# A New Algorithm for Automatically Calculating Noise, Spatial Resolution, and Contrast Image Quality Metrics

# A New Algorithm for Automatically Calculating Noise, Spatial Resolution, and Contrast Image Quality Metrics

## *Proof-of-Concept and Agreement With Subjective Scores in Phantom and Clinical Abdominal CT*

*Cécile R.L.P.N. Jeukens, PhD,* Maikel T.H. Brauer, MSc,*†‡ Casper Mihl, MD, PhD,*§ Emmeline Laupman, MD,|| Estelle C. Nijssen, PhD,* Joachim E. Wildberger, MD, PhD,*§ Bibi Martens, MD,*§ and Carola van Pul, PhD†‡*



**Objectives:** The aims of this study were to develop a proof-of-concept computer algorithm to automatically determine noise, spatial resolution, and contrast-related image quality (IQ) metrics in abdominal portal venous phase computed tomography (CT) imaging and to assess agreement between resulting objective IQ metrics and subjective radiologist IQ ratings.

**Materials and Methods:** An algorithm was developed to calculate noise, spatial resolution, and contrast IQ parameters. The algorithm was subsequently used on 2 datasets of anthropomorphic phantom CT scans, acquired on 2 different scanners (n = 57 each), and on 1 dataset of patient abdominal CT scans (n = 510). These datasets include a range of high to low IQ: in the phantom dataset, this was achieved through varying scanner settings (tube voltage, tube current, reconstruction algorithm); in the patient dataset, lower IQ images were obtained by reconstructing 30 consecutive portal venous phase scans as if they had been acquired at lower mAs. Five noise, 1 spatial, and 13 contrast parameters were computed for the phantom datasets; for the patient dataset, 5 noise, 1 spatial, and 18 contrast parameters were computed. Subjective IQ rating was done using a 5-point Likert scale: 2 radiologists rated a single phantom dataset each, and another 2 radiologists rated the patient dataset in consensus. General agreement between IQ metrics and subjective IQ scores was assessed using Pearson correlation analysis. Likert scores were grouped into 2 categories, "insufficient" (scores 1–2) and "sufficient" (scores 3–5), and differences in computed IQ metrics between these categories were assessed using the Mann-Whitney *U* test.

**Results:** The algorithm was able to automatically calculate all IQ metrics for 100% of the included scans. Significant correlations with subjective radiologist ratings were found for 4 of 5 noise ($R^2$ range = 0.55–0.70), 1 of 1 spatial resolution ($R^2$ = 0.21 and 0.26), and 10 of 13 contrast ($R^2$ range = 0.11–0.73) parameters in the phantom datasets and for 4 of 5 noise ($R^2$ range = 0.019–0.096), 1 of 1 spatial resolution ($R^2$ = 0.11), and 16 of 18 contrast ($R^2$ range = 0.008–0.116) parameters in the patient dataset. Computed metrics that significantly differed between "insufficient" and "sufficient" categories were 4 of 5 noise, 1 of 1 spatial resolution, 9 and 10 of 13 contrast parameters for phantom the datasets and 3 of 5 noise, 1 of 1 spatial resolution, and 10 of 18 contrast parameters for the patient dataset.

**Conclusion:** The developed algorithm was able to successfully calculate objective noise, spatial resolution, and contrast IQ metrics of both phantom and clinical abdominal CT scans. Furthermore, multiple calculated IQ metrics of all 3 categories were in agreement with subjective radiologist IQ ratings and significantly differed between "insufficient" and "sufficient" IQ scans. These results demonstrate the feasibility and potential of algorithm-determined objective IQ. Such an algorithm should be applicable to any scan and may help in optimization and quality control through automatic IQ assessment in daily clinical practice.

**Key Words:** computed tomography, objective image quality, noise, spatial resolution, contrast, automatic quantification, abdominal CT

(*Invest Radiol* 2023;58: 649–655)

Computed tomography (CT) has become an indispensable diagnostic tool in daily clinical practice,[1,2] and advances in the technology are ongoing.[3] A drawback of CT is the use of ionizing radiation, exposure to which may become high, in particular when accumulating from multiple scans.[4–6] Radiation dose optimization is necessary and even mandatory: radiation exposure per CT scan needs to be reduced while maintaining sufficient diagnostic image quality (IQ).[7,8] Finding the optimum balance between patient radiation exposure and IQ for each CT procedure is time consuming and not straightforward, particularly as there is no universal measure for clinical IQ.

Research in the field of IQ optimization commonly uses scoring by radiologists to determine diagnostic IQ. However, perceived IQ is subjective and depends on the clinical question at hand. Furthermore, scoring scans is time consuming. What would be ideal is a set of objective, quantitative IQ parameters that can be automatically determined. Obvious candidates are signal-to-noise ratio and contrast-to-noise ratio (CNR) in particular regions of interest, but these are difficult to automate.[9] More advanced metrics have been developed for quality control testing on physical phantoms, typically used in scanner performance quality checks.[10,11] Such phantoms contain structures for quantitative measurement of physical IQ aspects such as noise, spatial resolution, and contrast in a reproducible way and for different scanner settings, but they do not resemble human anatomy and are therefore not directly comparable to clinical images. Consequently, the relationship between phantom-based metrics and clinical IQ is not straightforward. Furthermore, iterative reconstruction measurements may be non-linearly influenced, hampering validation efforts.[10,12]

A promising approach to objectively quantify clinical IQ is the use of model observers. These are mathematical models developed for digital detection tasks such as a low-contrast object in phantom images that contain a certain background structure.[11] Detection by such a

Received for publication October 26, 2022; and accepted for publication, after revision, December 19, 2022.

From the *Department of Radiology and Nuclear Medicine, Maastricht University Medical Center, Maastricht; †Department of Applied Physics, Eindhoven University of Technology, Eindhoven; ‡Department of Medical Physics, Máxima Medical Centre, Veldhoven; §CARIM School for Cardiovascular Diseases, Maastricht University, Maastricht; and ||Department of Radiology, Máxima Medical Centre, Veldhoven, the Netherlands.

