UNIVERSITY OF GOTHENBURG

DEPARTMENT OF MEDICAL RADIATION SCIENCES
SAHLGRENSKA ACADEMY

Evaluation of deep learning image reconstruction for brain- and abdominal CT

A visual grading characteristics study

M.Sc. Thesis
Tuva Skarp

Essay/Thesis:	30 hp
Program and/or course:	Medical Physics
Level:	Second Cycle
Term/year:	Fall/2021
Supervisor:	Angelica Svalkvist and Maria Larsson
Examiner:	Magnus Båth
Keywords:	computed tomography, DLIR, VGC study, deep learning

Acknowledgements

I would first like to thank my supervisors Angelica Svalkvist and Maria Larsson for their support and guidance. Your inputs and ideas have been a great help to me and to the project. Thank you for always taking your time to answer my questions and providing encouraging comments!

Secondly, I am very grateful to the nurses Sara Ku, Helén Milde and Marit Johannesson for their help with image acquisition at the CT-lab and overall support. I want to thank Sara especially for her great help in this project.

I would also like to thank the physicians that participated in this study, Anna Molinder, Magnus Palmér, Tim Unnerstall, Maria Chatzopoulou, Christian Weiergang and Fredrik Thorén. I would like to thank Anna and Magnus for helping me with the image quality criteria used in the study and overall help with their knowledge of anatomy.

I am also grateful to Magnus Bard for helping to export the images from the CT and to Magnus Båth for his help with determining noise power spectrum.

Lastly, I would like to thank my friend Andreas for taking his time to proofread and provide insightful comments to the report, and my best friend Caroline for always supporting me!

Abbreviations

ALARA – As low as reasonably achievable

AP – anterio-posterior

ASIR-V – Adaptive statistical iterative reconstruction V

AUC – Area under the curve

BMI – Body mass index

CI – Confidence interval

CT – Computed tomography

$CTDI_{vol}$ – Computed tomography dose index-volume [mGy]

DICOM – Digital Imaging and Communication in Medicine

DLIR – Deep learning iterative reconstruction

DLP – Dose-length product [mGy*cm]

DNN – Deep neural network

FBP – Filtered back projection

HU – Hounsfield units

IR – Iterative reconstruction

LAT – Lateral

NI – Noise index

NPS – Noise power spectrum

ODM – Organ dose modulation

PACS – Picture Archiving and Communication system

ROC – Receiver operating characteristics

ROI – Region of interest

SD – Standard deviation

VGC – Visual grading characteristics

ViewDEX - Viewer for Digital Evaluation of X-ray Images

Abstract

Computed tomography (CT) is a common medical imaging method today. There are different ways of reconstructing the images from a CT examination. The standard method is the filtered back projection (FBP). However, a disadvantage of FBP is the need of a relatively high exposure in order to reduce the noise in the reconstructed images, which leads to high radiation doses to the patients. Therefore, iterative reconstruction (IR) methods have been developed which use statistical CT imaging models to estimate and reconstruct an image. This method reduces the radiation dose to the patient but also changes the image texture. Deep learning image reconstruction (DLIR) is a new reconstruction method based on deep neural networks (DNN). It has been shown that images reconstructed using DLIR have similar image texture as images reconstructed using FBP, while the image noise is reduced.

At Sahlgrenska University Hospital, there are three CT-systems (GE Revolution Apex™ CT, GE Healthcare, Milwaukee, USA) equipped with True Fidelity which is a DLIR software. The present study aims at evaluating the image quality for DLIR with different post-processing filters compared to ASIR-V for abdominal and brain examinations.

Images from 20 abdominal examinations and 20 brain examinations were collected retrospectively and reconstructed. The reconstructions made for the abdominal examinations were; *Stnd* ASIR-V 40% 3 mm, *Stnd* DLIR-M + E1 3 mm, *Stnd* DLIR-M + E1 0.625 mm, *Stnd* DLIR-H + E1 3 mm and *Stnd* DLIR-H + E1 0.625 mm. The reconstructions made for the brain examination were; *Soft* ASIR-V 50% EC1 3 mm, *Stnd* DLIR-M EC1 3 mm, *Stnd* DLIR-M EC2 3 mm, *Stnd* DLIR-H EC1 3 mm and *Stnd* DLIR-H EC2 3 mm. The reconstructed images were evaluated in a visual grading characteristics (VGC) study where the images were rated according to five image quality criteria. Three abdominal radiologists and three neuro radiologists participated in the study.

Three phantom studies were also performed. A Catphan phantom was used to evaluate if the spatial resolution, visibility of low contrast objects and linearity of HU were affected by the different reconstructions. An anthropomorphic phantom was used to show the effect on image quality when lowering the radiation dose. Lastly, a water phantom was used to compare the noise properties between the different reconstructions.

The results from the VGC study of the abdominal examinations showed that the images reconstructed using DLIR were rated significantly higher than the images reconstructed using ASIR-V 40%. The abdominal radiologists rated the DLIR-H E1 3 mm reconstruction the highest. However, for the brain examination, most of the images reconstructed using DLIR were rated similar to the images reconstructed using ASIR-V 50%.

The phantom studies showed that there were no significant differences for the spatial resolution, visibility of low contrast objects or linearity of HU between the different reconstructions. However, it was shown that the noise magnitude was lower in the images reconstructed using DLIR compared to the images reconstructed using ASIR-V.

The overall result showed that there might be a possibility to reduce the radiation dose for abdominal examinations when reconstructing the images using DLIR instead of ASIR-V.

Sammanfattning

Datortomografi (DT) är en vanlig medicinsk bildtagningsmetod idag. Det finns olika sätt att rekonstruera bilderna, en standardmetod är filtrerad bakåt projektion (FBP). En nackdel med FBP är dock behovet av en hög exponering, och därmed en hög stråldos till patienten, för att minska bruset i den rekonstruerade bilden. Därför har iterativa rekonstruktionsmetoder (IR) utvecklats som använder statistiska datortomografimodeller för att uppskatta och rekonstruera en bild. Denna metod minskar stråldosen till patienten men ändrar bildstrukturen. Deep learning image reconstruction (DLIR) är en ny rekonstruktionsmetod baserad på deep neural networks (DNN). Det har visats att bilder som rekonstrueras med DLIR har samma bildstruktur som de bilder som rekonstruerats med FBP, samtidigt som brusnivån i bilden sänks.

På Sahlgrenska Universitetssjukhuset finns tre CT-system (GE Revolution Apex™ CT, GE Healthcare, Milwaukee, USA) utrustade med True Fidelity som är en DLIR-mjukvara. Denna studie syftar i att utvärdera bildkvaliteten för DLIR med olika filter jämfört med ASIR-V för buk- och hjärnundersökningar.

