Evaluation of Low-Contrast Detectability of Iterative Reconstruction Across Multiple Institutions, Manufacturers and Exposure Levels

Ganesh Saiprasad, James Filliben, Adele Peskin, Eliot Siegel, Joseph Chen, Christopher Trimble, Zhitong Yang, Olav Christianson, Ehsan Samei, Elizabeth Krupinski and Alden Dima

1NIST, 2Univ. of Maryland Medical Center, 3Duke University, 4Univ. of Arizona, 5Univ. Maryland Baltimore

Manuscript Type: Original Research

Advances in Knowledge:

- Iterative Reconstruction (IR) yielded superior low-contrast detectability compared to Filtered Back Projection (FBP) across all three vendors and three exposure levels
- IR has the potential ability to reduce exposure by up to 60 % (12 mGy vs 7.2 mGy) without substantial loss of low-contrast detectability compared to FBP
- There is substantial variability among the major vendors not only in baseline image quality using FBP but also in the relative benefit of the use of IR.

Implication for Patient Care:

- CT radiation exposure to patients can be potentially reduced by using Iterative Reconstruction (IR) compared to Filtered Back Projection (FBP) without compromising low-contrast detectability

1 This contribution of NIST, an agency of the U.S. Government, is not subject to copyright.
Abstract

**Purpose:** To compare image resolution using Iterative Reconstruction (IR) with the resolution using Filtered Back Projection (FBP), for low-contrast objects in phantom images across vendors and exposure levels.

**Materials and Methods:** Randomized repeat scans of the Gammex 464 American College of Radiology (ACR) CT Accreditation Phantom (Module 2, Low-contrast) were performed for multiple radiation exposures, vendors and vendor IR algorithms. Eleven volunteers were presented with a total of 900 images using a custom-designed Graphical User Interface (GUI) to perform a task created specifically for this reader study. Results were analyzed using statistical graphics and analysis of variance.

**Results:** We found that across three vendors (blinded as A, B, and C) and across three exposure levels, the mean Correct Classification Rate (CCR) was higher for IR than FBP (p<0.01): 87.4 % IR and 81.3 % FBP at 20 mGy, 70.3% IR and 63.9 % FBP at 12 mGy, and 61.0 % IR and 56.4 % FBP at 7.2 mGy. There was a significant difference in mean CCR between vendor B and the other two vendors. Across all exposure levels, images obtained using vendor B’s scanner outperformed the other vendors, with a CCR of 74.4 % while the CCR for vendors A and C were 68.1 % and 68.3 % respectively. Across all the readers, the mean CCR for IR (73.0 %) was higher compared to the mean CCR for FBP (67.0 %).

**Conclusion:** The potential exists to reduce radiation dose without compromising low-contrast detectability by using IR compared to FBP. There is also a substantial variability across vendor reconstruction algorithms.

INTRODUCTION:
Computed Tomography (CT) is one of the most highly utilized medical imaging modalities. 70 million CT examinations are performed annually in the United States(1). There is great concern about the risks related to radiation-induced cancer. A recent study found an increase, from 0.4 % (1996) to 1.5-2.0 % (current), in the estimated cancer risk from CT radiation of all cancers in the United States(2). Pediatric CT radiation has been associated with the development of malignancies in children such as leukemia and brain cancer(3). Therefore, it is imperative for CT radiation exposure to be as low as reasonably achievable (the ALARA principle(4)).

The use of Iterative Reconstruction (IR) has been heavily promoted by vendors to reduce radiation exposure in CT imaging acquisition. IR is an iterative algorithm used to reconstruct 2D
and 3D images from projections of an object. The main advantages of IR over Filtered Back Projection (FBP) are the ability to incorporate attenuation corrections and reduction in image noise(5).

