Check for updates

Received: 23 August 2023 | Revised: 13 November 2023 | Accepted: 4 December 2023

DOI: 10.1002/mp.16901

RESEARCH ARTICLE

MEDICAL PHYSICS

Pediatric evaluations for deep learning CT denoising

Brandon J. Nelson[1] | **Prabhat Kc**[1] | **Andreu Badal**[1] | **Lu Jiang**[2] | **Shane C. Masters**[3] | **Rongping Zeng**[1]

[1]Center for Devices and Radiological Health, Office of Science and Engineering Labs, Division of Imaging, Diagnostics, and Software Reliability, U.S. Food and Drug Administration, Silver Spring, Maryland, USA

[2]Center for Devices and Radiological Health, Office of Product Evaluation and Quality, Office of Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland, USA

[3]Center for Drug Evaluation and Research, Office of Specialty Medicine, Division of Imaging and Radiation Medicine, U.S. Food and Drug Administration, Silver Spring, Maryland, USA

Correspondence
Brandon Nelson, 10903 New Hampshire Avenue, Silver Spring, MD 20993.
Email: brandon.nelson@fda.hhs.gov

Funding information
FDA National Center for Toxicological Research

Abstract

Background: Deep learning (DL) CT denoising models have the potential to improve image quality for lower radiation dose exams. These models are generally trained with large quantities of adult patient image data. However, CT, and increasingly DL denoising methods, are used in both adult and pediatric populations. Pediatric body habitus and size can differ significantly from adults and vary dramatically from newborns to adolescents. Ensuring that pediatric subgroups of different body sizes are not disadvantaged by DL methods requires evaluations capable of assessing performance in each subgroup.

Purpose: To assess DL CT denoising in pediatric and adult-sized patients, we built a framework of computer simulated image quality (IQ) control phantoms and evaluation methodology.

Methods: The computer simulated IQ phantoms in the framework featured pediatric-sized versions of standard CatPhan 600 and MITA-LCD phantoms with a range of diameters matching the mean effective diameters of pediatric patients ranging from newborns to 18 years old. These phantoms were used in simulating CT images that were then inputs for a DL denoiser to evaluate performance in different sized patients. Adult CT test images were simulated using standard-sized phantoms scanned with adult scan protocols. Pediatric CT test images were simulated with pediatric-sized phantoms and adjusted pediatric protocols. The framework's evaluation methodology consisted of denoising both adult and pediatric test images then assessing changes in image quality, including noise, image sharpness, CT number accuracy, and low contrast detectability. To demonstrate the use of the framework, a REDCNN denoising model trained on adult patient images was evaluated. To validate that the DL model performance measured with the proposed pediatric IQ phantoms was representative of performance in more realistic patient anatomy, anthropomorphic pediatric XCAT phantoms of the same age range were also used to compare noise reduction performance.

Results: Using the proposed pediatric-sized IQ phantom framework, size differences between adult and pediatric-sized phantoms were observed to substantially influence the adult trained DL denoising model's performance. When applied to adult images, the DL model achieved a 60% reduction in noise standard deviation without substantial loss in sharpness in mid or high spatial frequencies. However, in smaller phantoms the denoising performance dropped due to different image noise textures resulting from the smaller field of view (FOV) between adult and pediatric protocols. In the validation study, noise reduction trends in the pediatric-sized IQ phantoms were found to be consistent with those found in anthropomorphic phantoms.

Conclusion: We developed a framework of using pediatric-sized IQ phantoms for pediatric subgroup evaluation of DL denoising models. Using the framework, we found the performance of an adult trained DL denoiser did not generalize

 © 2023 American Association of Physicists in Medicine.

well in the smaller diameter phantoms corresponding to younger pediatric patient sizes. Our work suggests noise texture differences from FOV changes between adult and pediatric protocols can contribute to poor generalizability in DL denoising and that the proposed framework is an effective means to identify these performance disparities for a given model.

KEYWORDS
computed tomography, ct, deep learning, denoising, evaluations, image quality, medical imaging, pediatric imaging, phantoms

1 | INTRODUCTION

The use of x-ray computed tomography (CT) in pediatric patients presents benefits of improved diagnosis and treatment as well as risks associated with radiation exposure. This radiation exposure is particularly concerning for pediatric patients that are more radiosensitive per unit dose than adults and have a longer expected lifetime to accumulate associated cancer risks.[1] Therefore, appropriate pediatric CT radiation dose reduction techniques are important to minimize risks from radiation while maintaining the diagnostic utility of exams.

Image reconstruction can contribute to CT dose reduction by using computer algorithms to better separate useful information from noise in low dose CT exams. A new paradigm in CT image reconstruction has been the introduction of deep learning (DL)-based techniques. DL can be used to generate CT images directly from low dose x-ray projections. Alternatively, DL can be used to remove noise from low dose CT images reconstructed with conventional methods such as filtered backprojection (FBP).[2] Hybrid approaches have also been proposed that utilize DL in iterative reconstruction routines.[3]

Among the wide variety of DL implementations for image reconstruction, image-based DL denoising in CT has been most intensively studied and is available in commercial CT scanners.[4] These image-based DL denoisers are typically trained on many examples of paired low and high noise patient CT images. The performance of DL denoisers has been shown to be advantageous to other advanced reconstruction methods like model based iterative reconstruction (MBIR) due to faster computation times, preferable noise texture, along with comparable noise reduction and sharpness preservation.[5]

When using image reconstruction to reduce dose, performance assessments are essential to ensure that diagnostic quality is not impaired. DL denoisers are nonlinear meaning that their performance cannot be fully characterized by simple measures of noise standard deviation and high contrast sharpness. Additionally, DL algorithms are data driven and thus their performance decays when encountering patient populations, scanners, clinical protocols, or other situations or image features outside their training data distribution. Therefore, assessing the generalizability of DL reconstruction and denoising models to different patient characteristics and populations has become a growing concern.

Generalizability assessments of DL reconstruction are especially important for pediatric patients since DL models are typically trained on primarily adult patient data. DL models often lack pediatric data because pediatric patient data is so scarce. Despite making up roughly 20% of the US population, a recent survey of trends in medical imaging found only 4% of medical imaging exams were performed on pediatric patients.[6] This is also reflected in the lack of pediatric data in public datasets commonly used for DL model development.[7] Altogether, this general lack of pediatric imaging data makes accumulating sufficient pediatric imaging data to assess DL generalizability particularly challenging.

One aspect where pediatric patients characteristically differ from adults is body habitus and size (e.g., waist diameter). Patient size is a key characteristic influencing image quality in CT as it factors into x-ray path length, impacting noise and image artifacts. Furthermore, the size range of pediatric patients supersedes that of adult patients, from less than 100 mm effective diameter in newborns to over 350 mm in adolescents.[8,9] This is illustrated in Figure 1 by plotting mean waist effective diameter with age using data from AAPM Task Group report 204 and CDC Vital and Health Statistics data.[10,11] The pediatric subgroups shown (newborn, infant, child, and adolescent) are based upon FDA recommended age ranges.[12]

Initial studies of using DL denoising models on pediatric image data have demonstrated DL denoising models to be effective in clinical settings, but only with small patient numbers and a limited range of body sizes.[5,13,14] Without subgroup analyses[15] as well as comparisons against adult patients, it is unclear whether disparities in DL denoising performance exist among pediatric subgroups.