Bilder från 20 bukundersökningar och 20 hjärnundersökningar samlades in retrospektivt och rekonstruerades. De rekonstruktioner som gjordes för bukundersökningarna var; Stnd ASIR-V 40% 3 mm, Stnd DLIR-M + E1 3 mm, Stnd DLIR-M + E1 0,625 mm, Stnd DLIR-H + E1 3 mm och Stnd DLIR-H + E1 0,625 mm. Rekonstruktionerna som gjordes för hjärnundersökningen var; Soft ASIR-V 50% EC1 3 mm, Stnd DLIR-M EC1 3 mm, Stnd DLIR-M EC2 3 mm, Stnd DLIR-H EC1 3 mm och Stnd DLIR-H EC2 3 mm. Rekonstruktionerna användes i visual grading characteristics (VGC) studie där rekonstruktionerna betygsattes enligt fem bildkvalitetskriterier. Tre bukradiologer och tre neuroradiologer deltog i studien.

Tre fantomstudier utfördes. Ett Catphanfantom användes för att utvärdera om den spatiella upplösningen, synligheten av objekt med låg kontrast och linjäriteten hos HU påverkades av rekonstruktionerna. Ett antropomorft fantom användes för att se effekterna på bildkvaliteten när stråldosen sänktes. Slutligen användes ett vattenfantomför att studera brusegenskaperna mellan de olika rekonstruktionerna.

Resultatet från VGC-studien för bukundersökningen visade att bilderna som rekonstruerats med DLIR bedömdes signifikant högre än bilderna som rekonstruerats med ASIR-V 40%. Bukradiologerna bedömde DLIR-H E1 3 mm högst av alla rekonstruktioner. För hjärnundersökningen bedömdes de flesta av DLIR-rekonstruktionerna liknande som ASIR-V 50%.

Fantomstudierna visade att det inte var någon signifikant skillnad i spatiell upplösning, synbarhet av lågkontrast objekt och linjäritet av HU mellan de olika rekonstruktionerna. Däremot var magnituden av brus lägre i bilderna som rekonstruerats med DLIR jämfört med bilderna som rekonstruerats med ASIR-V.

Resultatet visade att det kan finnas en möjlighet att sänka stråldosen för bukundersökningar genom att rekonstruera bilderna med DLIR istället för ASIR-V.

Table of contents

Introduction

Today, computed tomography (CT) is a diagnostic imaging method that is useful for many different types of examinations. It is one of the most common imaging methods, among conventional X-ray, MRI and ultrasound. The advantage of a CT examination is that the imaging is very fast, usually a few minutes, which makes it optimal in situations that require answers quickly. A disadvantage is the high radiation dose to the patient. For example, a CT-Thorax examination gives an effective dose of 6.1 mSv, which should be compared to the effective dose of 0.1 mSv from a conventional chest X-ray examination [1].

Following the ALARA (As Low As Reasonably Achievable) principle [2], one concern in CT-imaging is to be able to keep the radiation doses low while still maintaining an image quality that is high enough for diagnosis. Usually, a lower radiation dose result in a poorer image quality. When reconstructing the CT images using filtered back projection (FBP), a lower radiation dose results in images containing a higher level of noise. To improve the image quality, iterative reconstruction (IR) methods have shown to be effective. In IR reconstructions, the algorithms are trained to recognize signal from noise and thereby suppresses the noise in the image. This means that a lower dose can be used for the CT-scan while still maintaining the image quality [3].

A phantom study from 2018, compared adaptive statistical iterative reconstruction V (ASIR-V) to filtered back projection (FBP). ASIR-V (GE Healthcare) combines IR with FBP to reconstruct images. The reconstruction comes with different strengths where the percentage implies how much of IR is used in the reconstruction, e.g. ASIR-V 40% is a combination reconstruction with 40% IR and 60% FBP [4]. It was found that ASIR-V resulted in a lower amount of noise in the images compared to FBP. Also, the higher the strength of ASIR-V, the lower the image noise. For ASIR-V 100%, the noise was reduced by 71.5% compared to the FBP images. The spatial resolution for high-contrast objects was found to be similar for both reconstruction methods FBP and ASIR-V, respectively. However, for low-contrast objects, ASIR-V showed to increase the visibility [5].

The issue with IR is that the resulting images sometimes are perceived as “plastic looking” since the image texture is altered. This can affect the diagnostic value of the images as the images may be harder to interpret [4]. A new reconstruction method called True Fidelity™ was thereby introduced by GE Healthcare. The method is a deep learning image reconstruction (DLIR). There are three strengths of DLIR; low, medium and high where the highest strength has the most reduction of noise [6]. In a phantom study from 2021, images reconstructed DLIR were compared to ASIR-V. The noise accepting level was set for 7.6 mGy for ASIR-V 50% and corresponding noise levels was calculated for different strengths of DLIR. A higher strength of DLIR enabled a higher level of reduction of radiation dose. By using the highest setting of DLIR instead of ASIR-V 50%, the radiation dose could be reduced by 48% without an increased noise level in the images [4].

In a third phantom study, the potential of dose reduction in abdominal examinations using DLIR was assessed. It was found that, compared to ASIR-V, DLIR enhanced the spatial resolution of low contrast objects in the images, especially at lower radiation doses [7]. Thus, there is an opportunity to lower the radiation dose and increase the detectability of low-contrast objects if DLIR is used for abdominal examinations.

Three CT systems (GE Revolution Apex™ CT, GE Healthcare, Milwaukee, USA) have been installed at Sahlgrenska University Hospital. These CT systems are all equipped with a new DLIR software called True Fidelity™. The software is currently used in the clinic for abdominal examinations but has not yet been thoroughly evaluated. Therefore, the optimal settings of True Fidelity™ might not be used. There is also a desire to start using True Fidelity™ for brain examinations. True Fidelity™ therefore needs to be evaluated for brain examinations as well as which strength that is appropriate to use.

Aim of the study

The overall aim of the study was to evaluate image quality for DLIR with different post-processing filters compared to ASIR-V for abdominal and brain examinations. The image quality was evaluated both subjectively, using a VGC study on clinical patient images, and objectively, using phantom studies. An additional aim was to evaluate the possibilities of dose reduction using DLIR.

Background

Computed tomography

Computed tomography (CT) was first introduced in 1971. The first generation of CT was a pencil-beam X-ray which was mounted opposite of a detector and scanned across the patient, then it rotated 1° and scanned the patient again, see Figure 1 (left). This was repeated until a view of 180° was collected [8]. Initially, the CT-scan was used for imaging of the brain. The first-generation CT was sensitive to patient motions, due to the long acquisition time of five minutes per examination. To reduce the scanning time, the second-generation CT used multiple pencil beams, positioned 1° from each other. For each acquisition, projections from multiple angles were collected which reduced the scan-time to under 20 seconds. A shorter scan-time reduced the motion artefacts which made it possible to use the CT for whole-body scanning [9].