Singh et al. (6) reconstructed CT data (chest) from 23 patients, using 30 %, 50 %, and 70 % Adaptive Statistical Iterative reconstruction (ASIR) - FBP blending. When they reduced exposure levels to 3.5 mGy, they achieved acceptable image noise and diagnostic confidence in 70 % of studies using ASIR, but only 30 % using FBP. Leipsic et al. (7) reconstructed CT data from 62 patients using FBP, 20 – 80 % ASIR-FBP blending, and 100% ASIR. Image quality evaluation by two radiologists concluded that 40 % - 60 % ASIR – FBP blending improved the overall image quality significantly and reduced noise. Hara et al. (8) scanned 12 patients at routine-dose and low-dose CT, with and without ASIR. They also scanned the American College of Radiology (ACR) CT accreditation phantom, Gammex 464 (Gammex, Inc, Middleton, WI). Their work suggests that 32 – 65 % CT dose index (CTDI) reductions may be achieved by using ASIR.

Previous research has suggested that significant dose reduction can be achieved with the use of IR without compromising the image quality. However, to our knowledge, the degree of variability among different vendors in their implementation of IR and the relative effectiveness of IR in minimizing the impact of low exposure has not been documented. Our quantitative approach was designed to provide initial data to investigate these questions by attempting to evaluate the impact of IR on low-contrast detectability utilizing three different CT exposure levels, and acquiring the images using three different CT vendors at three different institutions. Eleven readers were asked to analyze 900 randomly displayed images for a total of 9,900 individual responses. These responses were collected and analyzed to help estimate the impact of IR on low-contrast detectability when compared to FBP.

**MATERIALS AND METHODS:**
IRB approval was not required because the study did not involve human subjects.

**Phantom:**
The ACR CT accreditation phantom (Gammex 464) is designed for routine quality assurance (QA) testing of CT scanners (9). Image quality is evaluated using multiple modules to test specific parameters designed into the phantom. Module 2 is used to evaluate low-contrast resolution, shown in Figure 1(10). The 25 mm cylinder at the 12 noon position is used to assess contrast-to-noise ratio (CNR). Cylinders inside the phantom range from 2 mm to 6 mm in diameter. The cylinders are arranged in clusters which consist of four individually aligned cylinders. There is a 0.6 % (6 HU) difference in density between the cylinders and the
background material, which measures approximately 90 HU. All CT scanners data is output in Hounsfield units (HU), a quantitative but non-SI unit of measure for radiodensity.

**Scanning Protocol:**
The ACR phantom was scanned using three scanners: GE Discovery CT750 HD (GE Healthcare, Waukesha, WI), Siemens SOMATOM Definition Flash (Siemens Healthcare, Malvern, PA) and Philips Brilliance iCT 256 (Philips Healthcare, Andover, MA). Excluding the Computed Tomography Dose Index (CTDI) value, all the other scan parameters were kept constant and as similar to each other as technically possible (see Table 1). Three exposure levels: 7.2 mGy, 12 mGy, and 20 mGy, were selected to approximate the lower end of the range, typically used for abdominal CT. We chose to vary the CTDI rather than varying the X-ray tube current because CTDI provides scanner output radiation, enabling accurate comparisons of radiation output from different scan protocols or scanners(11). CTDI is used to quantify, consistently and reproducibly, the radiation output of a CT scanner. Since our study involved multiple institutions and multiple scanners, CTDI was the most effective way to achieve consistency with respect to robustly measuring radiation output. Each scan was reconstructed using the vendor-specific FBP and IR algorithm commercially available at the time of the study. In order to maintain anonymity of the vendors, no association of a particular vendor to a particular set of measurements is provided.

**Experimental Design:**
A 4-factor full factorial design was employed using 3 exposure levels (7.2, 12, and 20 mGy), 3 vendors (A, B, and C), 2 algorithms (FBP and IR) and 11 readers. The ACR phantom Module 2 was scanned and the resulting sinogram was reconstructed using FBP to produce 10 center image slices. Each scan was replicated 5 times for a total of 50 images. The sinogram data were post-processed using each vendor’s IR algorithm to yield 50 additional images. This was done for each of the 3 exposures and for each of the 3 vendors, producing 900 post-processed images, and thus 9,900 image quality evaluations were obtained from the 11 readers.