This work introduces a set of pediatric-sized IQ phantoms along with a computational framework to evaluate the performance of DL-based denoising across pediatric subgroups. To demonstrate the use of this framework, a study was performed on an adult-trained

24734209, 2024, 2, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.16901 by Us Fda, Wiley Online Library on [19/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

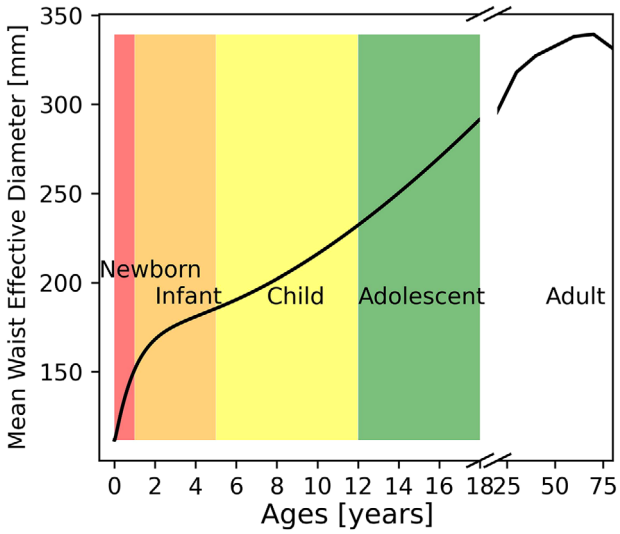

FIGURE 1 Mean waist effective diameters by age compiled from AAPM Task Group report 204 and the CDC Vital and Health Statistics data.[10,11] Shaded regions indicate age ranges of pediatric subgroups recommended by FDA for evaluating abdominal x-ray imaging devices.[9]

DL denoising model. The results of this study show how our developed framework can be used to complement existing scarce pediatric data in assessing the performance of a given DL denoiser in pediatric patients.

2 | METHODS

This investigation of pediatric generalizability of DL denoising models for CT starts by introducing the developed pediatric-sized image quality assessment framework, summarized in Figure 2. Newly developed pediatric-sized digital phantoms are used as inputs to existing CT simulation frameworks generating realistic noisy images. These images are then processed by the DL denoising model being evaluated. A series of objective and task-based image quality assessments then compare performance between pediatric subgroups and adults both before and after denoising.

A case study using the framework is shown to assess the performance of an adult trained REDCNN DL denoising model on different pediatric subgroups.

2.1 | Pediatric-sized image quality phantom assessment framework

The proposed pediatric-sized image quality (IQ) phantom assessment framework consists of two parts: 1) a set of newly developed pediatric-sized digital phantoms and 2) a set of image quality assessments. The digital phantoms are virtually imaged with a CT simulation framework to provide test images for a DL denoising model. These denoised test images are then assessed by the evaluation framework to identify performance disparities based on patient size.

2.1.1 | Pediatric-sized digital phantom design

To assess different aspects of image quality performance, three sets of IQ phantoms were simulated. These phantoms were modeled after standard CT image quality evaluation phantoms but with varying cross-sectional diameters. The diameters were simulated to match the mean effective diameters of pediatric subgroups ranging from newborns to adults. Figure 1 illustrates how these effective diameters relate to patient age and subgroup. A subset of these phantoms is shown in Figure 2 as the CTP404, Uniform Water, and MITA-LCD CCT189 phantoms.

The first computer modeled IQ phantom is a modified version of the sensitometry module CTP404 from the Catphan 600 phantom (The Phantom Laboratory, Salem, NY). This cylindrical phantom has eight unique contrast inserts ranging from -1000 to +900 HU in a uniform background of 0 HU. In its standard size, CTP404 has a diameter of 150 mm with 12 mm diameter inserts. Due to the sharp intersection between the phantom background and multi-contrast inserts, this module was used to evaluate contrast-dependent image sharpness using the contrast-dependent modulation transfer function.[16]

The second IQ phantom is the MITA-LCD (Catphan CCT189) phantom used to evaluate low contrast detectability. The standard MITA-LCD phantom is 200 mm in diameter and includes four cylindrical low contrast inserts of 14, 7, 5, and 3 HU, which at its standard adult size have diameters of 3, 5, 7, and 10 mm respectively.

Finally, a set of uniform cylinder water-equivalent phantoms was created to characterize noise properties.

These three sets of phantoms were defined analytically such that they can be generated at any desired size and resolution. For this study, eight phantom sizes were generated that span the range of pediatric effective diameters: 112, 131, 151, 185, 200, 216, 292, and 350 mm. While scaling the CTP404 sensitometry and MITA-LCD phantom diameters, their inserts can either be kept constant at their standard diameter or scaled proportionately. In this study the insert diameters were proportionately scaled to reflect the varying size of pediatric anatomy. For example, in the standard 200 mm diameter MITA-LCD phantom the largest insert is 10 mm in diameter, but when scaled down to 112 mm in phantom diameter, the largest insert diameter became 5.6 mm.

24734209, 2024, 2, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.16901 by Us Fda, Wiley Online Library on [19/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

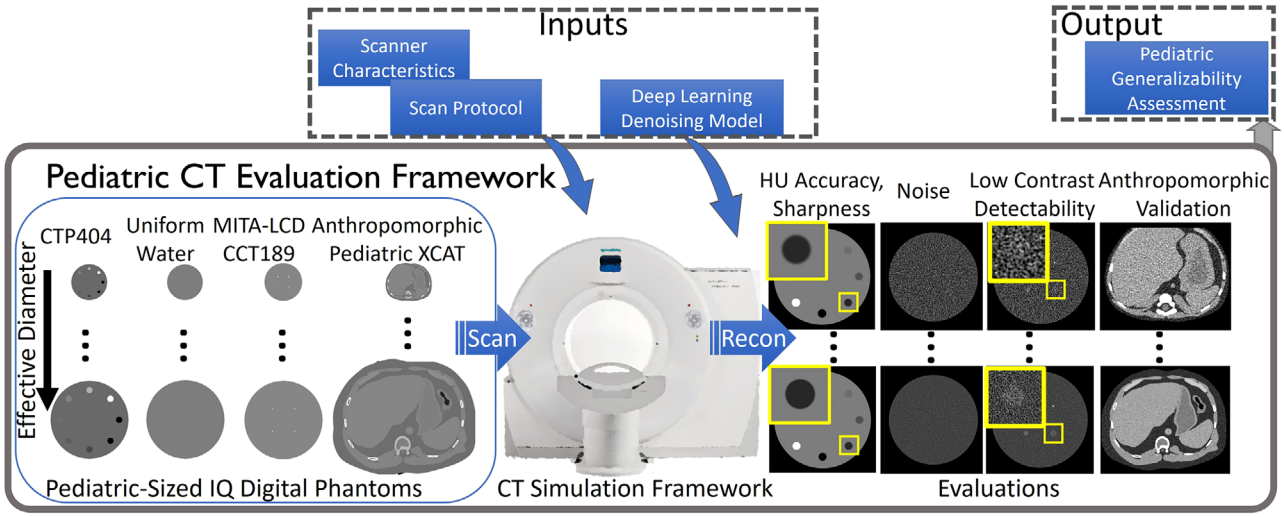

FIGURE 2 Pediatric CT Evaluation Framework summary diagram. The framework consists of a set of digital pediatric-sized image quality (IQ) and anthropomorphic phantoms virtually scanned on a CT simulation framework. The resulting scan projection data is then either reconstructed or given directly as inputs to a deep learning reconstruction or denoising model. The deep learning model and scan parameters are the primary inputs to the assessment framework. The framework output is a summary of image quality metrics as a function of phantom size that describe the model's pediatric generalizability.

2.1.2 | Image quality assessments

Image quality assessments were divided into task-independent physical image quality assessments and task-based assessments of low contrast detectability.

Physical image quality assessments such as CT number accuracy, sharpness, and noise properties are standard in the evaluation of image reconstruction and denoising. These assessments are the most basic yet very important tests in the domain of CT imaging.

When evaluating sharpness, nonlinear denoisers, including those utilizing DL, are known to perform variably under different contrast conditions. Thus, for each age-based phantom size investigated, contrast-dependent modulation transfer functions (MTF)[16] were calculated to assess sharpness at each contrast difference. The MTF curves were derived from radially averaged line profiles extending from the center of different contrast disks for each size of the CTP404 module.

The eight unique contrast inserts of the CTP404 module were also used in assessing CT number accuracy.