The third generation CT, which is used today, consists of detector elements that are arranged in an arc opposite of the x-ray tube, inside the gantry. The patient is scanned with fan-beam projections, see Figure 1 (right). This has reduced the rotation time to under one second and thereby made it useful for many different examinations [9].

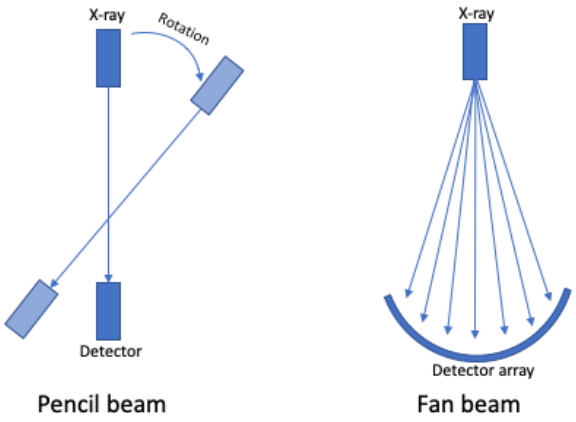

Figure 1: A pencil beam (left), consisting of one x-ray tube and one detector and a fan beam (right) with an array of detectors.

Projections

When the beam passes through the patient, the signal is detected and registered as a line integral. The line integral describes the change of intensity in the beam as it passes through the different tissues in the body. A set of parallel, uniform line integrals creates a projection, see Figure 2 [9].

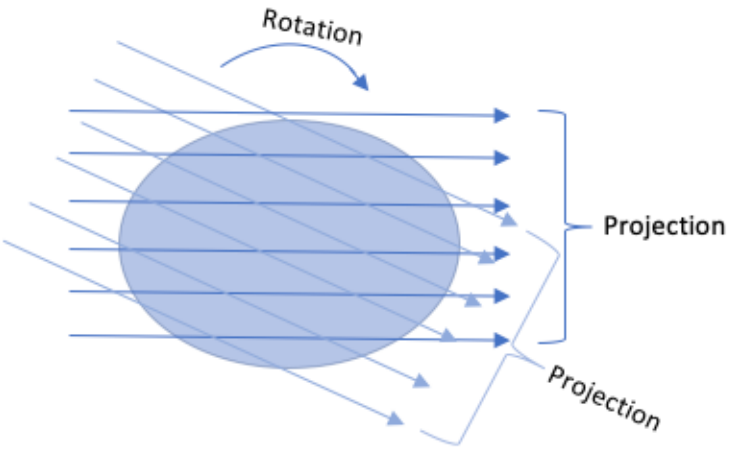

Figure 2: A set of projections, where one projections is represented by parallel uniform line integrals across the patient.

Dose measurements

The most common dose measurement in CT is $CTDI_{100}$ (Computed tomography dose index 100) [mGy] which shows the absorbed dose in a polymethyl methacrylate phantom [9]. $CTDI_{100}$ is determined by integrating the dose D(z) over 100 mm, see Equation 1, where n is the rows of detectors used for the scan and T is the thickness of each row.

$$\mathrm{CTDI}_{100} = \frac{1}{nT}\int_{-50}^{50} D(z)dz \qquad [\mathrm{mGy}] \qquad (1)$$

To take into account for the variation of the dose given from a CT-examination, a weighted dose measurement $CTDI_w$ (Computed tomography dose index weighted) [mGy] is used. The weighted dose combines $CTDI_{100}$ from the central part of the phantom ($CTDI_{100,c}$) and the peripheral part of the phantom ($CTDI_{100,p}$), see Equation 2.

$$\mathrm{CTDI}_w = \frac{1}{3} CTDI_{100,c} + \frac{2}{3} CTDI_{100,p} \qquad [\mathrm{mGy}] \qquad (2)$$

The mean radiation dose in an examined volume is presented as $\mathrm{CTDI}_{\mathrm{vol}}$ (Computed tomography dose index volume) [mGy] and shows the mean radiation dose for a specific scan protocol. The $\mathrm{CTDI}_{\mathrm{vol}}$ takes the pitch into account when measuring the mean radiation dose, see Equation 3.

$$\mathrm{CTDI}_{\mathrm{vol}} = \frac{\mathrm{CTDI}_{\mathrm{w}}}{\mathrm{pitch}} \qquad [\mathrm{mGy}] \qquad (3)$$

The integrated dose for the entire scan length *L* is given by the dose-length product (DLP), see Equation 4 [9].

$$\mathrm{DLP} = \mathrm{CTDI}_{vol} \cdot L \qquad [\mathrm{mGy*cm}] \qquad (4)$$

Hounsfield Unit

The Hounsfield unit (HU), or CT-number, is a unit used in computed tomography to describe the linear attenuation of the tissue. The unit is named after Godfrey Hounsfield, one of the inventors of the CT technology. Different tissues have different linear attenuation, however these differences can be very small. To enhance the difference in attenuation between tissues, HUs are used. The HU for different tissues is based on the linear attenuation coefficient of water μ_{water}, which is defined as 0, see equation 5 [9], and the linear attenuation coefficient of the specific tissue μ_{tissue}.

$$\mathrm{HU} = \frac{\mu_{tissue} - \mu_{water}}{\mu_{water}} \cdot 1000 \qquad (5)$$

Reconstruction methods

Filtered Back Projection

Filtered back projection (FBP) has long been the standard method of reconstruction in CT. The method has the advantage of being fast and robust. In back projection, the signal registered for one gantry angle (1) is projected back in the opposite direction of the beam (2), see Figure 3. All the back projections for a gantry rotation are added to one another to reconstruct the image [3].

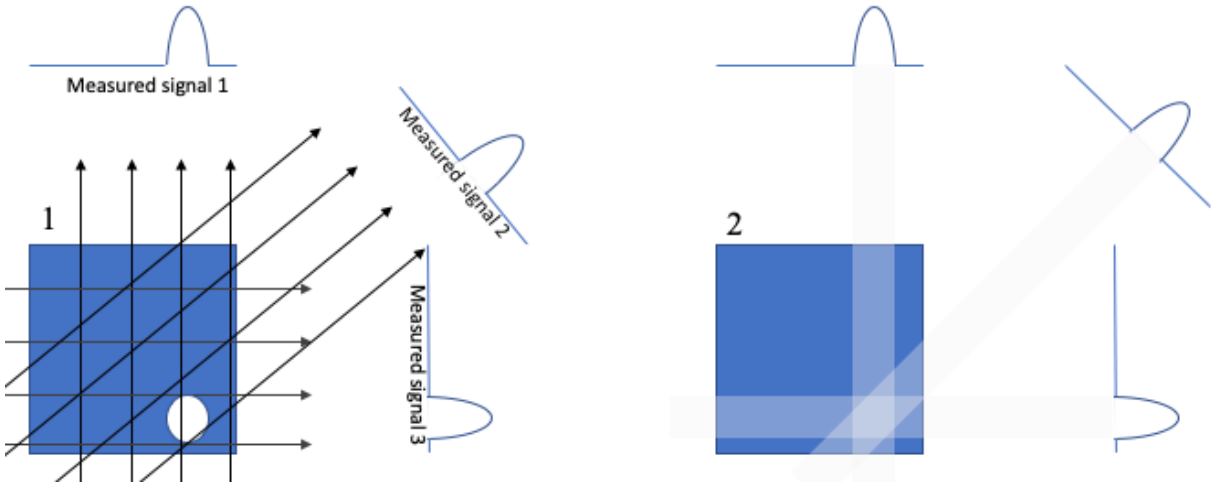

Figure 3: Signal is registered from different angles (1), then the signal is back projected to reconstruct the image (2).

