfield Unit; MTF, Modulation Transfer Function. [Color figure can be viewed at wileyonlinelibrary.com]

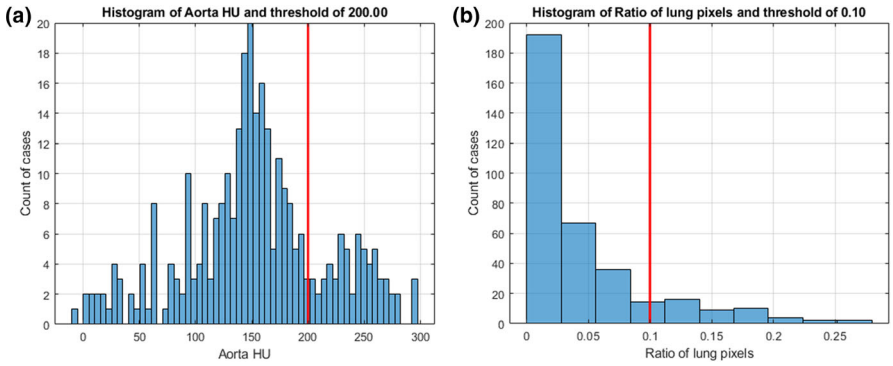

FIG. 2. Histogram plot of aorta HU values. The single vertical bar in red color marks the HU value of 200 (a). Histogram plot of ratio of lung pixels to body pixels. The single vertical bar in red color marks the threshold value of 0.1 (b). HU, Hounsfield Unit. [Color figure can be viewed at wileyonlinelibrary.com]

2.C. Selecting images for observer study

The selection of images for the observer study consisted of two parts. In the first part, all available clinical CT images were checked to ensure that they come from comparable imaging conditions. This was key to control confounding effects when comparing images from different clinical imaging conditions. After a bank of images from like clinical conditions was determined, those images were stratified into sets which varied in only one quality metric (liver parenchyma HU, clarity, and noise magnitude) at a time. Both steps are described below.

2.C.1. Ensuring consistent imaging conditions for image comparison

The slice location of the liver CT image was identified from the CT series by the algorithm of measuring liver parenchyma HU. As it identifies the liver parenchyma via matching the distribution of HU values within an algorithmically placed ROI to a Gaussian distribution, the ROI could be placed on a hepatic lesion and HU measurement from such an ROI is likely to be off the actual liver parenchyma HU. To limit the impacts of such unreliable measurements on algorithmic image quality assessment, the distribution of liver parenchyma HU values from all liver CT images were analyzed as shown in Figure 1(b) and those images with liver parenchyma HU values beyond the [2.5, 97.5] percentile range of the population were identified.

Besides the liver parenchyma HU measurement on the liver CT image, the algorithm also provided aorta HU measurement on a CT image at another slice location in the same series. The aorta HU measurement was also used for estimating the contrast phase of the liver CT image. A liver CT image was considered to be in the portal venous phase if the measured aorta HU value was lower than 200 [Fig. 2(a)]. To diminish the confounding effects due to less attenuation on image quality from a significant presence of patient lungs, a metric in terms of ratio of lung pixels to body pixels was evaluated for all liver CT images and the thresholding so that the liver CT images to be selected for the observer study did not have a significant presence of patient lungs [Fig. 2(b)]. After passing all checks noted above, 242 liver CT images remained to be selected for the observer study.

2.C.2. Stratifying images into consistent-quality sets

Our observer study aimed to ascertain the correspondence of algorithmic measurements of image quality metrics with observer preferences, one metric at a time. To