Noise properties were assessed with both noise magnitude and noise power spectra (NPS). Noise magnitude measurements were calculated as the standard deviation (std or σ) of pixels in a central circular region of interest (ROI) in the uniform water phantom for each phantom size with an ROI diameter equal to 1/3 the phantom diameter. 2D NPS measurements were taken from these same ROIs and calculated as the averaged 2D discrete Fourier transform (DFT_{2D}) of the noise only image[17]

$$NPS = \frac{\sum_{i=1}^{N_{sim}} \left| DFT_{2D} \left[I_i - \bar{I} \right] \right|^2}{N_{sim} N_x N_y} \frac{N_{sim}}{N_{sim} - 1}. \quad (1)$$

Here I_i refers to each of the N_{sim} repeat simulations of size N_x by N_y pixels within the ROI with different noise instances and $\bar{I}$ is the image averaged across these repeat scans. The difference of $I_i - \bar{I}$ removes the constant water component of the phantom leaving a noise-only image. Here $N_{sim}/(N_{sim} - 1)$ is a bias correction term. 1D radially averaged NPS curves were then extracted from these 2D NPS images. Note that NPS in this study, defined in **Eq**. 1, refers to the pixel NPS and is not scaled by the pixel dimensions, and thus is reported as a function of spatial frequency in units of cycles per pixel (cyc/pix). Reporting noise texture in units of cycles per pixel is more appropriate for characterizing the noise texture inputs to a denoising model as the model is not inherently knowledge of the length scale of each pixel.[18]

The final task-independent performance measure, CT number accuracy was calculated with each of the eight unique contrast inserts of the CTP404 module.

Task-based image quality refers to the imaging performance needed to accomplish a predefined clinical task. Well-designed image reconstruction and denoising methods should be able to allow a reduction of x-ray dose while maintaining or improving task performance for a fixed dose compared to the standard FBP reconstruction method.

24734209, 2024, 2, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.16901 by Us Fda, Wiley Online Library on [19/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

To assess the task-based performance of a DL denoiser in several pediatric subgroups, a low contrast detectability study was performed with mathematical model observers using the simulated MITA-LCD phantom with low contrast inserts of 14, 7, 5, and 3 HU. For each investigated diameter, 200 repeat simulated CT scans of the MITA-LCD phantom were performed to generate signal-present images. Another 200 repeat scans of the equal-size uniform water phantom were performed to generate a matched set of signal-absent images. These paired signal-present, signal-absent images were further cropped to ROIs around each low contrast insert to evaluate detectability as a function of lesion size and contrast level. From this dataset of paired signal-present and signal absent ROIs, 40% were used to train model observers and the remaining 60% were used for testing.

Different model observers can yield different results, thus two model observers were selected, a Laguerre-Gauss Channelized Hotelling observer (LG-CHO)[19] and a non-prewhitening observer (NPW)[20]. The LG-CHO is a linear approximation of an ideal observer capable of prewhitening, while the NPW is shown to be more responsive to noise reduction strategies.[5] Together this pair of efficient and inefficient model observers aim to establish bounds of detectability performance to be expected by human readers.

2.2 | DL denoising model and training

To demonstrate the use of this framework in identifying DL image denoising (DLID) sized-based performance, an initial investigation was performed using the image-based REDCNN DL denoising model.[2] REDCNN, like other image-based denoisers, takes as inputs filtered backprojection (FBP) reconstructed images yielding processed lower noise outputs. REDCNN is composed of ten convolutional layers followed by another ten deconvolutional layers each followed by rectified linear unit (ReLU) activations and with three skip connections between convolutional and deconvolutional layers. For more details on the implementation and design of REDCNN please refer to Chen et al.[2] and Zeng et al.[18] REDCNN was chosen as a DLID example for its relative simplicity and effective denoising performance. It is a widely used image-based deep learning denoiser by researchers in the CT field.[21] Furthermore, image-based denoisers are representative of the majority of DL denoising models available clinically.[21] It is also worth noting that the emphasis of this work is on the evaluation framework. We will discuss later that the currently implemented framework applies for any image-based denoising method and the generalizability performances in pediatric CT data are expected to depend on model architecture and training strategy.

For this case study the model was trained on patient data from the Low Dose Grand Challenge (LDGC) Dataset using low dose inputs and routine dose target images.[22] This dataset is composed of abdominal non-contrast CT exams from 10 adult patients with reconstructed fields of view ranging from 340 – 420 mm in diameter. Other major imaging parameters are summarized in Table 1. The dataset includes full- and simulated quarter-dose image pairs for model training and validation. Of the 10 patients included, 7 were used for model training with the remaining 3 used for model tuning. In this study only the 3 mm thick CT slices reconstructed with the Siemens D45f kernel, representing the sharp image series were used. Additional implementation details of the network optimization and parameter tuning process were described previously.[23]

2.3 | CT simulation parameters for test data generation

X-ray projection data for the simulated testing datasets were made using the Michigan Image Reconstruction Toolbox (MIRT)[25] by simulating a virtual 2D fan-beam CT scanner. Poisson noise was modeled at the detector, but electronic noise was not. MIRT was also used for FBP reconstruction of the digital phantom CT projection data to create images for evaluating the DL denoising model. The source to isocenter distance and source to detector distance, detector size, and reconstructed matrix size of the MIRT simulations (Table 1) were set to match the LDGC dataset used in the DL denoising model training.[18] To match the sharpness and noise texture of Siemens D45 sharp kernel used in the LDGC training dataset, the simulated test data was reconstructed with a Hanning filter with equivalent 50% and 10% modulation transfer function (MTF) cutoffs, referred to as a "D45 equivalent" kernel in Table 1.

X-ray flux in the CT simulations was varied to achieve a constant noise index of 23 HU at full dose for each sized phantom (Table 1). At full and quarter dose levels the noise magnitude measured at the center of the reference 200 mm uniform water digital phantom matched the mean noise in water of the LDGC dataset of 23 HU and 47 HU respectively. Then the tube output I_0 was scaled exponentially for different sized phantoms from its reference value (I_0, ref) in the d_{ref} = 200 mm uniform water phantom to yield constant noise magnitude relative to patient size:

$$I_0\ (d) = I_{0,ref}\ *\ exp(\mu_{water} * d)/exp\,(\mu_{water} * d_{ref})\,. \quad (2)$$

In the reference 200 mm water phantom an $I_{0,ref} = 3 \times 10^5$ photons/pixel yielded full dose noise magnitude of 23 HU. $I_{0,ref}$ was then proportionately scaled to achieve quarter dose simulated images.

24734209, 2024, 2, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.16901 by Us Fda, Wiley Online Library on [19/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 Dataset Imaging Parameters Description. The Low Dose Grand Challenge (LDGC)[22] was used as the training dataset for the denoising model evaluated using the Pediatric IQ and Anthropomorphic digital phantoms. The full dose noise index refers to the target noise level, measured in a water ROI, and was set to match the mean of the training dataset full-dose patient images. The noise index was tuned in simulations by varying the x-ray tube output with phantom size to achieve approximately constant noise magnitude according to Eq. 2.Water equivalent diameter (WED) was calculated based on patient attenuation.[24] In the simulated test datasets the reconstruction kernel was previously modeled as a Hanning filter found to best match the Siemens D45f kernel 50% and 10% MTF.[18]

Dataset	Full-Dose Noise Index (HU)	kVp (kV)	WED (mm) Range, Mean	FOV (mm) Range, Mean	Matrix Size (pixels)	Reconstruction Kernel
LDGC (train)	23	100-120	254-354, 309	340-420, 381	512×512	D45f
Pediatric IQ Phantoms (test)	23	120	112-350, 204	123-385, 254	512×512	D45f equivalent
Anthropomorphic Phantoms (test)	23	120	112-355, 212	123-392, 233	512×512	D45f equivalent

Where applicable, both pediatric and adult acquisition protocols followed abdomen/pelvis protocols recommended by the AAPM Alliance for Quality Computed Tomography.[26] The main difference in adult and pediatric protocols investigated in this study was reconstructed field of view (FOV). Adult acquisition protocols were defined as having a scanning and reconstructed field of view (FOV) of 340 mm in diameter or 110% the patient effective diameter depending on which was larger. Body fitting FOVs are routinely used in pediatric CT to make efficient use of dose and system spatial resolution. To reflect this practice, pediatric protocols had scanning and reconstructed FOVs defined as 110% the patient effective diameter. While all images were reconstructed in a 512 × 512 image matrix, this change in FOV size changes the reconstructed voxel size influencing noise texture and image sharpness. This is demonstrated in Figure 3 by comparing reconstructed images from an adult and newborn-sized IQ phantom. Images reconstructed in a smaller FOV have a finer voxel size and appear as a zoomed view. Features can thus appear blurrier if the pixel size is beyond the intrinsic image resolution, as observed in the newborn phantom image on top right of Figure 3. Additionally, while the noise magnitude is constant across phantom size due to the exposure control defined in **Eq**. 2, this zoomed view gives an apparent lower frequency noise texture. This can be seen in the bottom row of Figure 3 where the noise in inset image has a larger-grained appearance.