Cécile R.L.P.N. Jeukens and Maikel T.H. Brauer contributed equally to this article and are co–first authors.

Conflicts of interest and sources of funding: Siemens Healthineers provided the prototype software to perform this study (version 13.0.0.1, prototype software, Siemens Healthineers, Forchheim, Germany). In addition, the following authors declare relationships with the following companies: C. Mihl and B. Martens receive personal fees (speakers bureau) from Bayer, all outside the submitted work. J.E. Wildberger reports institutional grants from Bard, Bayer, Boston, Brainlab, GE, Philips, and Siemens and personal fees (speakers bureau) from Bayer and Siemens, all outside the submitted work.

Correspondence to: Cécile R.L.P.N. Jeukens, PhD, Department of Radiology and Nuclear Medicine, Maastricht University Medical Center, P. Debyelaan 25, PO Box 5800, 6202 AZ, Maastricht, the Netherlands. E-mail: Cecile.jeukens@mumc.nl.

Supplemental digital contents are available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.investigativeradiology.com).



ISSN: 0020-9996/23/5809–0649
DOI: 10.1097/RLI.0000000000000954

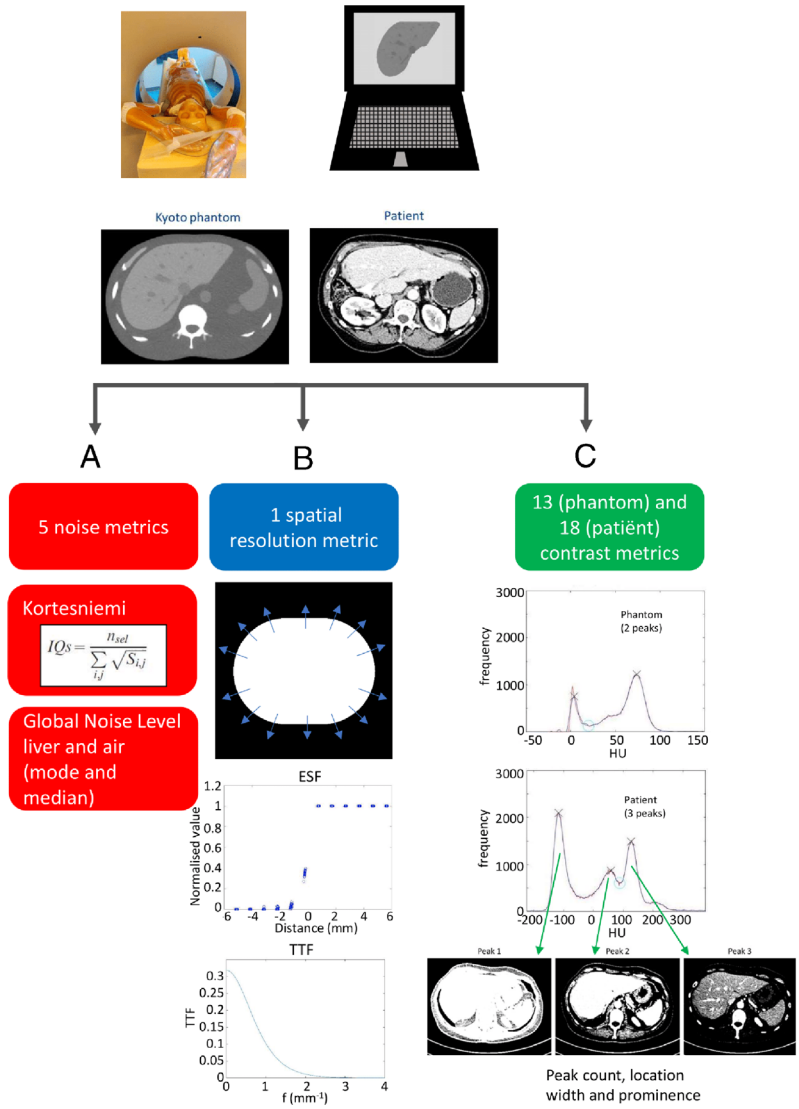


**FIGURE 1.** Pipeline of the algorithm for automatic image quality assessment: noise, spatial resolution, and contrast metrics (see text for more details). Phantom and patient-based image datasets were used. A slice in the liver region was manually selected. Five noise metrics (a) and 1 spatial resolution metric (b) were based on known values from literature; multiple contrast metrics (c) were newly developed for this study. The bottom 3 images in (c) illustrate the soft tissues building up the 3 peaks. The gray voxels are within the peak; the black and white voxels are outside the peak.

computer model is objective and will more closely mimic clinical tasks than measuring scanner performance with quality control phantoms. The technique allows for varying scanner settings, but existing models have been developed for simple tasks, and more studies are necessary to determine their accuracy in emulating radiological diagnosis.[11,13–16]

A method enabling direct, automatic, and objective IQ analysis of clinical images would overcome the above limitations and would constitute a major step forward in the field of IQ optimization. Such a method would provide opportunities for larger-scale automatic IQ assessment in daily clinical routine, enabling direct detection of decreasing IQ. Furthermore, artificial intelligence software for diagnostic support is becoming more prominent in clinical practice and requires sufficient IQ to function optimally; therefore, hospitals are required to monitor IQ. Finally, any protocol optimization would benefit from such an objective and time-efficient IQ assessment. Several studies have investigated automatically calculated IQ metrics for noise[17–20] or spatial resolution[21,22] of clinical images. These studies show a relationship between such automatically determined metrics and radiologist IQ assessment; however, most studies included only 1 aspect of IQ, whereas it is generally accepted that subjective IQ depends on a combination of the 3 aspects: noise, spatial resolution, and contrast.[23]

The aims of this study were to develop a proof-of-concept computer algorithm able to automatically calculate objective IQ metrics covering all 3 IQ aspects (noise, spatial resolution, and contrast) of clinical portal venous abdominal CT scans and to assess agreement between resulting objective IQ metrics and subjective radiologist IQ ratings.