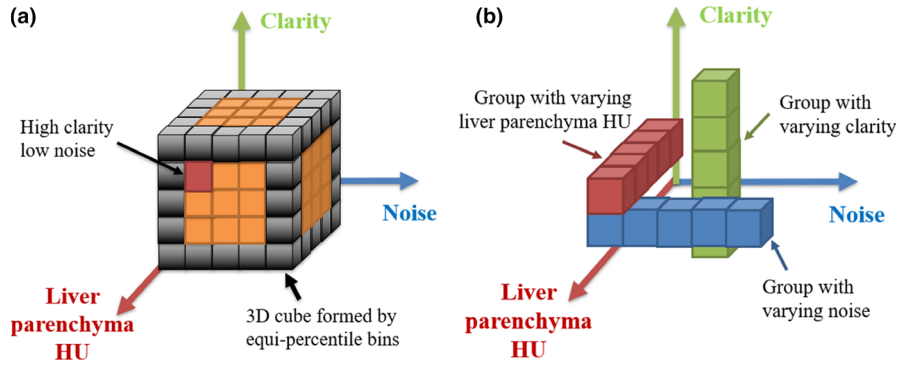

FIG. 3. The stratification scheme of the liver CT images by their image quality metric values. Each cube in the figure has a unique 3D stratification index (a). How a group of images that vary in terms of only one of the image quality metrics at a time is identified (b). CT, Computed Tomography. [Color figure can be viewed at wileyonlinelibrary.com]

diminish the confounding effects from multiple varying image quality metrics on the results, CT images were selected for the observer study from a group of images that vary in terms of only one image quality metric at a time. In this stratification process, along each of the three algorithmic image quality metrics (liver parenchyma HU, clarity, and noise magnitude), the range of metric values were divided into quintile bins and each bin included 20% of liver CT images. The collection of the three metric bin indexes of an image assigned the image to a 3D cube location. For example, if one liver CT image had liver parenchyma HU, noise magnitude, and clarity all within their respective 0%–20% bin (i.e., the first quintile), the image was assigned to the 3D cube location (1,1,1). If another liver CT image had liver parenchyma HU, noise magnitude, and clarity all within their respective 80%–100% bin (i.e., the fifth quintile), the image was assigned to the 3D cube location (5,5,5). In this way, each liver CT image was assigned to one and only one 3D cube location. The number of liver CT images assigned to a certain 3D cube location was counted and used as the basis for computing weights for data analysis. A set of liver CT image samples was drawn from a group of 3D cubes along a line parallel to the axis of the varying metric. Different sets of images were drawn from other groups of cubes where the two non-varying metric indexes took different pairs of values.

Figure 3 illustrates the stratification scheme of the liver CT images by their image quality metric values and how a group of images that vary in terms of only one image quality metric at a time is identified. Typically liver CT images in the lower extreme or upper extreme metric bins (e.g., bin index of 1 or 5) had considerably a wider range of metric values than those in the middle three bins due to existence of outliers. In this study, liver CT images were only drawn from cubes where the two nonvarying dimensions were not on either extreme (i.e., bin index of 2,3,4) and 151 images met the criteria. The metrics were tagged as low, medium, and high with bin index of 2, 3, and 4 respectively. If a cube had an index of 4 in clarity and index of 2 in noise, it was tagged as high clarity low noise.

When a set of image samples was drawn from the identified group that varied in terms of only one image quality metric, the set of images represented the full range of the varying metric in the group. The metric value difference between two adjacent images in the same set was kept above a minimum threshold (15% of the median metric value across images in the same group) to provide unambiguous perceptual difference to observers. Each set contained up to six liver CT images. If no more than one image could be drawn from a group, no image was drawn from that group. As a result, 23 sets of liver CT images were drawn (9 for noise magnitude, 9 for liver parenchyma HU, and 5 for clarity). All images were saved in an uncompressed format (PNG files) and displayed at the original image dimension (512 × 512 pixels) at the same abdominal window (window width 500 HU and window level/center 55 HU).

2.D. Observer study

An observer study with clinical radiologists was carried out to investigate the correspondence of algorithmic image quality metrics (noise, liver parenchyma HU, and clarity) and the observer rankings of perceptual image quality. The observers were instructed to rank the lesion-free images with respect to their quality for detecting small hypothetical noncalcified hepatic lesions. Figure 4 shows the user interface of the observer study. It was a modified version of the web interface used in a previous image quality investigation of chest radiographs.[23] The 23 sets of image files were divided into three investigation classes, varying image quality in each class with respect (a) to noise, (b) liver parenchyma HU, and (c) clarity. For example, there were nine sets for noise. In each set, the web interface displayed a set of reduced-size CT images that varied only in noise in random order. The observer was instructed to sort the displayed CT images by dragging one image with the computer mouse at a time and placing it in the order of their perception of noise level from more preferred to less preferred. The observer was instructed to sort images by noise only and not any other image quality features. The observer moved the mouse cursor over any reduced-size