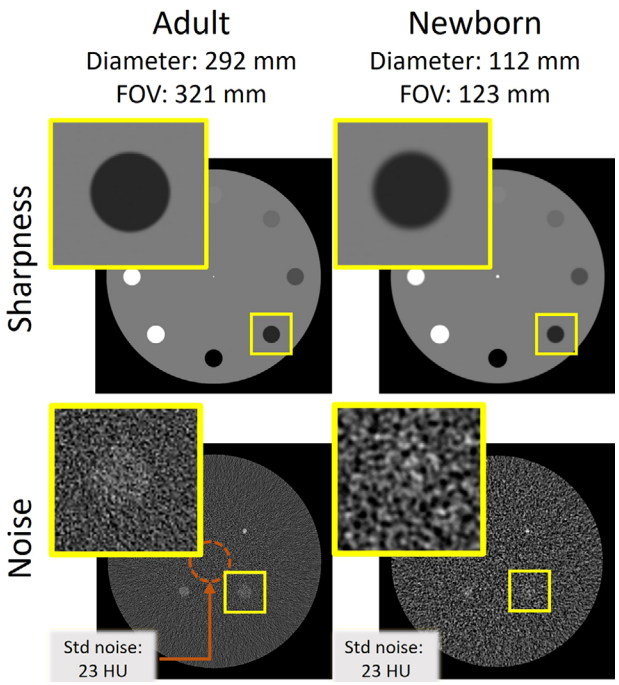

FIGURE 3 Representative full dose filtered backprojection (FBP) image quality after CT simulation of in silico pediatric IQ phantoms. When reconstructed in a field of view (FOV) 110% the body diameter, the image reconstructed in a smaller FOV with finer pixel size is like a zoomed view and can appear blurrier if the pixel size is beyond the intrinsic image resolution, as observed in the Newborn phantom image on top right of the figure.

2.4 | Anthropomorphic phantom validation

Since the DL denoising models are trained on patient data rather than phantoms, the pediatric XCAT cohort was included in the study validation to ensure that results found on the pediatric-sized IQ phantoms are representative of those to be expected in patients. In this simulation study, ground truth is available, which is the FBP reconstructed image from a noiseless sinogram. Noiseless FBP images were chosen as the ground truth because they contain the same streak and blurring artifacts induced by the deterministic imaging system transfer function as in the noisy test images. This choice better reflects the noise removal task of these denoisers which cannot be expected to remove other artifacts that they were not trained to remove. Thus, in this validation experiment, noise reduction was defined as reduction in root mean square error (RMSE) between the noisy reconstructed images x and the ground truth noiseless FBP y:

$$RMSE\ (x, y) = \sqrt{\sum_{i}^{N} \frac{(x_i - y_i)^2}{N}}\ . \qquad (3)$$

24734209, 2024, 2, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.16901 by Us Fda, Wiley Online Library on [19/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

RMSE reductions measured on the pediatric-sized phantoms were compared against RMSE reduction values determined from the XCAT anthropomorphic phantoms. This demonstrated that noise reduction trends with phantom size measured on IQ phantoms were representative of the noise reduction expected in pediatric subgroups of matching effective diameter.

3 | RESULTS

The pediatric-sized IQ phantom assessment framework was used to assess the pediatric generalizability of an adult-trained REDCNN DL denoising model. Below are the physical image quality assessments and task-based low contrast detectability results each reported as a function of patient size, followed by the validation experiment results using anthropomorphic phantoms.

3.1 | Physical image quality assessment

The physical image quality assessment results include noise, image sharpness and CT number accuracy assessments.

Figure 4 assesses the influence of patient size on noise image quality performance. Figure 4a features an ROI comparison between the FBP and DL denoised images. Despite having the same 24 HU noise standard deviation across phantom sizes in the FBP images, the noise has a finer texture in the larger diameter phantom due to differences in FOV size. Additionally, between the small newborn-sized and large adult-sized phantoms, only the adult-sized DLID processed image saw a reduction in standard deviation noise.

Figure 4b presents noise in spatial frequencies as 2D NPS. The finer noise textures in the adult-phantom images correspond to more pronounced higher frequency noise components towards the periphery of the 2D NPS. Following denoising of the adult-phantom images, the noise power is lower at all frequencies compared to its FBP input. As for the newborn-sized phantom, there is little change in the ROI image or 2D NPS following denoising by the model.

In Figure 4c 1D radially averaged NPS curves are plotted for several phantom sizes to show NPS trends versus phantom size. In the smaller phantom sizes NPS curves appear to peak at primarily lower frequency before denoising, while larger phantom sizes have higher frequency noise components. In the smaller phantom sizes with lower mean frequency NPS there is little difference following DL denoising. However, in the larger diameter phantoms the gap between FBP and DLID NPS curves increases as REDCNN primarily removes higher frequency noise components.

This size-dependent noise reduction is further demonstrated by plotting the measured noise reduction as a percent of FBP vs phantom diameter in Figure 4d. Age groups with equivalent waist mean effective diameters are overlaid to demonstrate which pediatric subgroups are represented with each phantom diameter. These results suggest that while larger phantoms, corresponding to larger and older subgroups benefit from noise reductions over 60%, smaller patients under 5 years old see less benefit, with almost no noise reduction expected for newborns and infants based upon their size.

Nonlinear denoising methods are known to better preserve sharpness at high contrasts. To assess the influence of patient size on a DL denoiser's image sharpness performance, MTF curves were calculated at six different contrast levels of the CTP404 phantom and compared between the original FBP image and the DL denoised image. These MTF curves were measured from 20 repeat scans at 100% relative dose level. Sharpness was assessed as 50% and 10% MTF, the spatial frequencies where 50% and 10% of signal are retained in an ideal edge measurement. The differences in 50% MTF and 10% MTF frequencies between FBP and the DL denoised images ($\Delta MTF50$ and $\Delta MTF10$) were then used to assess the change in sharpness. These changes in sharpness were then plotted against phantom diameters in Figure 5. At high contrasts the DL denoised images showed little loss in sharpness at 50% or 10% MTF, but sharpness was observed to degrade in lower contrasts. These trends are consistent with other nonlinear reconstruction and denoising methods[4] that tend to smooth lower-contrast features. Furthermore, these trends of reduced sharpness at low contrast were exacerbated in larger phantom sizes. This indicates a greater reduction in sharpness in larger patients following DL denoising. This finding is consistent with Figure 4c that shows denoising performance also increasing with phantom diameter. Together these suggest that at phantom diameters and FOVs far from the DL training distribution the DL denoiser acts more like an identity operator, leaving the image unchanged. However, at larger diameters and FOVs the denoiser has greater influence on both the noise and sharpness of the resulting image.

CT number accuracy was also evaluated as a function of contrast level and phantom diameter at 25% dose level. Across contrast levels and diameters, mean HU differences between FBP and DLID were observed to be within 2 HU and not found to be substantially influenced by contrast level or phantom size.