## MATERIALS AND METHODS

### Developing an Algorithm for Automatic IQ Assessment

A proof-of-concept computer algorithm was developed using MATLAB (version R2019a, copyright The MathWorks, Inc, New York, NY) to automatically calculate a set of IQ metrics. Multiple metrics were implemented to measure IQ aspects related to noise, spatial

resolution, and contrast (Fig. 1). For both phantom and patient CT scans, a single liver slice was analyzed to calculate these IQ metrics; to this end, the slice containing the largest area of liver tissue was manually selected by visual scrutiny. The noise and spatial resolution metrics used were based on metrics reported in literature; all contrast metrics, except for CNR, were newly developed for this study.



### Automatic IQ Metrics

Five noise metrics were implemented (Fig. 1a): the Kortesniemi IQ score[17] and 4 global noise level (GNL) values.[19,20] The Kortesniemi IQ score is calculated by shifting a 3 × 3 mask around 1 pixel to create 9 masks. For each mask, the standard deviation (noise) is calculated and the lowest value is retained. This is repeated for all pixels excluding areas of air and sharp transitions, that is, excluding pixels in the top 5% standard deviation and pixels below −500 Hounsfield units (HU). The Kortesniemi IQ score is inversely proportional to the sum of the square root of the calculated standard deviations. A lower Kortesniemi IQ score translates to increased noise in the image. Global noise level values are determined by calculating the standard deviation (noise) for a 7 × 7 mask around pixels in predefined HU ranges: soft tissue range, defined as 0–100 HU, and air range, defined as <−500 HU. A histogram of the standard deviations is plotted and GNL is derived from the median and mode of the histogram, resulting in 4 GNL values. Higher GNL values translate to increased image noise in the image.

One spatial resolution metric was implemented (Fig. 1b), calculated by combining 2 methods described in the literature,[21,22] to which normalization and averaging steps were added to generate a single IQ metric. First, the outline of the phantom or patient was automatically segmented. An edge spread function (ESF) was calculated along the body-air interface in the direction perpendicular to the interface, according to the method proposed in Sanders et al.[21] Small HU gradients, such as those caused by clothes, were excluded. Each ESF was normalized from 0 to 1 and shifted such that the ESF value was 0.5 in the middle of the transition. The ESFs were then stored into 12 equally spaced bins based on the radial distance between the body-air interface and the isocenter of the image. In each bin, the ESFs were stacked to obtain an oversampled ESF and resampled to 10% of the pixel width. For each bin, a function

$$\mathrm{ESF}(r) = \frac{1}{1 + \exp\left(-\frac{r}{m}\right)}$$

was fitted, where $r$ is the radial distance and $m$ is a fitting parameter.[22] Following the method of Ott et al,[22] all ESFs were analytically differentiated and Fourier transformed, from which the full width at half maximum (FWHM) of the task transfer function (TTF) was calculated as a function of the fitting parameter $m$:

$$\mathrm{FWHM} = \frac{2.17732}{\pi^2 m}$$

The spatial resolution metric was defined as the average FWHM of all bins. Higher values translate to better spatial resolution in the image.

A set of contrast metrics was developed based on HU histogram analyses (Fig. 1c). After removal of background air pixels, histograms of pixel values were generated using bin width of 2 HU and a Savitzky-Golay smoothing filter.[24] The histogram of the anthropomorphic phantom liver slice shows 2 peaks, whereas those of patient-based liver slices show 3 peaks (Fig. 1c).

These peaks encompass the soft tissues present, such as liver tissue, spleen, muscle, and subcutaneous fat. The range of the soft tissue peaks was empirically determined to be −224 to 376 HU by visual histogram analysis. Because contrast metrics are based on generated histograms, this means fewer contrast metrics can be developed for phantom than patient images. For each peak in the soft tissue region, the following metrics were calculated using the MATLAB function *findpeaks*[25]: location, prominence, pixel count, and full width measured at half prominence (FWHP). The same metrics were calculated for the minimum between the 2 right-most histogram peaks by inverting and repeating the analysis using the *findpeaks* function. In addition, CNR was calculated based on the following[26]:

$$\mathrm{CNR} = \frac{\mathrm{Contrast}}{\mathrm{Noise}} = \frac{|\mu_1 - \mu_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}},$$

with $\mu_1$ and $\mu_2$ representing the location of the last 2 peaks respectively and $\sigma_1$ and $\sigma_2$ representing the measured FWHP values of these peaks. In patient scans, the number of pixels with gray values between 200 and 300 HU was determined as this can be related to the amount of visible contrast resulting from the intravenously administered contrast agent. In total, this resulted in 13 and 18 contrast metrics per phantom and patient scan, respectively. The meaning of high or low values of the above contrast metrics for contrast in the image depends on peak characteristics.

## Evaluation of the Algorithm in Phantom and Patient CT Scans

To evaluate the developed algorithm, the above IQ metrics were determined for CT scans of an anthropomorphic phantom (5 noise, 1 spatial resolution, and 13 contrast metrics) and for clinical CT scans of patients (5 noise, 1 spatial resolution, and 18 contrast metrics). Two phantom datasets (1 phantom, 2 different CT scanners) and 1 patient dataset (multiple patients, 1 scanner) were collected, where each dataset contained multiple CT scans. The phantom datasets were used to evaluate algorithm performance on CT scans acquired using a wide range of mAs and kV values, which result in a wide range of IQs on the same phantom without introducing anatomical and pathological variation. The patient dataset enabled the testing of algorithm performance in a clinical setting. For thorough testing of the algorithm, broad IQ ranges were incorporated in the datasets, ranging from high to low.