Five replicated scans were used to provide a reasonable estimate of within-process variability and provided vital information about natural run-to-run variation, as well as information about potential machine anomalies (e.g., device drift, dead spots, and other position-related issues). Although we designed a fully balanced experiment (5 repeat scans at each exposure level), due to a minor acquisition complication for vendor A, we inadvertently had 4 repeat scans for the 7.2 mGy level and 6 repeat scans for the 12 mGy for vendor A. This resulted in a slightly unbalanced experimental design, which we believe had no significant effect on the conclusions from our study.
Reader Study Design:
Workstations for the study used optimal lighting routinely used for diagnostic interpretation. A Two-Alternative-Forced-Choice (2AFC) reader study design (12,13) was used. As part of the forced choice study design, we randomly flipped half of the images along the vertical axis. For each pair of displayed images, readers identified whether the 5 and 6 mm cylinders appeared on the left (not-flipped) or right (flipped) side of the phantom, and their choices were recorded as correct or incorrect. Readers worked at their own speed and response time was recorded. Figure 2 demonstrates the pattern on the left and the flipped version of the same image on the right.

In a preliminary assessment, cylinders with 2, 3 and 4 mm diameter cylinders were very hard to detect at lower exposures. Therefore, the 5 and 6 mm diameter cylinders were chosen for our study. They were relatively easy to identify at higher exposure levels, and challenging at lower exposure levels. The highest exposure level was set to 20 mGy, used clinically for studies such as abdominal CT. The two lower exposure levels were 12 mGy and 7.2 mGy, 60 % and 36 % of the highest level, used clinically for abdominal and pelvic CT exams.

Graphical User Interface (GUI):
To perform the reader study, a Graphical User Interface (GUI) was designed using Matlab (Mathworks, Natick, MA) (14), and is shown in Figure 3. Default display settings were set to: Window Width = 100 and Window Level = 100, recommended for evaluating the low-contrast module of the ACR Phantom. The GUI had various functionalities including changing a previous selection and pausing the study.

Statistical Analysis:
For our analysis, we employed a series of Exploratory Data Analysis (EDA) (15) statistical graphics (16) to ascertain the overall effect of each factor and/or combinations of factors. Statistical graphics were augmented with classical analysis of variance (ANOVA) statistics. We used NIST (National Institutes of Standards and Technology) Dataplot (17) software to generate the graphical results.

RESULTS:
Figure 4, a main effects plot (18), shows the relative importance of the three factors of interest: Exposure, Vendor and Algorithm. The vertical axis is the mean Correct Classification Rate (CCR). The horizontal axis displays the factors with their corresponding levels. The plot points represent the CCRs for a given level of a given factor. If a factor has no effect, the mean response has little or no change across the different levels of that factor. The size of the effect
for a factor is seen relative to the size of the change in the mean values. Conversely, if a factor does have an effect, then the mean response for at least one level within the factor will differ significantly from the remaining levels; the larger the difference, the more significant the factor.

The vertical bars on each point represent 95% confidence bounds. Towards the bottom of the plot is a list of relevant 1-factor model ANOVA statistics: 1) the “effect” for each factor (difference between the largest and smallest means); 2) the “relative effect” for each factor (the percentage ratio of the effect and the global mean accuracy, here = 70.24, for the response), and 3) the p-value from the series of 1-way model ANOVAs (all of which were statistically significant at the 1% level). In terms of relative importance, it is evident from the plot (and the statistics) that the most important factor is exposure (mean change of 26 units, or 36%), followed by the remaining two factors (changes <= 7 units, or 10%). A full 3-factor model ANOVA yields the same ranking with all p values < 0.01.