3.2 | Task-based image quality assessment

To assess the impact of the DL denoiser on a low contrast detectability task, two different model observers were used: a Laguerre-Gauss Channelized Hotelling

24734209, 2024, 2, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.16901 by Us Fda, Wiley Online Library on [19/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

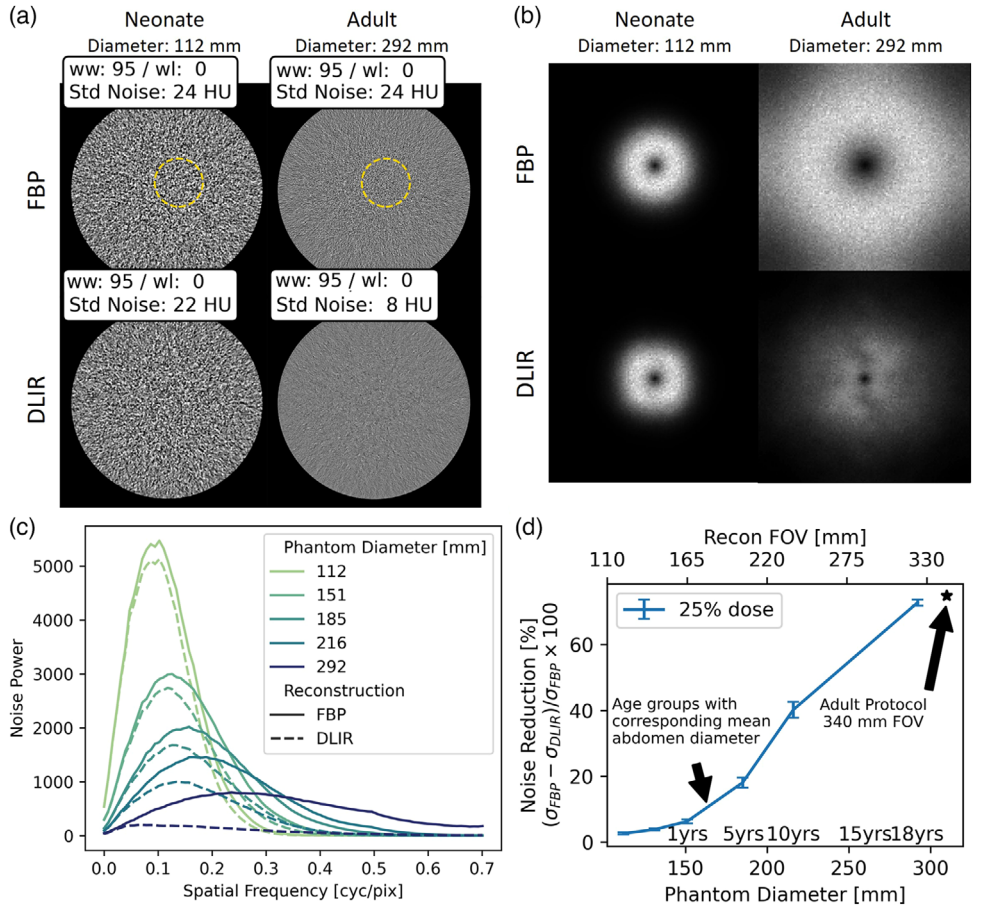

FIGURE 4 Noise assessment. a) Region of interest (ROI) comparison of uniform phantom images with standard deviation noise measurements. b) 2D noise power spectra images from a), note that brightness has been increased for improved visualization. c) 1D radial averaged NPS profiles show apparent increasing NPS, in cycles per pixel, at larger phantom diameter. DL models are oblivious to pixel size and are sensitive to these apparent changes in NPS. This can be observed also in d), which compares noise reduction performance versus phantom diameter, and by extension, fields of view (FOV). The pediatric reconstruction protocols have reconstructed FOV 110% their effective diameter while the plotted adult protocol has a fixed 340 mm reconstructed FOV. Noise reduction is calculated as the percent difference following denoising measured in the uniform water phantom.

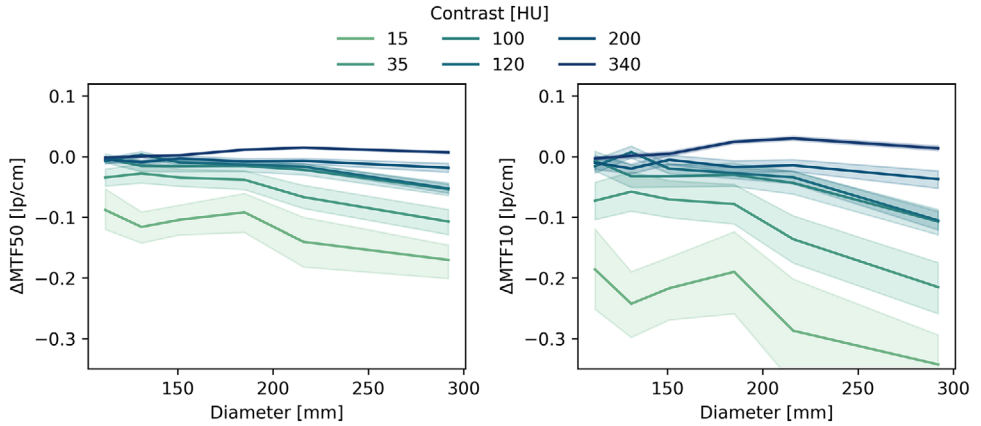

FIGURE 5 Evaluation of image sharpness as a function of phantom diameter and contrast level as measured with contrast dependent Modulation Transfer Function (MTF). Relative difference in sharpness is defined as the difference in 50% MTF frequency of DLID relative to FBP (**ΔMTF**50 [lp/cm]) and is plotted against phantom diameter for different contrast levels. Both ***ΔMTF***50 and higher frequency ***ΔMTF***15 results are shown. Bands indicate 95% confidence interval following 20 repeat measurements at 100% dose.

24734209, 2024, 2, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.16901 by Us Fda, Wiley Online Library on [19/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

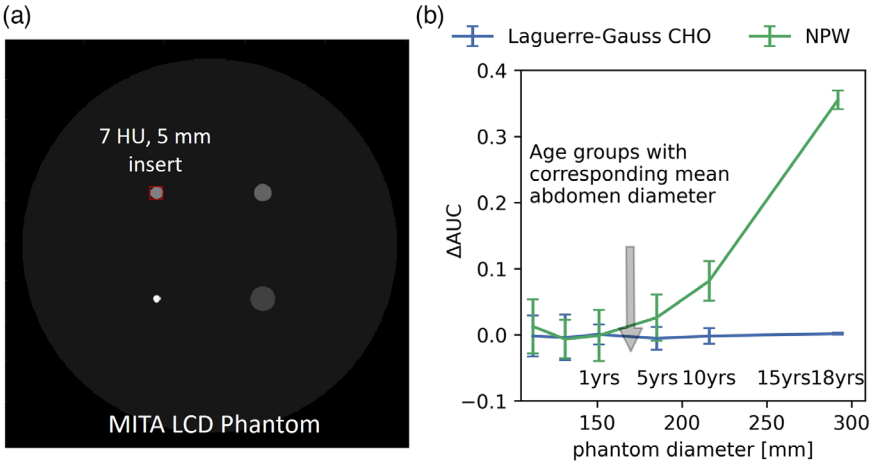

FIGURE 6 Low contrast detectability image quality assessment measured as a function of diameter of the pediatric-sized standard QA phantom (MITA-LCD), shown in a). Low contrast detectability performance is reported in b) as the difference in the area under the receiver operating characteristic curve (**Δ*AUC***) following DLID. Two types of model observer were used. First, a Laguerre-Gauss Channelized Hotelling Observer (CHO) able to decorrelate noise (prewhiten) and second, a nonprewhitening (NPW) observer. Different age groups in years (yrs) are overlaid on the plot where the phantom diameter matches the mean effective diameter of each subgroup.

Observer (CHO) able to decorrelate noise (prewhiten) and a nonprewhitening (NPW) observer. Task performance was calculated as the model observer's area under the receiver operating characteristic curve (AUC). To show potential improvements in the task following DL denoising, the difference in AUCs between the DL denoised images and standard FBP reconstruction (Δ*AUC*) is reported in Figure 6 at 25% dose. Using this measure of performance, task-based performance was shown to decrease progressively in smaller phantom diameters when measured with the less efficient NPW model observer. Similar trends were observed across the four inserts so only the 7 HU, 5 mm insert is shown in Figure 6.