### Anthropomorphic Phantom Datasets

An anthropomorphic phantom (PBU-60 phantom, Kyoto Kagaku Co, Ltd, Kyoto, Japan) was used for the acquisition of 2 datasets: 57 scans performed on a 256-slice CT scanner (Ingenuity, Philips Healthcare, Best, the Netherlands [CT1]) and 57 scans performed on a third-generation dual-source CT (DSCT) scanner (Somatom Force, Siemens Healthineers, Forchheim, Germany [CT2]). Varying scanner settings (tube voltage, tube current, reconstruction algorithm) ensured that the datasets included a range of high to low IQ scans. Scanner settings were identical on both scanners as much as possible: scan range pulmonary diaphragm to pelvis; helical scanning, tube voltage varied in 3 steps (80, 100, and 120 kVp); mAs values varied in 7 steps (200, 180, 160, 140, 100, 60, and only at 120 kV, 40); no automated tube current modulation; slice collimation 128 × 0.625 mm (CT1) and 192 × 0.6 mm (CT2); and pitch 1.4 for mAs ≤160, 1.3 for 180 mAs, 1.2 for 200 mAs (CT1) and 0.9 (CT2). Scans were reconstructed using filtered back projection and different iterative reconstruction strengths (iDose 3 and 6 [CT1] and Advanced Modeled Iterative Reconstruction [ADMIRE] 3 and 5 [CT2]) using a kernel B (CT1) and Br40d (CT2), and a slice thickness of 3 mm, with a reconstruction increment of 3 mm (CT1) and 2 mm (CT2), respectively.

### Patient Dataset

The clinical patient dataset was based on 30 consecutive CT liver scans of 30 unique patients in the portal venous phase. The dataset was collected for a previous study by our team.[27] Patients were scanned between September 2019 and February 2020 on a third-generation DSCT scanner (Somatom Force, Siemens Healthineers [CT2]). The contrast

medium protocol was determined using injection software (P3T; Bayer Healthcare, Berlin, Germany; dosing factor 0.4 g I/kg, injection duration 30 seconds). Inclusion criteria were as follows: helical scanning, use of automated tube current modulation (150 $mAs_{ref}$) and automated tube voltage selection (120 $kV_{ref}$), 192 × 0.6 mm slice collimation, pitch = 0.9, kernel Br40d, 3 mm slice thickness with a reconstruction increment of 2 mm, and performed at 90 kV.

To obtain images with a lower IQ and ensure that the dataset contained a range of low to high IQ, raw data of the 30 CT scans were transferred to ReconCT software (version 13.0.0.1, prototype software, Siemens Healthineers) and reconstructed as if acquired at a lower mAs, simulating CT scans containing more noise. This software was validated in a phantom study by Ellmann et al[28] and by our team.[27] Tube current (mAs) values were varied in 4 steps (60%, 70%, 80%, and 90% of the original mAs value), and at each mAs level, filtered back projection and iterative level 2, 3, and 4 reconstructions were made (ADMIRE, Siemens Healthineers). Thus, 16 simulated scans were reconstructed per patient, complemented by the original scans, yielding a total of 510 scans.

## Ethical Considerations

For the retrospective evaluation of anonymized patient CT scans, a waiver of consent was obtained from the local research ethics committee and the institutional review board (ref METC 2017-0250).

## Agreement With Subjective Radiologist IQ Assessment

Subjective IQ was scored for the diagnostic task malignancy detection (primary or follow-up) in an outpatient population using visual grading on a 5-point Likert scale (1 = very poor, 2 = poor, 3 = moderate, 4 = good, 5 = excellent). The datasets of anthropomorphic phantom scans were scored by abdominal radiologist (E.L.) with 8 years' clinical experience (CT1) and an abdominal radiologist (C.M.) with 9 years' experience (CT2). The patient dataset was scored in consensus by 2 abdominal radiologists (B.M. and C.M.) with 4 and 9 years' experience, respectively. Four anthropomorphic phantom and 20 patient test images were used to familiarize radiologists with the range of IQs present in the datasets. In the final scoring test, 57 phantom and 510 patient images were presented in random order (including the test images).

**TABLE 1.** Agreement Between Automatically Calculated Metrics and Subjectively Scored Image Quality in Phantom and Patient Abdominal CT Scans

| | Anthropomorphic Phantom Images* | | Patient Images† | |
|---|---|---|---|---|
| | Correlation With Subjective IQ Scores | Sufficient vs Insufficient IQ | Correlation With Subjective IQ Scores | Sufficient vs Insufficient IQ |
| | Adjusted $R^2$ (CT1/CT2) | $P$ (CT1/CT2) | Adjusted $R^2$ | $P$ |
| Noise metrics | | | | |
| Kortesniemi IQ score | 0.70/0.55 | <0.001/<0.001 | 0.019 | NS |
| GNL mode | 0.66/0.65 | <0.001/<0.001 | 0.062 | <0.001 |
| GNL median | 0.67/0.66 | <0.001/<0.001 | 0.020 | NS |
| GNL air mode | NS/NS | NS/NS | NS | 0.031 |
| GNL air median | 0.70/0.65 | <0.001/<0.001 | 0.096 | 0.003 |
| Resolution metric | | | | |
| FWHM TTF | 0.21/0.26 | 0.023/0.006 | 0.110 | <0.001 |
| Contrast metrics | | | | |
| Count peak 1 | NS/0.61 | NS/<0.001 | 0.053 | 0.013 |
| Count peak 2 | 0.70/0.30 | <0.001/<0.001 | NS | NS |
| Count peak 3 | – | | 0.015 | 0.004 |
| Peak 1 location | 0.24/0.14 | <0.001/0.008 | 0.071 | 0.002 |
| Peak 2 location | 0.14/NS | 0.012/NS | 0.018 | NS |
| Peak 3 location | – | | 0.055 | NS |
| Peak 1 FWHP | 0.11/0.64 | NS/< 0.001 | 0.029 | 0.007 |
| Peak 2 FWHP | 0.40/0.66 | <0.001/<0.001 | 0.008 | NS |
| Peak 3 FWHP | – | – | 0.116 | 0.011 |
| Peak 1 prominence | NS/0.58 | NS/<0.001 | 0.066 | <0.001 |
| Peak 2 prominence | 0.73/0.37 | <0.001/<0.001 | 0.017 | 0.035 |
| Peak 3 prominence | | – | 0.014 | 0.003 |
| Count minimum | 0.47/0.26 | <0.001/0.009 | 0.020 | NS |
| Minimum location | NS/NS | NS/NS | 0.041 | NS |
| Minimum FWHP | 0.22/NS | 0.010/NS | 0.012 | NS |
| Minimum prominence | 0.38/0.11 | 0.010/0.003 | 0.030 | NS |
| N pixels 200–300 HU | – | | NS | <0.001 |
| CNR | 0.61/0.43 | <0.001/<0.001 | 0.115 | <0.001 |