**Exposure Effect:**
The three exposure levels were 7.2 mGy, 12 mGy, and 20 mGy. Figure 5 shows that the CCRs increased with increasing exposure. The mean CCRs were 59%, 67%, and 84%, respectively. The effect of exposure was significant (p < 0.01) not only globally across all the three vendors, but also separately within each of the 3 vendors, and separately within each of the 3 vendors x 2 algorithms. Regardless of vendor and algorithm, 20 mGy was an improvement over 12 mGy, which in turn was an improvement over 7.2 mGy.

**Vendor Effect:**
Figure 6 shows that vendor was a significant (p < 0.01) predictive factor for correct classification, but was less significant than exposure. The mean CCRs for the three vendors, A, B, and C, were (rounded): 68% (A), 74% (B) and 68% (C). Vendor B was significantly different from Vendors A and C, but the two vendors A and C themselves were not significantly different from each other. Figure 6 shows that the effect of the vendor depends on the choice of algorithm (and vice versa). For the FBP algorithm, vendor B is superior to vendors A and C, but is statistically significant in only one of three cases (exposure = 20); whereas for the IR algorithm, B is statistically superior (p < 0.01) to A and C for all 3 exposures.

**Algorithm Effect:**
The two reconstruction algorithm classes being evaluated were Iterative Reconstruction (IR) and Filtered Back Projection (FBP). Figure 7 shows that IR and FBP are significantly different, with IR having the higher CCR (73%) compared with FBP (67%). The algorithm effect was examined for each of the 9 (3 vendor x 3 dose) combinations. Although the mean CCR for IR exceeded the mean CCR for FBP for all 9 cases, the difference was not significant for some
combinations, especially those with lower doses. For the highest dose (20 mGy), IR was significantly higher than FBP for each of the 3 vendors (vendor A was significant at the 2 % level and B and C at the 1 % level). The IR-FBP differences were greatest for vendor B for which IR statistically exceeded FBP, universally and regardless of the dose (highlighted in Figure 7).

**Reader Effect:**
There were a total of 11 readers who participated in the study (Figure 8). Across all readers, the mean CCR for IR (73 %) was higher than the mean CCR for FBP (67 %). For 7 out of the 11 readers, the mean CCR for IR was higher than for FBP with significance at the 5% level. The response/classification time is the time each reader took, to look at an image and make a decision. This is a measure of the difficulty of the task, as longer times correspond to more difficult set of images to classify (19). Figure 9 shows the plot of mean classification time vs. exposure for the two algorithms. With the exception of near-equality for the 7.2 mGy images, the readers took longer to read the FBP images than the IR images.

**DISCUSSION:**
Previous studies have evaluated the effect of IR on perceived image quality and compared performance of IR vs. FBP for a single vendor. They have either quantitatively measured image noise in a phantom or asked radiologists to make qualitative judgments about image quality. For example, one recent study suggested a dose reduction of 46.4 % for chest CT and 38.2 % for abdominal CT based on analysis of 234 patients(20), but the reconstructed images were not compared at the same radiation dose levels.

The objective of our study was to provide a design that combined the quantitative phantom measurements with experienced human observer performance, focusing on low-contrast detectability, to provide a quantitative measure that takes the complexities and idiosyncratic nature of human perception into account. We also evaluated the performance of three vendors with the highest CT market shares to test our hypothesis that in addition to baseline differences in low-contrast detection using FBP, there would be differences in the relative improvement provided by specific algorithms and implementations. Our results suggest that the task was sufficiently difficult for experienced readers to allow good discrimination among different vendors, algorithms and exposure values.

IR yielded superior low-contrast detectability at all exposure-vendor combinations. A closer look at Figure 7, with the mean value blocks across all exposure-vendor combinations (shown in grey), revealed an interesting observation. **Low-contrast detectability at 7.2 mGy for an IR image was indistinguishable from FBP at an exposure of 12 mGy across all vendors,**
suggesting that the same image could be obtained at 60% reduced exposure using IR without compromising low-contrast detectability.