When low contrast detectability was measured using the Laguerre-Gauss CHO observer, no statistically significant change in detectability was observed following DL denoising in any phantom diameter. This is because the LCD difference before and after applying our trained denoising model was very small on the adult test data with LG-CHO, making the performance gap between the adult and pediatric CT hardly identifiable in this test. This suggests that for this evaluated REDCNN model, the denoising was found to most benefit an observer unable to decorrelate noise as is the case with the NPW model. More discussion on the tradeoffs and limitations of these model observers is discussed later in **Section 5**.

3.3 | Anthropomorphic phantom validation

Virtual anthropomorphic phantoms were used to validate the noise reduction estimates given by the pediatric IQ phantoms. Example images of adult and pediatric anthropomorphic phantoms both before and after DL denoising are shown in Figure 7a. When applied to the quarter-dose FBP adult images (Figure 7a top), the DL denoising model reduces noise across the image. Noise is both visibly reduced and has a lower measured voxel standard deviation in the homogenous liver, from 49 HU before denoising to 11 HU after. On the contrary, the DL denoising model has no perceptible denoising effect on the newborn patient image (Figure 7a bottom). There is visually no change in noise compared to the low-dose FBP input image and the measured standard deviation is the same. This outcome is consistent with the uniform phantom-based findings in Figure 4a that the adult-trained REDCNN model was more effective in larger patients. Figure 7b confirms these findings quantitatively by plotting the percent decrease in RMSE against the anthropomorphic patient diameter. RMSE reductions from the uniform pediatric IQ phantoms of matching diameter are overlaid on Figure 7b and trend consistently with the anthropomorphic phantom results. This suggests that even in a simple uniform phantom, size and reconstructed FOV have an outsized influence on DLID denoising performance and can reasonably estimate noise reduction in an anthropomorphic phantom or patient of matched size. However, patient size does not determine all the noise reduction performance as there is still some underestimation of RMSE reduction in the uniform phantom relative to the anthropomorphic phantom.

4 | DISCUSSION

Deep learning image reconstruction and denoising techniques are potentially valuable CT dose reduction tools for pediatric patients. However, pediatric patients are underrepresented in healthcare, receiving proportionately far fewer medical imaging exams than adults, making evaluating the safety and effectiveness of deep learning-enabled tools challenging. To address this gap, this work developed a set of digital pediatric image quality (IQ) phantoms and evaluation framework (Figure 2) to assess the efficacy of DL CT denoising algorithms on pediatric patients based on one simple yet critical difference from adult patients: body size. Using virtual

24734209, 2024, 2, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.16901 by Us Fda, Wiley Online Library on [19/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

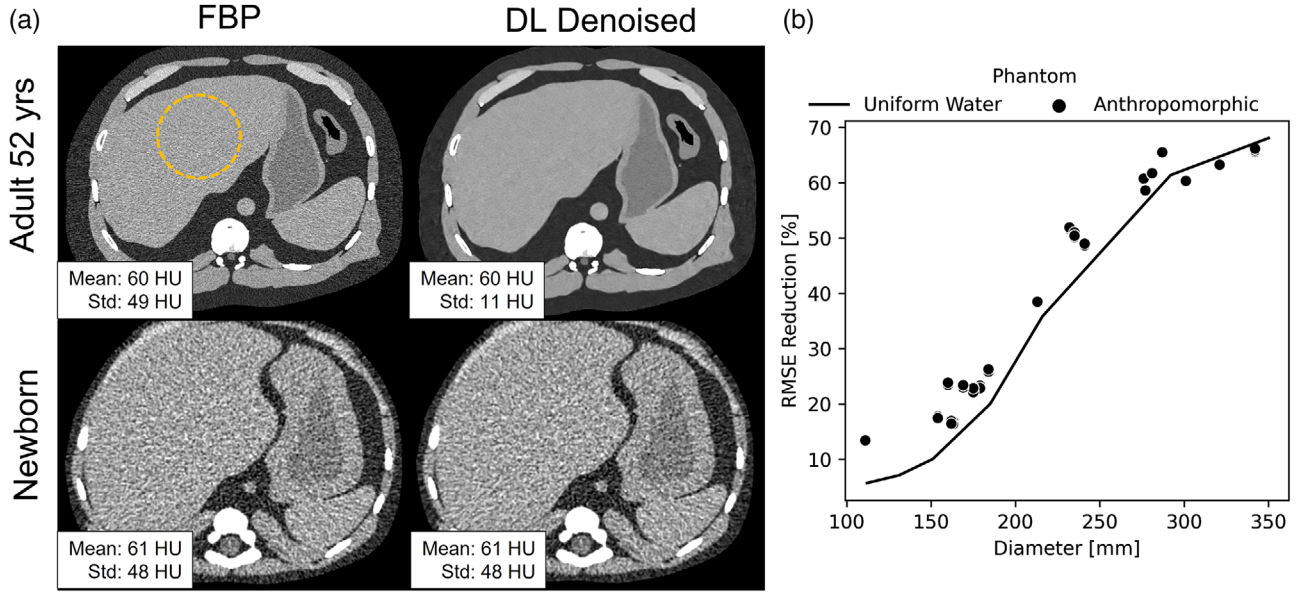

FIGURE 7 Adult and pediatric XCAT phantoms were imaged in our simulation framework at 25% dose level to validate our use of pediatric-sized image quality (IQ) phantoms to characterize DLID model performance in different pediatric subgroups. a) shows abdominal cross sections from a small newborn pediatric phantom and a large adult phantom. b) plots noise reduction, calculated as root mean square error (RMSE) reduction following DLID denoising as a function of diameter when applied to uniform water IQ phantoms and anthropomorphic phantoms.

pediatric-sized versions of standard image quality phantoms and associated image quality assessments, the proposed pediatric IQ phantom framework can assess pediatric performance on size-based subgroups.

The present study performed an initial investigation into the use of the proposed pediatric IQ phantom assessment framework to evaluate the performance of a REDCNN denoising model trained on adult CT image data and applied on pediatric images. Using the framework, we observed the REDCNN model to have superior noise reduction in large patient diameters and FOVs. These large sizes are consistent with the adult patients in the training dataset (FOV $\geq$ 340 mm). Noise reduction performance was observed to be >60% reduction in pixel standard deviation in larger patients corresponding to adolescents and adults. However, an inflection in performance was observed in smaller phantoms, starting at around 220 mm in effective diameter, equal to the mean-size of a 10-year-old based on reference data. Here performance dropped from >60% to around 40%. Performance was found to be worst, at less than 20%, for the smallest investigated diameters representing subgroups 5 years-old and younger. At these sizes, using body fitting reconstructed FOVs, the DL model was observed to have minimal influence on the input image.

While these findings are specific to the model investigated in this study, we anticipate the trend of degraded performance to exist for any DL denoising model tested on noise textures different from its training set. These findings of DL denoising performance being sensitive to changes in FOV are consistent with other studies of image-based DL denoising models that concluded convolutional neural network (CNN)-based denoising performance decays under imaging factors that change the noise texture of the image. These factors include FOV and recon kernel.[18,27] The present study further investigated the implications of this local noise texture and FOV sensitivity to pediatric imaging. We anticipate that by leveraging different model designs and training strategies, including a wider range of training FOVs, to likely improve the patient size-generalizability of denoising models which would also benefit the generalizability to pediatric patients.