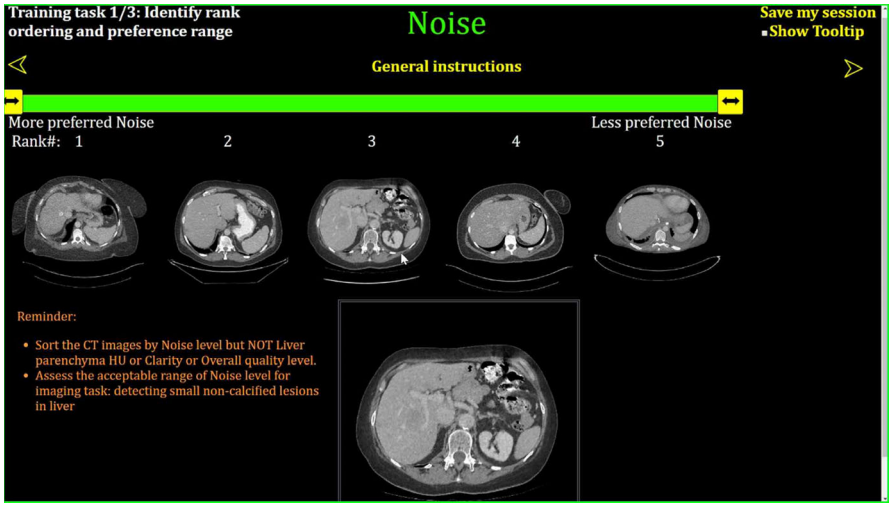

FIG. 4. Screenshot of the web interface for the observer study. The observer moved the mouse cursor over any reduced-size image to display the full-size image below. [Color figure can be viewed at wileyonlinelibrary.com]

image to view the full-size image in the box near the bottom of the web page. No further manipulation of the image was provisioned in the interface to provide a consistent presentation of the images across observers.

After sorting the images, the observer was instructed to further move the slider handles on top of the images to select which images fall within their acceptable range of the metric (e.g., noise). The slider handle could be placed right on top of a displayed image or between two displayed images. After finishing one set, the observer would advance to the next set. To get observers familiar with the web interface, there were three training sets at the beginning of the observer study, one for each of the three investigation classes.

Seven radiologists with considerable experience in reading CT images participated in the observer study using their diagnostic display workstation. Both observer data of rank ordering and acceptable ranges of metrics were recorded for the subsequent analysis.

2.E. Data analysis

Observer data of rank ordering of liver CT images were analyzed to establish the reference assessment for perceptual image quality. For each set of displayed liver CT images, rank-ordering data were averaged across all observers to obtain reference rankings of the images for that set. The agreement between reference and algorithmic rankings of these images were investigated in terms of rank correlation coefficient using Kendall's Tau method.[24] For example, nine Kendall's Tau rank correlation coefficients were computed for the sets of images with varying noise that span nine categories from low liver parenchyma HU and low clarity to high liver parenchyma HU and high clarity. These category-specific rank correlation coefficients were further combined with weights of the categories (proportional to the count of liver CT images assigned to a category) to compute a weighted mean rank correlation coefficient, which was the overall rank-order agreement for noise. The same analysis was also applied to assess overall rank-order agreements for liver parenchyma HU and clarity.

Observer data of acceptable ranges of metrics in terms of image ranks were analyzed to characterize clinically acceptable range of an image quality metric for abdominal CT examinations in terms of the algorithmic metric values. Provided the mapping between reference rankings and algorithmic metric measurements on images of a certain category, acceptable ranges in image ranks were translated to algorithmic metric values for each observer by linear interpolation. These observer-specific acceptable ranges of the same category were combined to obtain the mean and median observers' acceptable range of a metric. The 95% Confidence Interval (CI) on either limit of the mean range was estimated assuming the acceptable ranges of observer population were normally distributed. The end of the observers' acceptable range with lower metric value was taken as the lower bound threshold of the metric for an image category. The other end of the range was taken as the upper bound threshold for the same image category. These category-specific metric thresholds were further combined with weights of the categories to compute the overall clinically acceptable thresholds for image quality metrics in terms of algorithmic metric values.

The detectability index thresholds of different image categories were estimated using Eq. (3), combining the three metrics (liver parenchyma HU, noise magnitude and clarity). The lesion contrast was taken as 20% of the liver parenchyma HU for a noncalcified hepatic lesion. Detectability index estimated in this way would be proportional to one varying metric when the other two metrics are held unchanged. Overall clinically acceptable detectability index thresholds for detecting small non-calcified hepatic lesions were further computed by combining category-specific thresholds and weights of the categories.