The variability among vendors in FBP performance was expected and the relative amount of improvement with IR was particularly interesting. Vendor B had the best FBP performance and also demonstrated the greatest improvement with the use of IR. This resulted in a substantial superiority for vendor B in low-contrast detectability with IR. It is not certain whether this was due to superiority of the IR algorithm itself, or whether it was related to a higher quality sinogram which may have maximized the benefit of IR for this task.

There are several limitations of our study with regard to the clinical implications in the determination of optimal exposure using IR in comparison to FBP. For our specific task, the human observers seemed to perform as expected, and they demonstrated improved classification using IR. However, this technique of determining low-contrast detectability by flipping an image on its vertical axis has not otherwise been validated, since it differs substantially from subjective evaluation of an image by a radiologist or from the objective measurement by a physicist. Our hypothesis, that this technique better represents the impact of the human visual system, needs further testing.

The distance and angle from the observer to the monitor, and precise monitor calibration and matching was not performed and could have had an impact on our results. We utilized the commercially available IR algorithms and the implementation of these, as suggested by the vendors for abdominal imaging. However, most of the vendors have subsequently refined/improved their IR algorithms, and consequently the variability among vendors may have changed.

Only one CT scanner was utilized from each vendor for the acquisition of images. Variability among CT scanners likely exists due to the age of the scanner, preventative maintenance, variability in detector performance and noise and other factors. Ideally data should be acquired on more than one machine to determine this variability. Although our approach to combine human perception and evaluation of the low-contrast object task was successful in measuring low-contrast performance, it is likely that low-contrast performance is only one of multiple important variables in overall image quality. Other factors, such as the ability to accurately reproduce image texture, artifact reduction, and geometric distortion likely play a major role in overall image quality and variability in quality among vendors and among different IR algorithms.
SUMMARY STATEMENT:
Low-contrast detectability at 7.2 mGy for an IR image was indistinguishable from FBP at an exposure of 12 mGy across all vendors, suggesting that the same image could be obtained at 60% reduced exposure using IR without compromising low-contrast detectability.

DISCLAIMER:
Certain commercial equipment, instruments, materials or software are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose. The results of this study on low-contrast detectability lend support to the notion that routine IR usage can play a role in reducing patient radiation dose.

References:


<table>
<thead>
<tr>
<th>Scan Parameter</th>
<th>GE</th>
<th>Siemens</th>
<th>Philips</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tube Voltage (kVp)</td>
<td>120</td>
<td>120</td>
<td>120</td>
</tr>
<tr>
<td>Slice Thickness (mm)</td>
<td>1.25</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Acquisition Mode</td>
<td>Helical</td>
<td>Helical</td>
<td>Helical</td>
</tr>
<tr>
<td>Display Field of View (mm)</td>
<td>250</td>
<td>250</td>
<td>250</td>
</tr>
<tr>
<td>Pitch (mm)</td>
<td>0.984</td>
<td>1</td>
<td>0.93</td>
</tr>
<tr>
<td>Rotation Time (s)</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>Collimation Width (mm)</td>
<td>0.625</td>
<td>0.6</td>
<td>0.625</td>
</tr>
<tr>
<td>Reconstruction Kernel</td>
<td>Standard/ASIR&lt;sub&gt;50&lt;/sub&gt;</td>
<td>B30f/IRIS</td>
<td>B/iDose 4</td>
</tr>
</tbody>
</table>
Figure 1: ACR Phantom Module 2
Displaying the Pattern of Cylinders
Figure 2: Not Flipped (5 and 6 mm cylinders on the left) and Flipped (5 and 6 mm cylinders on the right) Images
Figure 3: Graphical User Interface (GUI) for the Reader Study
Figure 4: Main Effects Plot
Figure 5: Exposure Effect over Vendor and Algorithm
Figure 6: Vendor Effect over Exposure and Algorithm
Figure 8: Algorithm Effect over Sorted Reader
Figure 9: Algorithm Effect over Vendor and Exposure (Time)