*Two phantom datasets were acquired: the first using a 256-slice CT scanner (CT1) and the second using a third-generation dual-source CT scanner (CT2).

†The patient dataset was acquired using a third-generation dual-source CT scanner.

CT indicates computed tomography; IQ, image quality; NS, not significant. IQ indicates image quality; GNL, global noise level; FWHM, full width at half maximum; FWHP, full width at half prominence; TTF, task transfer function; HU, Hounsfield units; CNR, contrast-to-noise ratio.

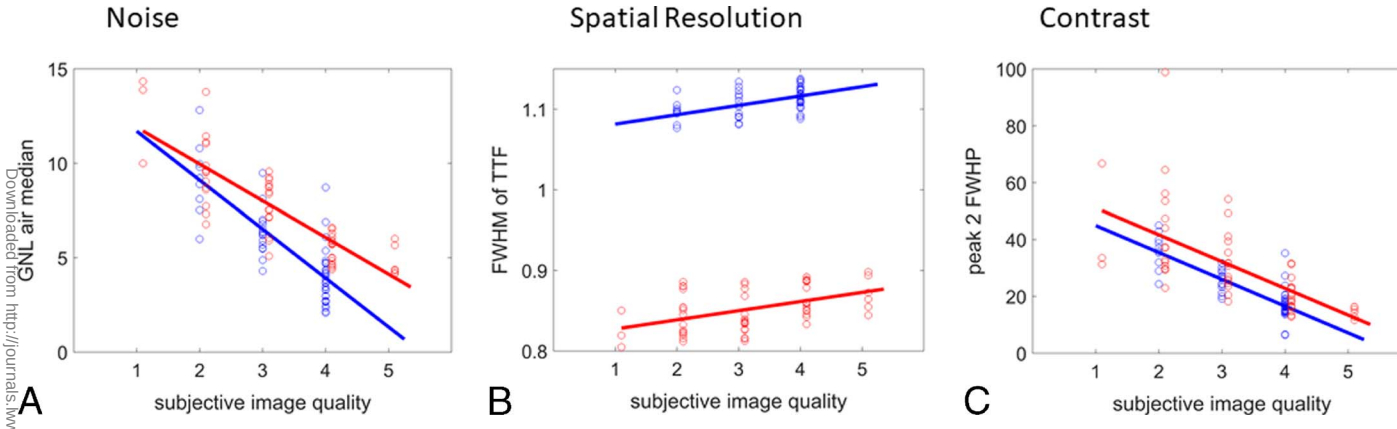


**FIGURE 2.** Correlations and significant linear regressions between automatically calculated parameters and subjective scores of image quality in 2 phantom datasets (n = 57 each). The best correlating parameter was selected for all 3 categories (see Supplemental Digital Content Figure 1, http://links.lww.com/RLI/A792, for all parameters). The 2 datasets were scanned on 2 different CT scanners: a 256-slice CT scanner (CT1, red) and a third-generation DSCT scanner (CT2, blue). A linear regression is shown for significant correlations only (adjusted $R^2$ values are given in Table 1). For visualization purposes, the red data points have been shifted up by 0.1 on the x-axis.



Radiologists were blinded to patient, scan, and reconstruction information. Scoring was conducted on a diagnostic workstation, with freedom provided to scroll through the scan volumes, zoom in or out, and change window settings.

Scans were later grouped into 2 categories based on resulting IQ scores: "insufficient" (Likert scores 1 and 2) and "sufficient" (Likert scores 3 to 5).

### Outcomes

Primary outcome for the proof of concept is the percentage of CT scans for which the algorithm can successfully calculate all IQ metrics.

The primary outcome for agreement with subjective radiologist ratings is the number of significant correlations between IQ metrics and subjective scores within each metric group (noise, spatial resolution and contrast). Secondary outcome on agreement is the number of IQ metrics within each metric group that differ between scans of "insufficient" and "sufficient" subjective IQ.

### Statistical Analysis

The correlation between each calculated IQ metric and subjective radiologist IQ score was determined using the Pearson correlation coefficient with a corresponding $P$ value. With $\alpha \leq 0.05$ and $\beta \leq 0.20$, the correlation was determined to be significant for the 57 phantom scans if the adjusted $R^2 \geq 0.14$ and for 510 patient scans if the adjusted $R^2 \geq 0.015$.

Significant correlations were inserted into figures as regression lines. Image quality metric differences between "insufficient" and "sufficient" categories were evaluated for each dataset using the Mann-Whitney $U$ test. Analyses were performed using MATLAB (MATLAB version R2019a, copyright The MathWorks, Inc). A $P$ value below 0.05 was considered to indicate statistical significance.

## RESULTS

### Proof of Concept

The developed algorithm was able to successfully calculate all noise, spatial resolution, and contrast IQ metrics in 100% of the CT scans of both phantom and patient datasets.