The observer data for one of the clarity investigation tasks were excluded from the data analysis as this task

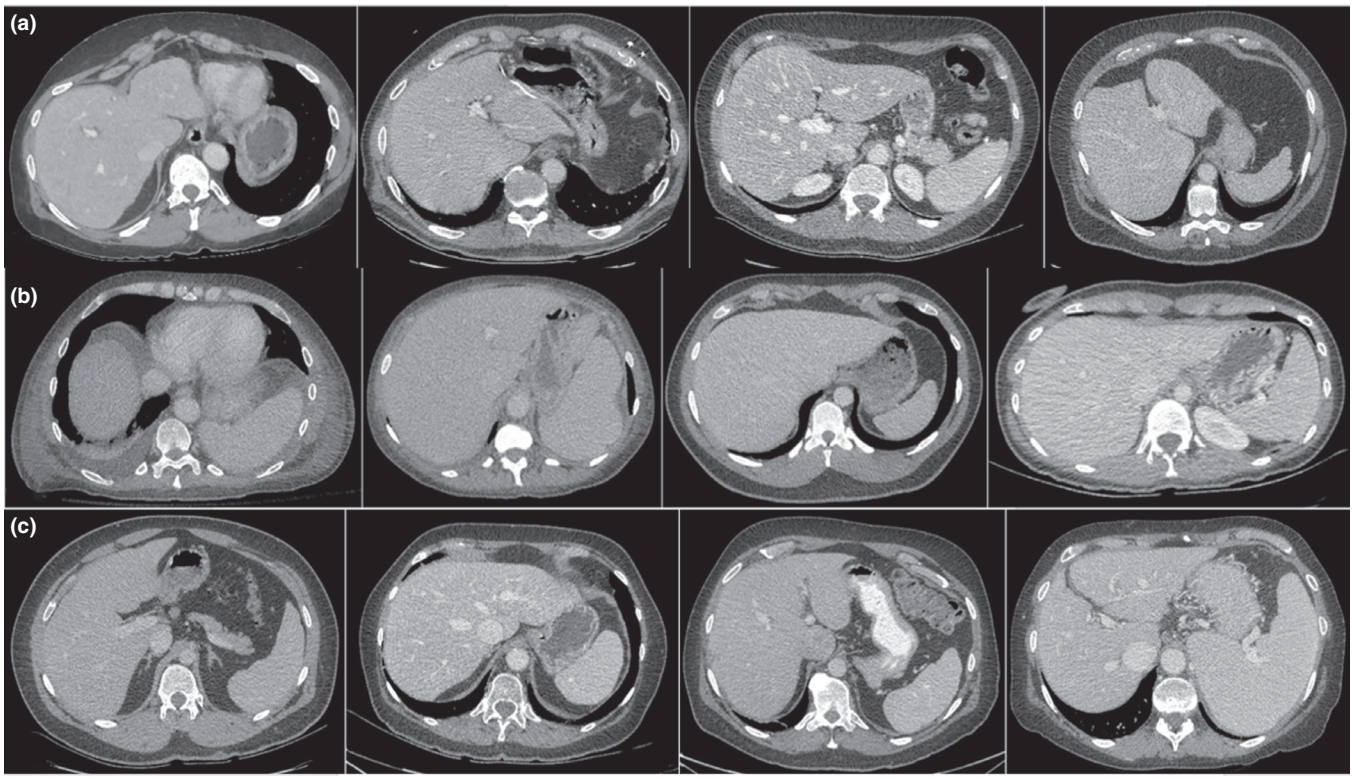

FIG. 5. Examples of liver CT images selected for observer study: varying noise magnitude (a), varying liver parenchyma HU (b), and varying clarity (c). HU, Hounsfield Unit.

had a set of liver CT images that showed strong perceptual difference in liver vasculature enhancement. Vasculature enhancement rather than image clarity had the biggest influence on perceptual image quality in this set of images. The resulting observer rankings reflected observer preference of vasculature enhancement, but not algorithmic clarity measurement and therefore were excluded.