The reconstructed image somewhat represents the object but with a blurring effect, see Figure 4. To get rid of the effect, each projection is Fourier transformed. In the Fourier space, each projection is multiplied with a ramp filter. The ramp filter is a high-pass filter that reduces the low frequencies causing the blurring effect. Then, the projections are inversed Fourier transformed, and an integration is made across all projections to reconstruct the FBP image [3].

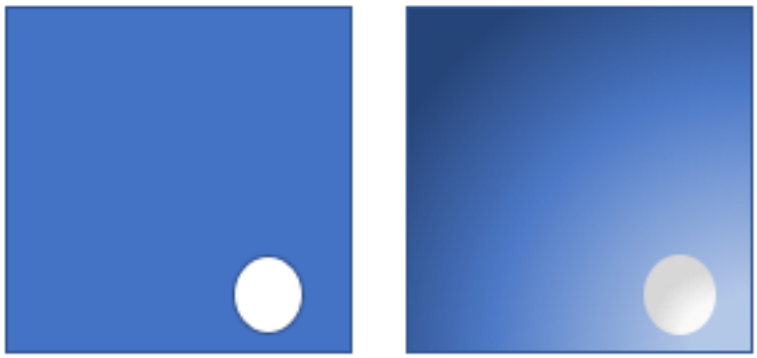

Figure 4: The original image (left) and the reconstructed image (right) with a blurring effect.

Iterative reconstruction method

A disadvantage of FBP is that the high-pass filter enhances the noise in the image. In order to reduce the noise in the FBP-images, a higher exposure level is needed, leading to higher patient radiation doses. To reduce the radiation dose, a different reconstruction method based on the use of iterative algorithms has been developed.

An iterative reconstruction (IR) method is signified by three steps; 1. the input, which is the measured projections and the initial estimated projections, 2. the iterative algorithm, and 3. the output, which is the reconstructed image, see Figure 5. The initial estimated projections are based on statistical CT imaging models. The iterative algorithm consists of a loop where the input first is compared to the initial estimated projections and will be processed until it meets predefined conditions. There are several variants of IR, and depending on the algorithm used, the conditions are based on various images [10].

In Figure 5, the iterative reconstruction method is illustrated. First, the measured projections are compared with the estimated projections from the initial estimated image, see step 1 in Figure 5. The difference between the measured projections and the estimated projections is used as a weight to correct and optimize the estimate. The weight works as an adjustable parameter which will alter the estimated projections [10]. The estimated projections are updated in the IR loop, step 2 in Figure 5, until it meets predefined conditions. When the conditions are fulfilled, an output image is reconstructed, from the updated projections, see step 3 in Figure 5.

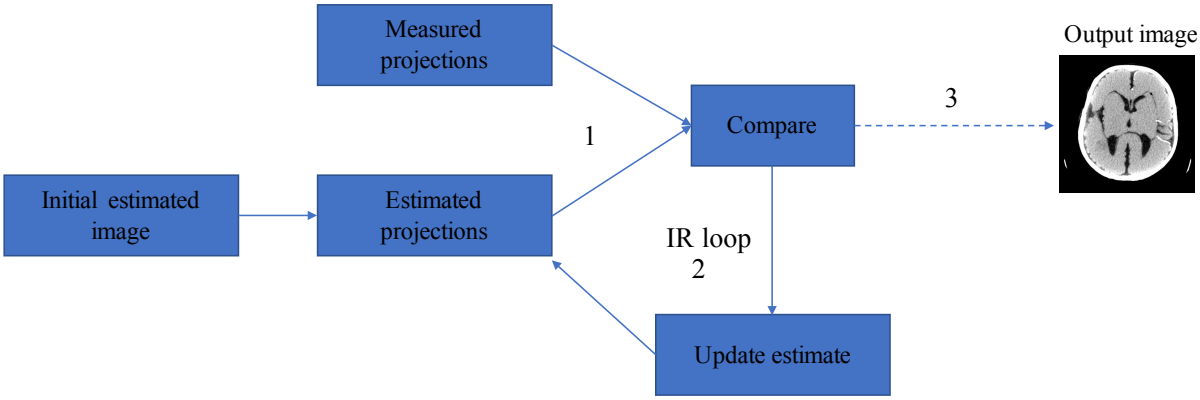

Figure 5: Schematic description of an iterative reconstruction algorithm.

Deep Learning Image Reconstruction

According to the White Paper [6], the Deep learning image reconstruction (DLIR) is an algorithm based on a deep neural network (DNN). A DNN consists of multiple layers of equations, so called artificial neurons, which are represented as blue circles in Figure 6. The artificial neurons are connected to each other, in similarity to the brain's cortex, where one neuron is connected to several other neurons via synapses. The network consists of an input, layers of neurons, and an output. The layers of neurons are called the hidden layers, where the output from one neuron in a layer is the input for all neurons in the next layer, see Figure 6. A network with at least two hidden layers is called a deep neural network DNN [11].

The DNN works by using the measured projection data from the image acquisition to form an input vector (1), see Figure 6, to the network. The input vector is passed through the neural network where it is transformed to an output vector (2). The output vector is then used to create the reconstructed image.

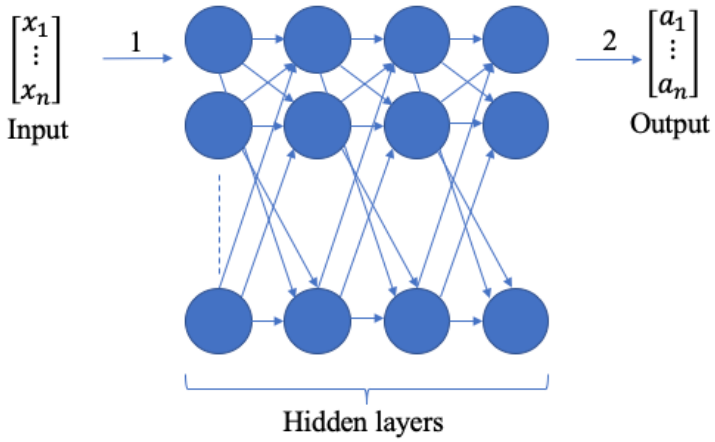

Figure 6: Schematic description of a deep neural network.