### Agreement: Correlations

The results of the correlation analyses between each IQ metric and subjective IQ scores are presented in Table 1. Figures 2 and 3 present the correlations and significant linear regression lines for a selected parameter in all 3 categories (see Supplemental Digital Content Figures 1

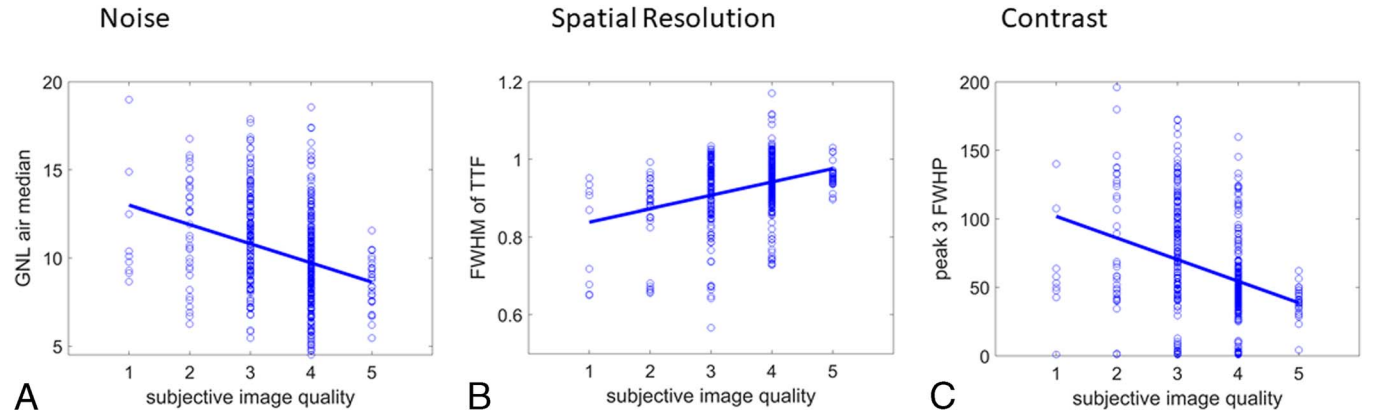


**FIGURE 3.** Correlations and significant linear regressions between automatically calculated parameters and subjective scores of image quality in a patient dataset of CT liver scans in the portal venous phase (n = 510). The best correlating parameter was selected for all 3 categories (see Supplemental Digital Content Figure 2, http://links.lww.com/RLI/A793, for all parameters). All scans were performed on a third-generation DSCT scanner. A linear regression is shown for significant correlations only (adjusted $R^2$ values are given in Table 1).

and 2, http://links.lww.com/RLI/A792 and http://links.lww.com/RLI/A793, for all parameters).

Significant correlations were found for 15 of 19 IQ metrics in both phantom datasets (4/5 noise with $R^2$ range = 0.55–0.70, 1/1 spatial resolution with $R^2$ = 0.21 and 0.26, and 10/13 contrast with $R^2$ range = 0.11–0.73) and for 21 of 24 IQ metrics in the patient dataset (4/5 noise with $R^2$ range = 0.019–0.096, 1/1 spatial resolution with $R^2$ = 0.11, and 16/18 contrast with $R^2$ range = 0.008–0.116).

Correlations observed in the phantom dataset were stronger than those in the patient dataset by almost a magnitude of 10. Strong correlations ($R^2 \geq 0.65$) in the phantom dataset are, in descending order, contrast metric peak 2 prominence ($R^2$ = 0.73), noise metrics Kortesniemi IQ score ($R^2$ = 0.70) and GNL air median ($R^2$ = 0.70), contrast metric count peak 2 ($R^2$ = 0.70), and noise metrics GNL median ($R^2$ = 0.67) and GNL mode ($R^2$ = 0.66) for the CT1 subgroup and noise metric GNL median ($R^2$ = 0.66), contrast metric peak 2 FWHP ($R^2$ = 0.66), and noise metrics GNL mode ($R^2$ = 0.65) and GNL air median ($R^2$ = 0.65) for the CT2 subgroup. The strongest correlations in the patient dataset are, in descending order, contrast metrics peak 3 FWHP ($R^2$ = 0.116) and CNR ($R^2$ = 0.115), spatial resolution metric FWHM TTF ($R^2$ = 0.110), and noise metric GNL air median ($R^2$ = 0.096).

The 2 phantom datasets showed some differences in strength and significance of correlations, but overall, the slope of 14 of the 19 regression lines was similar (Supplemental Digital Content Figure 1, http://links.lww.com/RLI/A792). For contrast metrics count peak 1, peak 2 location, peak 1 prominence, and minimum FWHP, correlations were found to be significant in only 1 of the phantom datasets, and for peak location 1, regression lines showed opposite slopes (Supplemental Digital Content Figure 1, http://links.lww.com/RLI/A792).

### Agreement: "Insufficient" Versus "Sufficient" IQ Scans

In the phantom dataset, 18 (CT1) and 10 (CT2) scans were scored as "insufficient" and 39 (CT1) and 47 (CT2) scans were scored as "sufficient." In the patient dataset, 43 scans were scored as "insufficient" and 465 scans as "sufficient". Of the 19 IQ metrics in the phantom datasets, 14 (CT1) and 15 (CT2) significantly differed between "sufficient" and "insufficient" categories (4/5 noise, 1/1 spatial resolution and 9 [CT1] and 10 [CT2]/13 contrast parameters), whereas 14 of the 24 IQ metrics in the patient dataset significantly differed between "sufficient" and "insufficient" categories (3/5 noise, 1/1 spatial resolution and 10/18 contrast parameters). Fourteen (CT1) and 15 (CT2) parameters in the phantom datasets, and 12 parameters in the patient dataset, showed significant correlations with subjective IQ scores as well as significant differences between "sufficient" and "insufficient" categories (Table 1).