3. RESULTS

Figure 5 shows example clinical images selected for the observer study based on their algorithmic image quality measurements. Images in the same row from left to right is for the same investigation class in the order of algorithmic metric values from low to high. Top row shows liver CT images with varying noise magnitude, middle row liver parenchyma HU, and bottom row clarity. They include images acquired with scanners from both manufacturers (GE Healthcare and Siemens Healthineers).

The overall rank-order agreement between algorithmic measurements and radiologists' assessments for the three image quality metrics was (perfect agreement has value of 1): noise magnitude 0.90, liver parenchyma HU 0.98, and clarity 1.00. As the overall agreements were greater than or equal to 0.9 across all metrics, the results indicated a strong rank-order agreement between algorithmic and observer assessments of image quality.

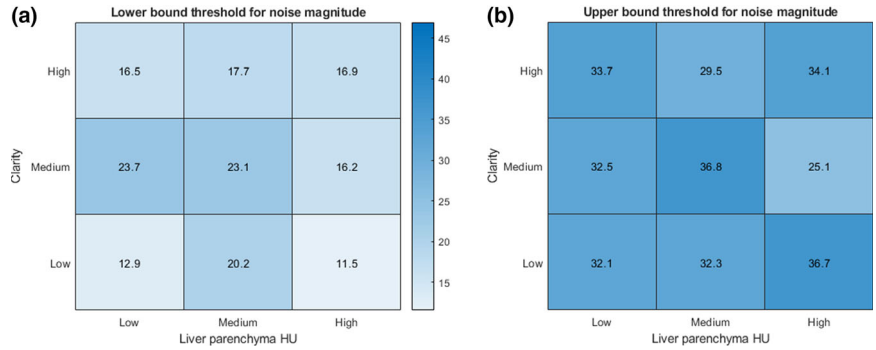

FIG. 6. Observers' mean acceptable ranges of noise magnitude across image categories (a) lower bound threshold (b) upper bound threshold. [Color figure can be viewed at wileyonlinelibrary.com]

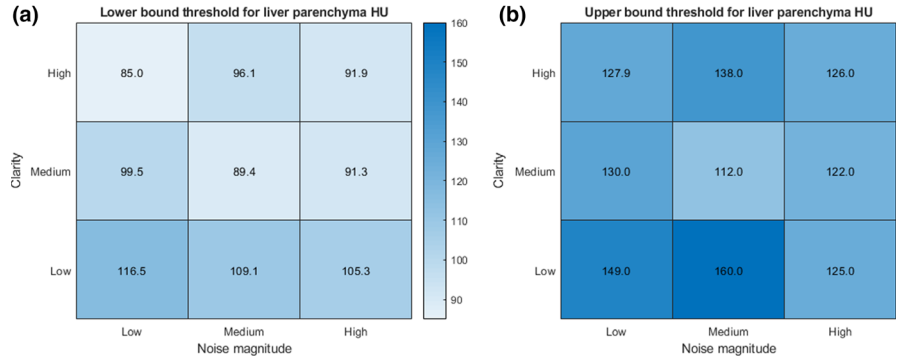

FIG. 7. Observers' mean acceptable ranges of liver parenchyma HU across image categories (a) lower bound threshold (b) upper bound threshold. HU, Hounsfield Unit. [Color figure can be viewed at wileyonlinelibrary.com]

Figure 6 shows observers' mean lower bound thresholds of noise magnitude across image categories as well as upper bound thresholds. The upper bound thresholds of noise magnitude in most of the categories were over 30 HU, while the lower bound thresholds in most of the categories were below 20 HU. The few categories that had upper bound thresholds below 30 HU/lower bound thresholds above 20 HU were due to the maximum/minimum noise magnitude of images in those categories showing a limited confounding effect of liver parenchyma HU and clarity on noise preferences of radiologists.

Figure 7 shows the value of mean thresholds of liver parenchyma HU across image categories. The lower bound thresholds of liver parenchyma HU in image categories with low clarity were over 100 while categories with medium and high clarity had thresholds below 100. The upper bound thresholds in all but one category was above 120. Clarity and noise of images across categories had a limited confounding effect on liver parenchyma HU preferences of radiologists.