To better understand why DL denoising models are more sensitive to changes in FOV consider that the smaller reconstructed FOV decreases the effective voxel size. This causes the influence of geometric blurring to become more apparent and spread across more voxels in the reconstructed matrix. This can be visualized by comparing the visual edge sharpness and extent of noise correlation (noise grain size) between images of the adult-sized and newborn-sized phantoms in Figure 3 reconstructed with the same FBP reconstruction kernel. This longer scale noise correlation (larger noise grain size) in the smaller FOV reconstructed newborn appears to the CNN, which has no understanding of physical pixel size, as lower spatial frequency noise components (smaller diameter solid curves in Figure 4c). The investigated REDCNN was trained exclusively on adult patient image data (Table 1)

24734209, 2024, 2, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.16901 by Us Fda, Wiley Online Library on [19/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

with noise power extending to higher spatial frequencies (in cycles per pixel), like the larger diameter solid curves in Figure 4c. This REDCNN then generalized poorly to small phantom diameter FBP inputs with lower mean apparent NPS frequencies. These findings suggest that the smaller reconstructed FOVs used in scanning smaller patients is a key image characteristic that can cause CNN-based models to generalize poorly. However, the use of body fitting FOVs is common practice in pediatric abdominal imaging. While a larger adult FOV could be used to reconstruct data from smaller patients for more consistent DLID performance, this would come at the cost of resolving small spatial features necessary to image the smaller patient anatomy. A better solution would be to utilize DLID models specifically designed for smaller pediatric patients or utilize data augmentation strategies that incorporate lower frequency noise textures in training data as would be found in smaller FOV pediatric images.

The developed set of pediatric IQ phantoms and evaluation framework proposed in this work can be used to assess the size-dependence of DL denoising models and thus better understand how the model would perform in patients of different sizes including small adult and pediatric patients. The pediatric evaluation framework uses geometric IQ phantoms routinely used in bench testing-based CT image quality evaluation and informs the fundamental imaging performance of DLID methods. Thus, the performance predicted with the simulated images can be validated several ways including with a small set of real phantoms. A set of small, medium, and large IQ phantoms covering the desired patient size range could be used to estimate or at least flag for poor performance in different patient size-based subgroups for a given DL denoising model. Some existing physical IQ phantoms, such as the CATPHAN600 insert without the housing, about 12 cm in diameter, can be utilized for the purpose of testing the generalizability of DLID on small-size pediatric CT. Similar to the MURCURY4.0 phantom (Sun Nuclear, Melbourne, FL) that consists of five sections of diameter from 16 and 36 cm, a pediatric IQ phantom consisting of multiple sections of diameter representing age groups from newborn to 12 years-old may be manufactured to allow the evaluation of the DLID performance across a broader size range of pediatric populations using physical scans. Additionally phantom printing using either iodinated inkjet printing methods[28] or PixelPrint 3D printing[29] could also be employed to physically reproduce a subset of the digital pediatric IQ phantoms for further physical validation of predicted performance from simulations.

The current study and framework have limitations. First, only a single image-based DL denoising model was evaluated. The principles behind the proposed framework are not exclusive to image-based methods and future work aims to conduct comparisons between different DL reconstruction methods including projection-based denoising models and MBIR with deep priors. This will allow us to determine the DL methods or training methods, such as data augmentation, that have better generalizability in pediatric and small patients.

A second limitation is that the CT simulations in the current implementation utilizing MIRT made several simplifying assumptions in favor of reduced simulation computation time. These include monoenergetic source and infinitesimally small x-ray focal spot, 2D fan-beam, and not incorporating detector blur, bowtie filter, and the effects of scatter. Further planned iterations of the framework's simulations will leverage existing CT simulation frameworks such as XCIST[30] that include these factors as well as helical acquisition and recon with different slice thicknesses to better match real scanner data characteristics.

Additionally, the use of standard image quality phantoms enables measurement of essential CT image performances including MTF and NPS, however their simple uniform background is not representative of the textures and features observed in real patients. This could be responsible for the remaining differences of REDCNN's noise reduction rate between the uniform and anthropomorphic phantoms observed in Figure 7. Later versions of the framework aim to further leverage the virtual patient cohorts, like XCAT,[31] to utilize more realistic patient anatomy and local textures for more representative assessments beyond body size.

A final limitation of our framework is on the use of model observers to estimate low-contrast lesion detectability. Model observers are useful for estimating image performance without the need of expensive and time-consuming reader studies with real human readers. A limitation of using model observers is that the conclusions drawn can vary widely based on the choice of model observer. In this relatively simple task of detecting a disk signal in flat background containing correlated noise, a nonprewhitening (NPW) model observer was shown to correlate well with human performance.[32,33] In contrast, Channelized Hotelling Observers (CHO) such as the Laguerre-Gauss CHO estimate the linear ideal observer and their performance more closely relates to the underlying information in the image.[19] Both observers were included in this study to see how their assessment of task performance varies with phantom size. The NPW has been used in other DLID studies to assess task performance.[5,34] Our findings with the NPW observer are consistent with those previous works in that task performance as measured by a NPW observer benefits from noise reduction. However, NPW models have also been reported as overestimating detection AUC following nonlinear noise reduction such as IR and DLID.[35] Meanwhile, when task-performance was measured with a Laguerre-Gauss CHO in this study, no task-benefit was measured, indicating that there was no improvement in the underlying information from the perspective of the estimated linear ideal observer. In

24734209, 2024, 2, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.16901 by Us Fda, Wiley Online Library on [19/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

conclusion, while NPW task results are likely an overestimate of the expected task-benefit of DLID and LG-CHO is likely an underestimate. Actual human reader performance may be somewhere in between as humans have been shown to be able to prewhiten noise but not as efficiently as ideal model observers.[32,36]

Medical devices are generally not designed initially for children, yet at some point most medical devices will be used on children regardless of whether that was their designed population or not.[37] For some applications the safety and effectiveness of a device might not be sensitive to patient age or size. However, as medical devices increasingly incorporate more data-driven methods, it is crucial to understand the safety and effectiveness in pediatric populations that are not thoroughly represented in training datasets. This work aims to support a culture of informed development and use of devices to better serve these patients. While evaluations relying on simulation reduce the burden of implementation, they should be used to supplement existing patient evaluations when pediatric data is scarce rather than replace it entirely. Simulations can produce large amounts of data representing key aspects of pediatric patients, but they still lack in realism and diversity available in real patient data and thus should still be validated with physical scan data. The development of large public repositories of anonymized pediatric data and improved generative methods in time may help address this shortcoming of current simulation methods. However, constant vigilance of device safety and performance in pediatric and other vulnerable populations should continue to be a priority in the development and use of medical devices.

5 | CONCLUSION

We developed a framework for pediatric evaluation of deep learning image reconstruction and denoising models. Using this framework, we tested the generalizability of an adult-data trained REDCNN model in different pediatric subgroups by size. The results show that the DL denoising model's noise reduction capability decreased with the simulated pediatric phantom size, in both the uniform and anthropomorphic backgrounds. FOV differences between adult and pediatric protocols were identified as contributing to reduced noise reduction and task performance in the evaluated model. Our work highlights the importance of assessing pediatric generalizability for DL denoising algorithms and demonstrates a means of performing these assessments.

ACKNOWLEDGEMENTS

Funding for this work was provided by an the National Center for Toxicological Research's Perinatal Health Center of Excellence within the Food and Drug Administration.

CONFLICT OF INTEREST STATEMENT

The authors have no conflicts to disclose.

DATA AVAILABILITY STATEMENT

The simulation code and simulated phantom CT images that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