## DISCUSSION

The developed computer algorithm was able to successfully calculate all IQ metrics for noise, resolution, and contrast on both phantom and clinical CT images, indicating the feasibility of IQ characterization of all 3 metric groups (noise, resolution, and contrast). The IQ metrics from each of these groups showed significant correlations with subjective radiologist IQ scores, demonstrating the potential for automatic IQ monitoring. Of the new contrast metrics developed in this study, 10 of 13 (phantom datasets) and 16 of 18 (patient dataset) correlated significantly with subjective radiologist IQ scores, indicating validity for contrast evaluation. The secondary analysis showed that in the phantom datasets, 14 and 15 of the 19 IQ metrics significantly differed between the 2 subjective "insufficient" and "sufficient" categories; in the patient dataset, significant differences between "insufficient" and "sufficient" categories were found for 14 of 24 IQ metrics. These metrics were distributed among all 3 aspects of IQ (noise, resolution, and contrast), indicating the potential for automatically determining whether CT scan IQ is diagnostically adequate.

Two phantom datasets were included to enable the acquisition of scans using a wide range of tube current and tube voltage values, to yield a wide range of IQs. The inclusion of a different scanner and a different radiologist demonstrates that the algorithm yields similar results and is robust even in a completely independent setting. Furthermore, the study design ensured that each radiologist evaluated images acquired on the scanner to which he/she was accustomed, to avoid the known bias in IQ assessment that is introduced when presenting images of a different/new scanner. In both phantom datasets, largely similar trends between automatic IQ metrics and subjective IQ were observed, indicating robustness. Only 5 contrast metrics showed differing trends, which may be due to differences in scanner performance, slight differences in protocols, and/or differences in reconstruction methods. Stronger correlations were seen between automatic IQ metrics and subjective scoring in phantom datasets as compared with the patient dataset, but this is to be expected due to lack of anatomical variation in the former. Although phantoms contain a reflection of the spectrum of human tissues, phantom CT images are limited as representations of the human anatomy, as complexity in the phantom is much reduced and anatomical variations and pathologies are lacking. Despite the lower coefficients seen in the patient dataset, correlations were significant, indicating the potential of this automatic IQ metric calculating algorithm.

The current study was unique in incorporating IQ metrics for noise, resolution, and contrast in the algorithm and in comparing results with subjective IQ scores of both phantom and clinical patient scans. Most previous studies developing IQ metrics focus on 1 aspect of total IQ.[17,19–22] For example, correlation between the Kortesniemi IQ score and visual grading was previously shown in thorax CT scans of cadavers.[18] Only Cheng et al[23] used 3 metrics: liver HU value, noise magnitude, and clarity (which is a combination of spatial resolution, noise texture and lesion contrast). The clinical images were subjectively evaluated per IQ aspect, and strong rank-order agreements between algorithm and subjective IQ assessment were found. The current study, however, relates the automated IQ metrics to an overall subjective IQ score and includes scans that are simulated in such a way that they are of insufficient IQ. We consider it more relevant to clinical practice to use a combination of the 3 aspects of noise, spatial resolution, and contrast and to be able to distinguish between scans of sufficient and insufficient IQ.

The current study was performed using transversal CT scans of the abdomen, in a patient dataset specifically focusing on the portal venous phase, but results are generalizable to other scan areas or reconstruction planes. Even though other anatomical areas have distinctly different appearances, the Kortesniemi IQ score, GNL, and FWHM of the TTF are not anatomy dependent and can be calculated for other body regions without adjustments. For the spatial resolution metric, however, care must be taken to avoid measuring at the edges between different body parts, such as the thorax-arm interface or the interface between the legs, as this metric quantifies sharp changes in HU values from tissue to air. The contrast metrics will also need some attention as they are based on histogram analyses of HU values and both location and number of peaks are specific to anatomical area and contrast injection protocol. However, once the histogram is known, it is straightforward to adapt contrast IQ metric calculations following the method presented in this study.

The subjective IQ was quantified using a Likert scoring system performed by a single radiologist per phantom dataset and by 2 radiologists in consensus for the patient dataset. This limited number of scoring radiologists prohibits interrater characteristic evaluation. Furthermore, subjective IQ scoring is strongly related to the diagnostic question at hand, which limits generalizability. Whether similar correlations exist when looking at images of other anatomies must therefore be investigated. Limitations of a statistical nature include low prevalence of the number of "insufficient IQ" scans in both datasets, which may have induced bias. Furthermore, several of the metrics described in this work are likely to be collinear. Nevertheless, these results are encouraging

and future studies can be undertaken to implement the algorithm for different anatomies and slice reconstructions, as well as to evaluate IQ metric performance when using CT scanner settings that might be expected to interfere. The algorithm can be further developed into a continuous IQ monitoring tool that may prove invaluable in radiation dose and contrast injection protocol optimization[29,30] or in retake scan decision making in daily clinical practice.[31]



## CONCLUSION

A computer algorithm was developed that successfully calculated IQ metrics quantifying noise, spatial resolution, and contrast from phantom and clinical abdominal CT scans. Correlations between subjective IQ and multiple IQ metrics of noise, spatial resolution, and contrast were significant. Furthermore, many of these metrics appear to differentiate between images that radiologists score as "sufficient" and "insufficient." These results can be leveraged to develop an algorithm to automatically evaluate objective IQ of any scan, regardless of vendor, patient characteristics, radiation, or contrast media dose. Such a tool could not only prove invaluable in optimization and quality control in clinical practice, but may also provide much sought-after objective outcomes missing in IQ research to date.