Figure 8 shows the same charts for mean thresholds of clarity across image categories. The lower bound thresholds of clarity were consistently over 0.43 in different categories while the upper bound thresholds were over 0.48. A few image categories had no reported thresholds of clarity because no observer data in those categories were available for data analysis either due to no clinical images selected for observer study as described in Section 2.C or data exclusion described in Section 2.E. If these image categories with missing data were omitted, the computed overall thresholds for clarity would be biased toward categories with reported thresholds. Instead we imputed the clarity threshold of such an image category from the median clarity value of clinical images in the same category. If for a certain category only one image had been displayed to the readers, their acceptable range of clarity would be the same as the clarity value of the displayed image, which is well represented by the median metric value of clinical images in the same category.

The clinically acceptable ranges of detectability index for detecting small noncalcified hepatic lesions were estimated to be (0.34, 0.64), (0.27, 0.39), and (0.32, 0.35) for varying noise magnitude, liver parenchyma HU, and clarity, respectively. The strong correlation between the varying metric and perceptual image quality also indicated a strong correlation between detectability index and perceptual image quality.

Table I summarizes the overall rank-order agreement and clinically acceptable range for the three image quality metrics for abdominal CT examinations investigated in this study. Both the median and mean acceptable range were reported in the form of (lower bound threshold, upper bound threshold) for each image quality metric. The 95% CI of both lower bound and upper bound thresholds for their mean across image categories were also reported.

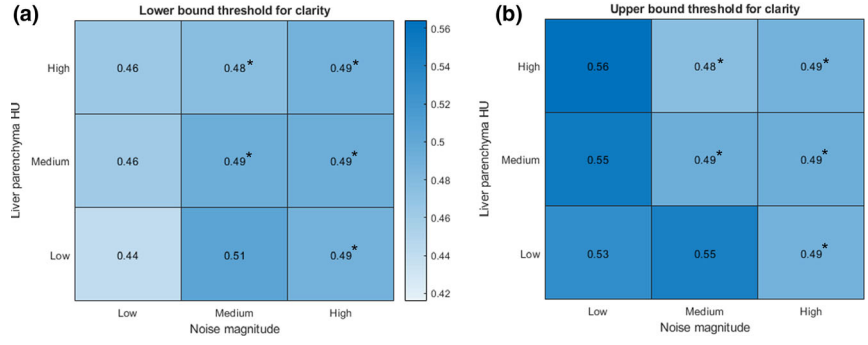

FIG. 8. Observers' mean acceptable ranges of clarity across image categories (a) lower bound threshold (b) upper bound threshold. Image categories with the imputed clarity thresholds have an asterisk mark next to the threshold value and have the same lower and upper bound threshold values. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE I. Overall rank-order agreement and clinically acceptable range of image quality metrics for abdominal CT examination in terms of algorithmic metric value.

	Overall rank order agreement		Clinically acceptable range of image quality metricsfor abdominal CT examination			
Metric	Mean	Std error	Median	Mean	95% CI Lower bound	95% CI Upper bound
Noise magnitude (HU)	0.90	0.06	(17.8, 32.6)	(17.8, 32.5)	(17.8, 17.8)	(28.0, 37.1)
Liver parenchyma HU	0.98	0.19	(92.1, 131.9)	(97.2, 131.8)	(82.8, 111.5)	(131.6, 132.1)
Clarity	1.00	0.22	(0.47, 0.52)	(0.48, 0.51)	(0.47, 0.49)	(0.51, 0.52)

4. DISCUSSION

Assuring quality of clinical CT images is essential for assuring their effective use in the care of patients. Techniques have been developed in the past to assess image quality of clinical CT images largely through inferring from measurements on images of standardized phantoms. While these techniques can be objective and straightforward to implement, their measurements are more reflecting the technical capabilities of imaging systems than the actual quality of clinical CT images. Image quality measurements on clinical CT images are indispensable for assuring their quality. However, given the massive number of clinical CT images generated in clinical practice every day, manually assessing image quality of clinical CT images as a day-to-day task is intractable and ineffective.