During the transformation from an input vector to an output vector, data are passed through neural layers (l) which is illustrated in Figure 7. Between two neurons there is an adjustable weight ω and connected to each neuron n there is a bias b, see Figure 7. The input $z(l)$ for one neuron in layer l is the weighted $\boldsymbol{\omega}(l)$ sum of the outputs $\boldsymbol{a}(l\text{-}1)$ from neurons in the previous layer (l-1) as well as the bias b_n for that specific neuron, see equation 2. In each neuron, there is an activation function $h(z)$ which transforms the input $z(l)$ to an output $a(l)$, see equation 3. The activation function is specific for each neuron [11].

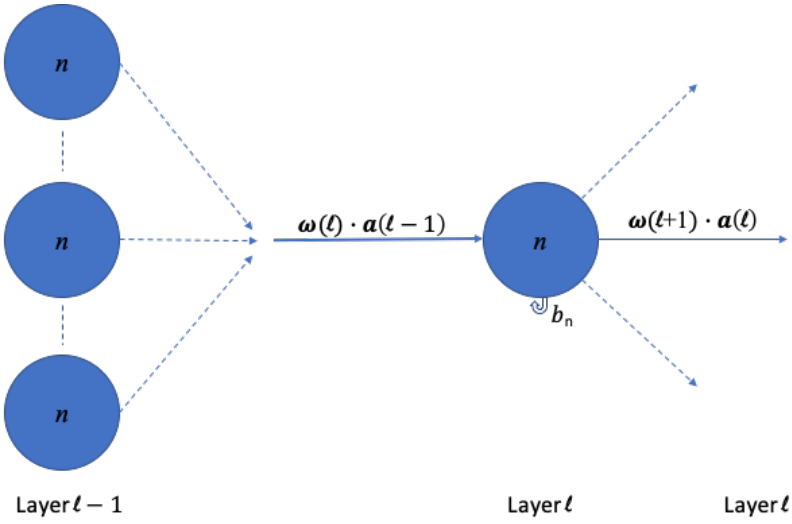

Figure 7: The input for the neurons in layer l is the sum of outputs from previous layers a(l-1) (left), weights ω and biases b_n.

$$z(l) = \boldsymbol{\omega}(l) \cdot \boldsymbol{a}(l-1) + b_n \tag{6}$$

$$a(l) = h(z(l)) \tag{7}$$

Before the network is ready to be used, it is trained by the manufacturer. The purpose of training the neural network is to optimize the parameters for the neurons, the weights and the bias. The network is trained using data from high- and low dose scans of the same objects. The data from a low dose scan is fed through the network and an output is given. The output vector is compared to the corresponding output vector from the high dose scan (the known

truth), and the difference between the output and the known truth is calculated. The difference is called the error function of the system. By optimizing the weights and biases, the error function is minimized. Since the truth is known, the output vector is fed backwards through the network to determine the desired values for the weights and the biases. The training is done when the output for any image acquisition is equal, or at an accepting level, to the known truth [11].

Convolutional neural network

Even though the White Paper uses the terminology DNN, a convolutional neural network (CNN) is commonly used when images are input data as this type of network maintains the spatial information. [12]

A CNN used for image processing consists of convolutional layers and pooling layers. The convolutional layers are used to extract components from the images while the pooling layers reduce the images. In the convolutional layer, a kernel consisting of different values, are applied to the image and multiplied with the image value for each pixel to create a feature map. A pooling layer will then reduce the feature map, for example, a max pooling layer will extract the maximum value of the pixels and remove the other values in order to extract detailed information from the images. The extracted information is then used as input to the next convolutional layer [12]. During training of the CNN, the kernel values in the convolutional layers are adjusted to obtain the desired result.

It is common to use a U-net in the neural network. The U-net has encoder and decoder units which will down- and up sample the image. The encoder unit consist of convolutional layers and an activation function that will down sample the image. The decoder also consists of convolutional layers, but transposed, and an activation function that will up sample the image [13].

Image quality

Filters and image enhancements

The purpose of using filters in medical imaging is to alter the image properties in different ways in order to facilitate the possibilities to identify anatomy and pathology in the images. Most often, the filters are adapted to specific examinations. For example, the filters used in brain examinations are not optimized to be used in lung examinations. Filters works in two domains, the spatial domain and the frequency domain [11].

There are two types of filters in the spatial domain, the linear and non-linear filters. Both types are based on the gray values for a group of pixels. Neighboring pixels tend to have similar intensity. There is also an occasional “intensity-jump” which comes from image noise. To reduce the noise in an image, a spatial filter can be applied which will use the neighboring pixel intensities to replace the pixel value, and thereby remove the higher intensity in a pixel that is due to noise. These types of filters are called smoothing filters, where the intensity value for one pixel is affected by the surrounding pixel intensities. The opposite of the smoothing filters are the sharpening filters. The purpose of these filters is to enhance the edges, with the result that also the noise in images is enhanced [11].

In the frequency domain, the filter is added to the image by a convolution. The filter works by suppressing low or high frequencies in the Fourier space. An image consists of different

frequencies. The low frequencies contain most of the information in the image and the higher frequencies represents the details, edges and noise. A low pass filter suppresses the higher frequencies in an image which makes the image appear more "smooth" and reduces the noise. The opposite is a high pass filter which filters out the low frequencies and reinforces edges, details and noise [11].

In a CT-scan, there are so-called reconstruction kernels that can be selected depending on which examination is to be performed. These kernels are frequency filters that are applied before the back projections are made, with the purpose of removing the unwanted frequencies from the image. For example, *Bone* (GE Healthcare) is an edge-enhancing kernel which reduces the low frequencies in the image and makes the image optimal for reviewing bone structures [14].

Noise power spectrum

The noise power spectrum (NPS) shows the frequencies of the noise in an image. The noise is Fourier transformed and plotted against the spatial frequencies. The purpose of analysing the NPS of an image is to find the magnitude and variance of the noise [15].

A problem with the NPS is that it is based on limited data. Thereby, uncertainties are introduced to the spectrum. The resolution of the frequencies depends on the size of the region of interest (ROI) used to make the measurements. A larger ROI gives a better resolution. However, by making the ROI larger, the uncertainty increases in finding a specific frequency [16].

Visual grading characteristics and statistical analysis

Visual Grading Characteristics

Visual grading is a study where the observer rates the visibility of structures in the images. The images are assessed according to specific image quality criteria and the observer grades how well the criteria are fulfilled using a multi-alternative grading scale. Visual grading characteristics (VGC) is a method used to analyse data from a visual grading study by comparing ratings from two conditions [17].