1. Hall EJ. Lessons we have learned from our children: cancer risks from diagnostic radiology. *Pediatr Radiol*. 2002;32(10):700-706. doi:10.1007/s00247-002-0774-8
2. Chen H, Zhang Y, Kalra MK, et al. Low-Dose CT With a Residual Encoder-Decoder Convolutional Neural Network. *IEEE Trans Med Imaging*. 2017;36(12):2524-2535. doi:10.1109/TMI.2017.2715284
3. Chen H, Zhang Y, Chen Y.et al. LEARN: Learned Experts' Assessment-Based Reconstruction Network for Sparse-Data CT. *IEEE Trans Med Imaging*. 2018;37(6). doi:10.1109/TMI.2018.2805692
4. Szczykutowicz TP, Toia GV, Dhanantwari A, Nett B. A Review of Deep Learning CT Reconstruction: Concepts, Limitations, and Promise in Clinical Practice. *Curr Radiol Rep*. 2022;10(9):101-115. doi:10.1007/s40134-022-00399-5
5. Brady SL, Trout AT, Somasundaram E, Anton CG, Li Y, Dillman JR. Improving Image Quality and Reducing Radiation Dose for Pediatric CT by Using Deep Learning Reconstruction. *Radiology*. 2021;298(1):180-188. doi:10.1148/radiol.2020202317
6. Smith-Bindman R, Kwan ML, Marlow EC, et al. Trends in Use of Medical Imaging in US Health Care Systems and in Ontario, Canada, 2000–2016. *JAMA*. 2019;322(9):843-856. doi:10.1001/jama.2019.11456
7. Moen TR, Chen B, Holmes III DR, et al. Low-dose CT image and projection dataset. *Med Phys*. 2021;48(2):902-911. doi:10.1002/mp.14594
8. McDowell MA, Fryar CD, Ogden CL, Flegal KM. Anthropometric Reference Data for Children and Adults: United States, 2003–2006: (623932009-001). Published online 2008. doi:10.1037/e623932009-001
9. Pediatric Information for X-ray Imaging Device Premarket Notifications - Guidance for Industry and Food and Drug Administration Staff. Published online. 2017.
10. Boone J, Strauss K, Cody D, et al. *Size-Specific Dose Estimates (SSDE) in Pediatric and Adult Body CT Examinations*. AAPM;2011. doi:10.37206/143
11. Vital and Health Statistics, Series 3, Number 46.:44.
12. Health C for D and R. Premarket Assessment of Pediatric Medical Devices. U.S. Food and Drug Administration. Published March 24, 2020. Accessed March 1, 2023. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/premarket-assessment-pediatric-medical-devices
13. Yoon H, Kim J, Lim HJ, Lee MJ. Image quality assessment of pediatric chest and abdomen CT by deep learning reconstruction. *BMC Med Imaging*. 2021;21(1):146. doi:10.1186/s12880-021-00677-2
14. Sun J, Li H, Wang B, et al. Application of a deep learning image reconstruction (DLIR) algorithm in head CT imaging for children to improve image quality and lesion detection. *BMC Med Imaging*. 2021;21(1):108. doi:10.1186/s12880-021-00637-w
15. Brendlin AS, Schmid U, Plajer D, et al. AI Denoising Improves Image Quality and Radiological Workflows in Pediatric Ultra-Low-Dose Thorax Computed Tomography Scans. *Tomography*. 2022;8(4):1678-1689. doi:10.3390/tomography8040140
16. Richard S, Husarik DB, Yadava G, Murphy SN, Samei E. Towards task-based assessment of CT performance: system and

24734209, 2024, 2, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.16901 by Us Fda, Wiley Online Library on [19/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

object MTF across different reconstruction algorithms. *Med Phys*. 2012;39(7):4115-4122. doi:10.1118/1.4725171
17. Siewerdsen JH, Cunningham IA, Jaffray DA. A framework for noise-power spectrum analysis of multidimensional images. *Med Phys*. 2002;29(11):2655-2671. doi:10.1118/1.1513158
18. Zeng R, Lin CY, Li Q, et al. Performance of a deep learning-based CT image denoising method: Generalizability over dose, reconstruction kernel, and slice thickness. *Med Phys*. 2022;49(2):836-853. doi:10.1002/mp.15430
19. Gallas BD, Barrett HH. Validating the use of channels to estimate the ideal linear observer. *J Opt Soc Am A*. 2003;20(9):1725. doi:10.1364/JOSAA.20.001725
20. Abbey CK, Barrett HH. Human- and model-observer performance in ramp-spectrum noise: effects of regularization and object variability. *J Opt Soc Am A, JOSAA*. 2001;18(3):473-488. doi:10.1364/JOSAA.18.000473
21. Wang G, Ye JC, De Man B. Deep learning for tomographic image reconstruction. *Nat Mach Intell*. 2020;2(12):737-748. doi:10.1038/s42256-020-00273-z
22. McCollough CH, Bartley AC, Carter RE, et al. Low-dose CT for the detection and classification of metastatic liver lesions: Results of the 2016 Low Dose CT Grand Challenge. *Med Phys*. 2017;44(10):e339-e352. doi:10.1002/mp.12345
23. Kc P, Zeng R, Farhangi MM, Myers KJ. Deep neural networks-based denoising models for CT imaging and their efficacy. In: Bosmans H, Zhao W, Yu L, eds. *Medical Imaging 2021: Physics of Medical Imaging*. SPIE;2021:16. doi:10.1117/12.2581418
24. McCollough C, Bakalyar D, Bostani M, et al. *Use of Water Equivalent Diameter for Calculating Patient Size and Size-Specific Dose Estimates (SSDE) in CT*. AAPM;2014. doi:10.37206/146
25. Fessler JA. Michigan Image Reconstruction Toolbox. Accessed November 10, 2023. https://web.eecs.umich.edu/~fessler/code/index.html
26. AAPM CT Scan Protocols - The Alliance for Quality Computed Tomography. Accessed 2023. https://www.aapm.org/pubs/ctprotocols/default.asp?tab=5#CTabbedPanels
27. Huber NR, Missert AD, Yu L, Leng S, McCollough CH. Evaluating a Convolutional Neural Network Noise Reduction Method When Applied to CT Images Reconstructed Differently Than Training Data. *J Comput Assist Tomogr*. 2021. doi:10.1097/RCT.0000000000001150
28. Ikejimba LC, Graff CG, Rosenthal S, et al. A novel physical anthropomorphic breast phantom for 2D and 3D x-ray imaging. *Med Phys*. 2017;44(2):407-416. doi:10.1002/mp.12062
29. Shapira N, Donovan K, Mei K, et al. PixelPrint: Three-dimensional printing of realistic patient-specific lung phantoms for CT imaging. *Proc SPIE Int Soc Opt Eng*. 2022;12031:120310N. doi:10.1117/12.2611805
30. Wu M, FitzGerald P, Zhang J, et al. XCIST-an open access x-ray/CT simulation toolkit. *Phys Med Biol*. 2022;67(19). doi:10.1088/1361-6560/ac9174
31. Segars WP, Norris H, Sturgeon GM, et al. The development of a population of 4D pediatric XCAT phantoms for imaging research and optimization. *Med Phys*. 2015;42(8):4719-4726. doi:10.1118/1.4926847
32. Barrett HH, Yao J, Rolland JP, Myers KJ. Model observers for assessment of image quality. *Proc Natl Acad Sci USA*. 1993;90(21):9758-9765. doi:10.1073/pnas.90.21.9758
33. Myers KJ, Barrett HH, Borgstrom MC, Patton DD, Seeley GW. Effect of noise correlation on detectability of disk signals in medical imaging. *J Opt Soc Am A, JOSAA*. 1985;2(10):1752-1759. doi:10.1364/JOSAA.2.001752
34. Nagayama Y, Goto M, Sakabe D, et al. Radiation Dose Reduction for 80-kVp Pediatric CT Using Deep Learning–Based Reconstruction: A Clinical and Phantom Study. *Am J Roentgenol*. 2022;219(2):315-324. doi:10.2214/AJR.21.27255
35. Chen B, Yu L, Leng S, et al. Predicting detection performance with model observers: Fourier domain or spatial domain? *Proc SPIE Int Soc Opt Eng*. 2016;9783:978326. doi:10.1117/12.2216962
36. Abbey CK, Samuelson FW, Zeng R, Boone JM, Eckstein MP, Myers K. Classification images for localization performance in ramp-spectrum noise. *Med Phys*. 2018;45(5):1970-1984. doi:10.1002/mp.12857
37. SECTION ON CARDIOLOGY AND CARDIAC SURGERY, SECTION ON ORTHOPAEDICS, Jenkins KJ, et al. Off-Label Use of Medical Devices in Children. *Pediatrics*. 2017;139(1):e20163439. doi:10.1542/peds.2016-3439

How to cite this article: Nelson BJ, Kc P, Badal A, Jiang L, Masters SC, Zeng R. Pediatric evaluations for deep learning CT denoising. *Med Phys*. 2024;51:978–990. https://doi.org/10.1002/mp.16901

24734209, 2024, 2, Downloaded from https://aapm.onlinelibrary.wiley.com/doi/10.1002/mp.16901 by Us Fda, Wiley Online Library on [19/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License