Automated algorithms assessing image quality of clinical images open the possibility of effective image quality management of clinical CT images provided that the correspondence from the output of the algorithms to clinical expectations is ascertained. In the present study, we aimed to address this challenge by validating algorithmic image quality measurements with clinical preferences, and further determining the clinically acceptable range of algorithmic measurements. Sets of clinical CT images were randomly selected from a multicenter dataset across image quality levels. Quality of clinical CT images (noise magnitude, liver parenchyma HU and clarity) were assessed by automated algorithms and the assessments were compared to rankings by clinical radiologists in an observer study. The results indicated a strong rank-order agreement between algorithmic and observer assessments of image quality. The study further determined clinically acceptable thresholds of image quality metrics in terms of algorithmic metric values. The clinically acceptable thresholds of one image quality metric were derived from observer data across different image categories formed by possible combinations of the other two metrics. The overall clinically acceptable thresholds of the metric are the weighted averages of its category-specific thresholds and independent on the values of the other metrics. For example, a clinical CT image that has its noise magnitude within the corresponding acceptable thresholds is deemed with clinically acceptable noise magnitude but not necessarily with acceptable liver parenchyma HU or clarity. These thresholds can be utilized for establishing diagnostic reference levels[25,26] in terms of clinically acceptable perceptual image quality. These quantitative definitions of image quality acceptance can support optimization of clinical image acquisition parameters.

This study investigates the correspondence between clinically preferred CT image quality and algorithmic measurements and has its limitation. Its reference assessments of CT image quality are in terms of perceptual quality of an image as assessed by clinicians and the reference assessments are not based on clinical task performance. We rely on the general assumption that CT images with quality preferred by clinicians will also likely be optimal for their task performance. We are not aware of any studies that contradict this assumption. The clinically acceptable ranges of image quality for abdominal CT examinations were determined from random samples of adult patients and would generalize to adult patient population. However, for another type of CT examination on pediatric or smaller patients, a different acquisition protocol (e.g., with a lower kV) is typically used, which would impact the clinically acceptable range of image quality (e.g., different organ HU range may allow the observer to tolerate wider range of noise magnitude). It will be beneficial extending the investigated method to a wider patient population and different types of examinations to determine their corresponding clinically acceptable range of image quality.

Based on the findings of this study, a few additional investigations are of merit. As the algorithmic assessments of several essential image quality metrics of clinical CT images were validated with preferences of radiologists, it will be interesting to investigate the correlation of the algorithmic assessments with clinical task performance across a variety of tasks. Furthermore, the algorithmic measurements of image quality metrics may be combined with dose metrics to support the optimization of dose and image acquisition parameters to minimize clinical risks.[2] They have the potential to greatly facilitate data-driven monitoring and optimization of CT image acquisition protocols.

5. CONCLUSION

Assuring quality of clinical CT images is essential for assuring their effective use in the care of patients. In this study, we investigated the utility of several computer algorithms to measure image quality on clinical CT images and compared the algorithmic measurements with observer

rankings. The observer study results indicated that these algorithms can robustly assess the perceptual quality of clinical CT images in an automated fashion. Clinically acceptable ranges of algorithmic measurements were determined. The correspondence of these image quality assessment algorithms to clinical expectations was established. The correspondence has the potential to greatly facilitate data-driven optimization of CT image acquisition protocols and to assure quality clinical CT images are continuously delivered.

ACKNOWLEDGMENTS

The authors thank Dr. David Enterline, Dr. Donald Frush, Dr. Rendon Nelson, Dr. Francesca Rigiroli, and Dr. Qi Wang for their passionate participation in the observer study, and Dr. Aiping Ding and Dr. Justin Solomon for facilitating access to clinical CT series. This work was supported in part by Siemens Healthineers.

CONFLICTS OF INTEREST

The authors have no conflicts to disclose.

[a)]Author to whom correspondence should be addressed. Electronic mail: ehsan.samei@duke.edu.