The conditions are usually one reference condition, for example what is currently used in the clinic, while the other condition is a test condition. The observer ratings for each condition are used to create a VGC-curve, where the area under the curve (AUC) shows which condition is preferred. If AUC is larger than 0.5 the test condition is preferred, while if AUC is less than 0.5, the reference condition is preferred. When AUC is equal to 0.5, the conditions are rated equally [18].

When multiple observer's ratings are used to plot the VGC-curve, the final AUC is determined by averaging each observers AUC [18]. Since the observers have interpreted the step-scale in different ways and judged the images based on that, the ratings will be very individual. However, this does not affect the AUC. The VGC-curve shows the proportion of the observer's ratings and is therefore independent of the interpretation of the grading scale [19].

The software VGC Analyzer can be used to make a statistical analysis of the VGC data [20]. The ratings, given by the observers, are used as an input to generate the VGC-curve. VGC-

analyzer also provides the AUC, the p-value and the confidence interval (CI) for the comparison.

VGC analyzer uses the bootstrapping technique to calculate the CI for the AUC [18]. The bootstrapping technique picks random data from the original set of ratings and creates a new set of ratings. The number of rating sets that can be created is n^n, where n is the total number of data in the original set of ratings. Another resampling technique used in VGC analyzer is the permutation resampling technique which is used to find the p-value. The p-value describes how probable an outcome is to occur. The technique combines the data from both conditions and selects random ratings from the combination. The selected ratings are used to create two new sets of data for two "fake" conditions and the AUC for the conditions is determined. This is repeated many times to create a distribution of the AUC. The p-value is found based on where the AUC for the original data set is placed in the distribution [20].

The AUC can be calculated using either a binormal fit or the trapezoidal rule. The selection should be made based on the number of cases included in the comparison. For larger numbers of cases, both methods work equally in finding the AUC. When the number of cases is smaller, the binormal fit was found to overestimate the CI while the trapezoidal rule underestimated the CI. The binormal curve also had a larger deviation in the AUC compared to the trapezoidal rule. Therefore, when the number of cases is small (i.e. fewer than 20), the trapezoidal rule is preferable when calculating the AUC [18]. The statistical analysis can be performed both for a fixed reader situation or a random reader situation. In the fixed reader situation, each resampled data set contain data from all original observers. In the random reader situation a bootstrapping of observers is also applied [20].

ViewDEX

Viewer for Digital Evaluation of X-ray Images (ViewDEX) is a software used for viewing images and performing observer studies. ViewDEX is compatible with DICOM images from multiple modality sources. The images and/or image series are shown in a unique random order for each observer, without any image information present. The observers will review each image and/or image series and perform different tasks according to the outline of the study. For example, the tasks can be related to answering questions regarding image quality or marking suspicious pathology in the images. The results from image review are saved in a unique log file for each observer [21].

ViewDEX include different functionality that might be useful for the observers during image review. For example, the observers can zoom, pan the images, and change the windows settings (window width and window level). It is also possible to do measurements in the images, like distance, area and pixel value. Another feature is the ability to write notes about the images during the review. In ViewDEX 3.0, the layout can consist of up to four canvases on up to four different monitors. This makes it possible to display, for example, different image reformations (coronal, sagittal and axial) to the observer simultaneously [21].

The set-up of a study in ViewDEX is made by altering a study property file. A property file is a simple text file in which the properties of each study is defined. The observers enter the study by signing in with unique usernames that is assigned to them in a user property file. Before starting the review of the actual study, the observers have the option to perform a demo study in order to get familiar with the ViewDEX 3.0 software and the study setup [21].

Method

Revolution Apex CT

Three GE Revolution Apex™ CT from GE Healthcare (Milwaukee, USA) installed at the Sahlgrenska University Hospital were used in the study. The X-ray tube used in the system is a Quantix™ 160 tube with a tube voltage range of 70 – 140 kV. The tube current has a range of 10 – 1 300 mA. By using dose modulation, the tube current will be adapted to the patient size to achieve an optimized image quality and patient radiation dose. The detector is a Gemstone Clarity Detector, a scintillator detector, with a coverage of 160 mm [22].

True Fidelity

The GE Revolution Apex CT systems are all equipped with GE's new reconstruction software True Fidelity™. True Fidelity™ is an AI-based software which reconstructs images with deep learning image reconstruction (DLIR). GE has trained the network with high dose filtered back projection (FBP) images. Three strengths of DLIR; low, medium and high can be used, where the highest strength has the most reduction of noise [6].

Scan protocol and reconstructions

The kernels used in the acquisitions depends on the type of examination. In GE Revolution Apex there are multiple different kernels available. At Sahlgrenska University Hospital, a *Soft* kernel is used for brain imaging, while for abdominal imaging a *standard (Stnd)* kernel is used. The *Soft* kernel is primarily used for soft tissue while the *Stnd* kernel is preferred for abdominal or thorax examinations [23]. The only kernel applicable when reconstructing with DLIR is the *Stnd* kernel. However, in the future DLIR might also be able to run with other kernels.

Reconstructions to be evaluated

The reconstruction setting evaluated in the present study were selected by one neuro radiologist and one abdominal radiologist. Both radiologists are specialists in each field and have several years of experience reviewing images. The selection included the clinically used ASIR-V and different settings of DLIR. The different reconstructions were evaluated both in a patient study and in phantom studies.

For the abdominal examination, a total of five different reconstructions were evaluated. At Sahlgrenska University Hospital, abdominal examinations are usually reviewed using a slice thickness of 3 mm. However, as DLIR can be expected to lower the image noise, it might be possible to reduce the slice thickness from 3 mm to 0.625 mm if DLIR is used instead of ASIR-V when reconstructing the images. Therefore, the present study included the evaluation of both 3 mm and 0.625 mm slice thickness for the images reconstructed using DLIR. The reconstructions evaluated in the present study were;

- *Stnd* ASIR-V 40% 3 mm,
- *Stnd* DLIR-M + E1 3 mm,
- *Stnd* DLIR-M + E1 0.625 mm,
- *Stnd* DLIR-H + E1 3 mm,
- *Stnd* DLIR-H + E1 0.625 mm.

For the brain examination, a total of five different reconstructions were evaluated. These were;

- *Soft* ASIR-V 50%,
- *Stnd* DLIR-M + EC1 3 mm,
- *Stnd* DLIR-M + EC2 3 mm,
- *Stnd* DLIR-H + EC1 3 mm,
- *Stnd* DLIR-H + EC2 3 mm.

Both axial and coronal slices were reconstructed for the brain examination. No thinner slices than 3 mm were evaluated since a slice thickness of 3 mm are used for review of brain examinations at Sahlgrenska University hospital and due to the increase of noise in thinner slices, it is not expected that the slice thickness used for image review will be changed in the near future.

For the brain exam, the enhanced contrast filter 1 and 2 were chosen (EC1/EC2). The purpose of EC is to enhance the difference between white and grey matter based on the Hounsfield unit (HU). It comes with three levels of enhancement, EC1, EC2 and EC3. EC3 has the highest differentiation while EC1 has the lowest [24].