## REFERENCES


1. Brenner DJ, Hall EJ. Computed tomography—an increasing source of radiation exposure. *N Engl J Med*. 2007;357:2277–2284.
2. Smith-Bindman R, Miglioretti DL, Johnson E, et al. Use of diagnostic imaging studies and associated radiation exposure for patients enrolled in large integrated health care systems, 1996–2010. *JAMA*. 2012;307:2400–2409.
3. Lell MM, Kachelrieß M. Developments in computed tomography high speed, low dose, deep learning, multienergy. *Invest Radiol*. 2020;55:8–19.
4. Jeukens CRLPN, Boere H, Wagemans BAJM, et al. Probability of receiving a high cumulative radiation dose and primary clinical indication of CT examinations: a 5-year observational cohort study. *BMJ Open*. 2021;11:e041883.
5. Rehani MM, Hauptmann M. Estimates of the number of patients with high cumulative doses through recurrent CT exams in 35 OECD countries. *Phys Med*. 2020;76:173–176.
6. Lumbreras B, Salinas JM, Gonzalez-Alvarez I. Cumulative exposure to ionising radiation from diagnostic imaging tests: a 12-year follow-up population-based analysis in Spain. *BMJ Open*. 2019;9:e030905.
7. Euratom. Council Directive 2013/59/EURATOM. 2014. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:02013L0059-20140117&from=EN. Accessed July 13, 2022.
8. ICRP. Managing patient dose in computed tomography. A report of the International Commission on Radiological Protection. *Ann ICRP*. 2000;30:7–45.
9. Song JS, Lee JM, Sohn JY, et al. Hybrid iterative reconstruction technique for liver CT scans for image noise reduction and image quality improvement: evaluation of the optimal iterative reconstruction strengths. *Radiol Med*. 2015;120:259–267.
10. AAPM. *Performance Evaluation of Computed Tomography Systems*. Alexandria, VA: American Association of Physicists in Medicine; 2019. Available at: https://aapm.org/pubs/reports/detail.asp?docid=186. Accessed July 13, 2022.
11. Verdun FR, Racine D, Ott JG, et al. Image quality in CT: from physical measurements to model observers. *Phys Med*. 2015;31:823–843.
12. Mileto A, Guimaraes LS, McCollough CH, et al. State of the art in abdominal CT: the limits of iterative reconstruction algorithms. *Radiology*. 2019;293:491–503.
13. Noferini L, Taddeucci A, Bartolini M, et al. CT image quality assessment by a channelized Hotelling observer (CHO): application to protocol optimization. *Phys Med*. 2016;32:1717–1723.
14. Yu L, Chen B, Kofler JM, et al. Correlation between a 2D channelized Hotelling observer and human observers in a low-contrast detection task with multislice reading in CT. *Med Phys*. 2017;44:3990–3999.
15. Zhou W, Michalak GJ, Weaver JM, et al. A universal protocol for abdominal CT examinations performed on a photon-counting detector CT system: a feasibility study. *Invest Radiol*. 2020;55:226–232.
16. Racine D, Mergen V, Viry A, et al. Photon-counting detector CT with quantum iterative reconstruction, impact on liver lesion detection and radiation dose reduction [published online ahead of print September 12, 2022]. *Invest Radiol*. doi:10.1097/RLI.0000000000000925.
17. Kortesniemi M, Schenkel Y, Salli E. Automatic image quality quantification and mapping with an edge-preserving mask-filtering algorithm. *Acta Radiol*. 2008;49:45–55.
18. Franck C, De Crop A, De Roo B, et al. Evaluation of automatic image quality assessment in chest CT—a human cadaver study. *Phys Med*. 2017;36:32–37.
19. Christianson O, Winslow J, Frush DP, et al. Automated technique to measure noise in clinical CT examinations. *AJR Am J Roentgenol*. 2015;205:W93–W99.
20. Malkus A, Szczykutowicz TP. A method to extract image noise level from patient images in CT. *Med Phys*. 2017;44:2173–2184.
21. 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;43:5330–5338.
22. Ott JG, Becce F, Monnin P, et al. Update on the non-prewhitening model observer in computed tomography for the assessment of the adaptive statistical and model-based iterative reconstruction algorithms. *Phys Med Biol*. 2014;59:4047–4064.
23. Cheng Y, Abadi E, Smith TB, et al. Validation of algorithmic CT image quality metrics with preferences of radiologists. *Med Phys*. 2019;46:4837–4846.
24. MathWorks. Savitzky-Golay filtering. 2020. Available at: https://nl.mathworks.com/help/signal/ref/sgolayfilt.html. Accessed September 3, 2021.
25. MathWorks. Prominence. 2020. Available at: https://nl.mathworks.com/help/signal/ug/prominence.html. Accessed September 3, 2021.
26. Timischl F. The contrast-to-noise ratio for image quality evaluation in scanning electron microscopy. *Scanning*. 2015;37:54–62.
27. Martens B, Bosschee JGA, Van Kuijk SMJ, et al. Finding the optimal tube current and iterative reconstruction strength in liver imaging; two needles in one haystack. *PLoS One*. 2022;17:e0266194.
28. Ellmann S, Kammerer F, Brand M, et al. A novel pairwise comparison-based method to determine radiation dose reduction potentials of iterative reconstruction algorithms, exemplified through circle of Willis computed tomography angiography. *Invest Radiol*. 2016;51:331–339.
29. Martens B, Gregor J, Mihl C, et al. Individualized scan protocols in abdominal computed tomography, radiation versus contrast media dose optimization. *Invest Radiol*. 2022;57:353–358.
30. Tilman E, O'Doherty J, Schoepf UJ, et al. Reduced iodinated contrast media administration in coronary CT angiography on a clinical photon-counting detector CT system: a phantom study using a dynamic circulation model [published online ahead of print September 13, 2022]. *Invest Radiol*. doi:10.1097/RLI.0000000000000911.
31. Rose S, Viggiano B, Bour R, et al. Applying a new CT quality metric in radiology: how CT pulmonary angiography repeat rates compare across institutions. *J Am Coll Radiol*. 2021;18:962–968.