REFERENCES

1. Barrett HH, Myers KJ, Hoeschen C, Kupinski MA, Little MP. Task-based measures of image quality and their relation to radiation dose and patient risk. *Phys Med Biol*. 2015;60:R1–75.
2. Samei E, Jarvinen H, Kortesniemi M, et al. Medical imaging dose optimisation from ground up: expert opinion of an international summit. *J Radiol Prot*. 2018;38(3):967–989.
3. Paolicchi F, Faggioni L, Bastiani L, et al. Optimizing the balance between radiation dose and image quality in pediatric head CT: findings before and after intensive radiologic staff training, AJR. *Am J Roentgenol*. 2014;202:1309–1315.
4. Niiniviita H, Kiljunen T, Kulmala J. Comparison of Effective Dose and Image Quality for Newborn Imaging on Seven Commonly Used CT Scanners. *Radiat Prot Dosimetry*. 2017;174:510–517.
5. Bushberg JT, Seibert A, Leidholdt EM, Boone JM. *The essential physics of medical imaging*. 3rd edn. Philadelphia, PA: LWW; 2012.
6. Swets JA. Measuring the accuracy of diagnostic systems. *Science*. (New York, N.Y.). 1988;240:1285–1293.
7. American College of Radiology (ACR). *CT Accreditation Program Requirements*. http://www.acraccreditation.org/~/media/ACRAccreditation/Documents/CT/Requirements.pdf?la=en. Accessed September 10, 2018.
8. Larson DB, Wang LL, Podberesky DJ, Goske MJ. System for verifiable CT radiation dose optimization based on image quality. part I. Optimization model. *Radiology*. 2013;269:167–176.
9. Ria F, Wilson JM, Zhang Y, Samei E. Image noise and dose performance across a clinical population: patient size adaptation as a metric of CT performance. *Med Phys*. 2017;44:2141–2147.
10. Euler A, Solomon J, Marin D, Nelson RC, Samei E. A third-generation adaptive statistical iterative reconstruction technique: phantom study of image noise, spatial resolution, lesion detectability, and dose reduction potential, AJR. *Am J Roentgenol*. 2018;210:1301–1308.
11. Samei E, Richard S. Assessment of the dose reduction potential of a model-based iterative reconstruction algorithm using a task-based performance metrology. *Med Phys*. 2015;42:314–323.
12. Gomez-Cardona D, Li K, Hsieh J, et al. Can conclusions drawn from phantom-based image noise assessments be generalized to in vivo studies for the nonlinear model-based iterative reconstruction method? *Med Phys*. 2016;43:687–695.
13. Malkus A, Szczykutowicz TP. A method to extract image noise level from patient images in CT. *Med Phys*. 2017;44:2173–2184.
14. Gong H, Yu L, Leng S, et al. A deep learning- and partial least square regression-based model observer for a low-contrast lesion detection task in CT. *Med Phys*. 2019;46:2052–2063.
15. Abadi E, Sanders J, Samei E. Patient-specific quantification of image quality: An automated technique for measuring the distribution of organ Hounsfield units in clinical chest CT images. *Med Phys*. 2017;44:4736–4746.
16. Chen B, Christianson O, Wilson JM, Samei E. Assessment of volumetric noise and resolution performance for linear and nonlinear CT reconstruction methods. *Med Phys*. 2014;41:071909.
17. Christianson O, Winslow J, Frush DP, Samei E. Automated technique to measure noise in clinical CT examinations, AJR. *Am J Roentgenol*. 2015;205:W93–99.
18. Sanders J, Hurwitz L, Samei E. Patient-specific quantification of image quality: An automated method for measuring spatial resolution in clinical CT images. *Med Phys*. 2016;43:5330.
19. Solomon J, Mileto A, Nelson RC, Roy Choudhury K, Samei E. quantitative features of liver lesions, lung nodules, and renal stones at multi-detector row CT examinations: dependency on radiation dose and reconstruction algorithm. *Radiology*. 2016;279:185–194.
20. Smith TB, Solomon J, Samei E. Estimating detectability index in vivo: development and validation of an automated methodology. *J Med imaging* (Bellingham, Wash.). 2018;5:031403.
21. Dolly S, Chen HC, Anastasio M, Mutic S, Li H. Practical considerations for noise power spectra estimation for clinical CT scanners. *J Appl Clin Med Phys*. 2016;17:392–407.
22. McCollough C, Bakalyar DM, Bostani M, et al. Use of water equivalent diameter for calculating patient size and size-specific dose estimates (SSDE) in CT: the report of AAPM task group 220. *AAPM report*. 2014;2014:6–23.
23. Samei E, Lin Y, Choudhury KR, McAdams HP. Automated characterization of perceptual quality of clinical chest radiographs: validation and calibration to observer preference. *Med Phys*. 2014;41:111918.
24. Gibbons JD, Chakraborti S. *Nonparametric Statistical Inference*, 5th edn. Boca Raton, FL: CRC Press; 2010.
25. International Atomic Energy Agency. *Radiation Protection of Patients (RPOP) – Diagnostic Reference Levels (DRLs),* https://www.iaea.org/resources/rpop/health-professionals/radiology/diagnostic-reference-levels. IAEA, 2017.
26. Vano E, Miller DL, Martin CJ, et al. ICRP Publication 135: Diagnostic Reference Levels in Medical Imaging. *Annals of the ICRP*. 2017;46:1–144.