For the abdominal examination, an edge enhancement filter E1 was chosen. Image edge enhancement filter (E1, E2 and E3) are used to enhance the structure of the anatomy and to make the image sharper. E1 has the least amount of edge sharpening while E3 has the most [24].

Clinical protocol

The clinical scan protocol for abdominal and brain examinations using the GE Revolution Apex™ CT at Sahlgrenska University Hospital can be seen in Table 1.

The radiation dose to the patient is directly related to the mAs-value, a higher mAs-value results in a higher radiation dose to the patient. By using the dose modulation, smartmA (GE Healthcare) the mAs-value will be adapted to patient size to optimize the exposure over different parts of the patient's body. Noise index (NI) is a GE specific parameter that determines the amount of noise in the image. The higher the NI, the more noise is present in the image. Using smartmA, the mAs-value will be modulated to achieve the desired noise level given by the NI over the entire examined volume [25]

Table 1: Clinical scan protocol for the abdominal and the brain examinations.

	Abdominal	Brain
Scan type	Helical	Helical
kV	120	120
mAs	smartmA (80-650)	smartmA (100-340)
NI	14	3.3
Pitch	0.992	0.516
Collimation (mm)	80	40
Slice width (mm)	3	3

Visual grading characteristics study

Ethical approvement

The use of patient images in this study was approved by Sahlgrenska University Hospital. The images were pseudonymized before they were extracted from PACS (picture archiving and communication system).

Collecting Image Material

The patient images were collected retrospectively from one of the GE Healthcare Revolution Apex™ CT systems at Sahlgrenska University Hospital. All examinations were pseudonymized and patient size and age was noted. The examinations included in the study were chosen based on a desire to cover the clinical range of patients. Therefore, patients of different sizes (abdominal examinations) and ages (brain examinations) were included in the study.

Abdomen

Examinations from 20 patients (13 women and 7 men) were included in the study see Table 2. The patients were divided into two groups based on body mass index (BMI), one group containing patients with BMI < 25 and one group of patients with BMI ≥ 25. The size of the patients was also estimated by measuring the lateral (LAT) and anterio-posterior (AP) diameters, see Figure 8. The diameters were measured in the CT scout images, in the level above the iliac crest in line with the L4 lumbar vertebrae, see Figure 8.

Table 2: Patient data from the collected material from the abdominal examination. The mean values of the age, $CTDI_{vol}$, weight, length, BMI, AP and LAT is shown and the range is shown inside the parenthesis.

	Smaller patient BMI < 25	Larger patient BMI ≥ 25
Number of patients	11	9
Male/female	7/4	0/9
Age	60	68
$CTDI_{vol}$ (mGy)	6.3 (4.9-8.2)	9.9 (5.2-21.7)
Weight (kg)	66 (50-80)	76 (57-110)
Length (cm)	171 (146-182)	158 (152-166)
BMI	22 (20-24)	30 (25-44)
AP (mm)	232 (213-251)	298 (191-387)
Lat (mm)	343 (300-376)	388 (295-505)

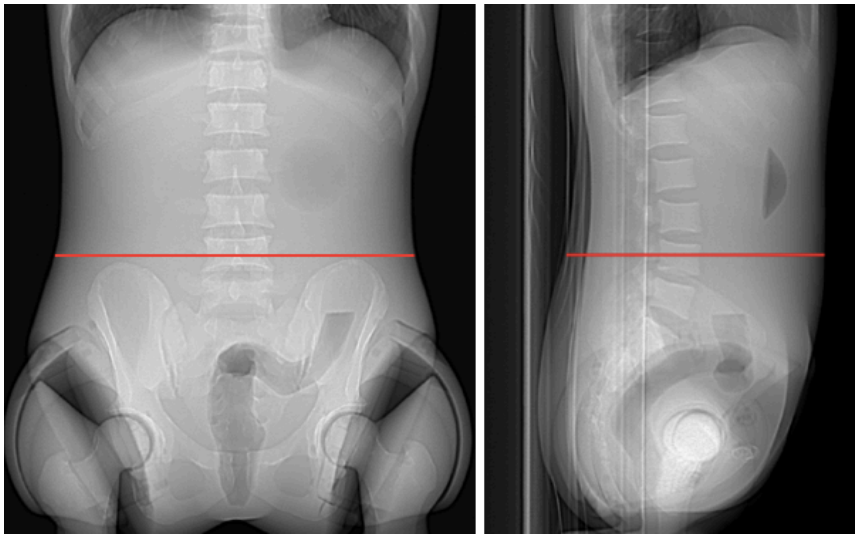

Figure 8: Scout of anthropomorphic phantom showing how the LAT (left) and AP (right) were measured above the iliac crest at L4, represented by the red line in the scout.

Brain

Images from 20 patients (11 women and 9 men) were included in the study. The patients were divided into three groups according to age as the anatomy in the brain changes due to ageing, e.g. for older patients the cerebrospinal fluid is easier to distinguish due to atrophy of the cerebral cortex. One group consisted of older patients (born before 1940) and one group consisted of younger patients (born after 1970), see Table 3. A third group of four patients who were born after 1940 and before 1970 were also included in the study, see Table 3.

Table 3: Patient data from the collected material from the brain examination. The mean values of the age and $CTDI_{vol}$ are shown and the range is shown inside the parenthesis.

	Born before 1940	Born after 1970	Born between 1940 and 1970
Number of patients	9	7	4
Male/female	3/6	5/2	1/3
Age	89 (82-95)	29 (16-42)	63 (52-74)
$CTDI_{vol}$ (mGy)	34.4 (29.7-41.3)	36.7 (31.0-40.2)	38.3 (34.2-41.3)

After image reconstruction, the brain images had to be corrected in the CT software in order to be anatomically aligned. This was done, in the axial plane, by placing the reference axes in a section where the ear canals and the lens of the eye were visible. The axes were then placed symmetrically in the ear canal and between the eyes. For the coronal plane, the brain stem was used as a reference and the axes were placed according to its location.

ViewDEX

Two studies were created in ViewDEX 3.0, one for the abdominal examinations and one for the brain examinations. For the abdominal examinations, only axial reformats were shown to the observers during image review. For the brain examinations, both axial and coronal reformats were shown. The image quality criteria used in the studies were determined together with the experienced neuro- and abdominal- radiologist, respectively. Three radiologist who were specialists within each field reviewed the images and evaluated them according to the

image quality criteria determined for each anatomical region, see “Image quality criteria for abdomen” and “Image quality criteria for brain” below. The image quality criteria were rated according to a five-step grading scale, see “Answer alternatives” below. For illustrations of the structures that were evaluated, see the Appendix under heading “Evaluated anatomy”.

The studies were reviewed on a DICOM-calibrated monitor (EIZO, RadiForceRX320). The review station was set up in a room with low ambient light and the review period was set to six weeks.

Image quality criteria for abdomen

Following criteria were used for evaluating the image quality of the abdomen examinations;

1. Visual reproduction of the liver parenchyma and intrahepatic vessels
2. Visual reproduction of the differentiation of the right adrenal gland from adjacent structures
3. Visual reproduction of the perirenal fat
4. Visual reproduction of the terminal ileum
5. Overall quality of the examination

Image quality criteria for brain

Following criteria were used for evaluating the image quality of the brain examinations;

1. Visual reproduction of the border between white and grey matter
2. Visual reproduction of the basal ganglia
3. Visual reproduction of the cerebrospinal fluid space around the mesencephalon
4. Visual reproduction of the posterior fossa structures
5. Overall quality of the examination

Answer options

Following options were used to answer the image criteria for brain and abdomen.

- Excellent
- Good
- Acceptable
- Poor
- Unacceptable

Visual grading characteristics analysis

Two different VGC analyses were performed. In all the analyses, the ratings from the different observers for each quality criterion and reconstruction method were pooled and compared to the corresponding pooled ratings for ASIR-V, which was used as the reference condition. The analyses focused on evaluating differences in image quality between the reconstructions. Sub-analyses were also performed, focusing on different patient groups. For the brain examination, the image quality was evaluated separately for older and younger patients. For the abdominal examination, the image quality was evaluated separately for patients of different sizes.

The rating data was used as an input to VGC Analyzer, in which the ratings of the test condition (the DLIR images) were compared to the ratings of the reference condition (the ASIR-V images). In the VGC Analyzer software, the calculations of AUC and the 95% CI were based on the fixed reader situation and the trapezoidal rule for curve fitting. If the AUC is larger than 0.5, the DLIR reconstruction is preferred over the ASIR-V reconstruction. If the

CI covers 0.5, there is no statistically significant difference in the ratings of the image quality criteria between the compared reconstructions.

Phantom studies

Catphan phantom study

A Catphan phantom (Catphan 600, *The Phantom Laboratory*) was used in the study. The phantom consists of multiple layers (modules) with different elements used to study the spatial resolution, low contrast sensitivity and the linearity of CT-numbers (HU). There is also a homogenous section which can be used to study the noise. The phantom was mounted on the couch and scanned using the GE Healthcare Revolution Apex™ CT system.

Abdomen

The Catphan phantom was scanned with the clinical protocol for the abdominal examination and the same five reconstructions as evaluated in the patient study were made; *Stnd* ASIR-V 40% 3 mm, *Stnd* DLIR-M + E1 3 mm, *Stnd* DLIR-M + E1 0.625 mm, *Stnd* DLIR-H + E1 3 mm and *Stnd* DLIR-H + E1 0.625 mm.

Brain

The Catphan phantom was scanned with the clinical protocol for the brain examination and the same five reconstructions as evaluated in the patient study were made; *Soft* ASIR-V 50%, *Stnd* DLIR-M + EC1 3 mm, *Stnd* DLIR-M + EC2 3 mm, *Stnd* DLIR-H + EC1 3 mm and *Stnd* DLIR-H + EC2 3 mm.

Image analysis

The reconstructed images were analyzed in the image viewing software *MicroDicom* (v. 3.9.5). The following analysis of the images were made; linearity of the CT-numbers, spatial resolution and low contrast visibility.

To analyze the linearity of the HU values, multiple ROIs were placed in the reconstructed image located in module CTP404 of the phantom, containing multiple materials. This was done both for the images collected using the brain protocol and for the images collected using the abdominal protocol. The ROIs were placed in water, air, polystyrene and Teflon. A ROI was also placed in the homogenous part in the center of the image. The HU values and the SD was measured in each material.

The evaluation of the effect of the different reconstructions on spatial resolution was made using the reconstructed image located in module CTP528 of the phantom, containing different line pairs. The evaluation was done both for the images collected using the brain protocol and for the images collected using the abdominal protocol. The number of visible line pairs were compared between the different reconstructions.

The evaluation of low contrast resolution made using the reconstructed image located in module CTP515 of the phantom, containing low contrast structures. The evaluation was done both for the images collected using the brain protocol and for the images collected using the abdominal protocol (original dose level). The number of visible low contrast objects were compared between the different reconstructions.

Anthropomorphic phantom study

An anthropomorphic body phantom (CT Whole Body Phantom PBU-60 KYOTO KAGAKU) was used. The phantom had extra tissue padding that could be placed around the phantom in order to simulate a larger patient. Without the padding, the phantom represents a patient with a BMI of 19. When adding the padding, the phantom represents patients with BMI 32 and BMI 40. The phantom was scanned using the GE Healthcare Revolution Apex™ CT.

Abdomen

The phantom was placed on the couch and scanned with the clinical protocol, with noise index (NI) 14, for the abdominal examination. This was done for the phantom representing patient sizes of both BMI 19 and BMI 40. The images were reconstructed with *Stnd* DLIR-H + E1 3 mm. Since a previous study has shown that it might be possible to lower the radiation dose when DLIR-H is used [7], the effect of dose reduction on image quality was investigated. Therefore, the phantom was scanned again, reducing the dose by approximately 50%, by increasing the NI to 18.

Water phantom study

A homogenous water phantom (LightSpeed phantom, GE Healthcare) was scanned using the GE Healthcare Revolution Apex™ CT system. The phantom was mounted on the couch and scanned with the clinical brain protocol and the abdominal protocols (original dose level, i.e. NI 14, and 50% dose level, i.e. NI 18). The same reconstructions as evaluated in the patient study were made for each examination. However, as the edge enhancement filters might lead to an enhancement of image noise, the standard deviation and noise power spectrum (NPS) were determined in images reconstructed both with and without the spatial post-processing filter E1.

Standard deviation and noise power spectrum

Fifty consecutive image slices from every reconstruction series were used for the evaluation of standard deviation and noise power spectrum (NPS). In each of the 50 image slices an ROI of 128 x 128 pixels was placed in the middle of the phantom. The standard deviation in each ROI was determined and the resulting standard deviation for each reconstruction was given by the average standard deviation in the 50 image slices. The 2D NPS was obtained by the discrete Fourier transform of the ROI in each image slice and then averaging the 2D NPS over the 50 image slices. The 1D NPS was obtained as the radial average of the 2D NPS. This was done by sorting the 2D NPS into frequency bins, and then averaging the value for each bin. [26]

Results

DLIR-reconstructions

The visual difference between the evaluated reconstructions can be seen in Figure 9 and in Figure 10. For both the abdomen and brain examinations, the ASIR-V reconstructions were observed as noisier than DLIR-M and DLIR-H.

Figure 9: Axial 3 mm slices showing the evaluated reconstructions for the abdominal exam, (from left to right); ASIR-V 40%, DLIR-M E1 and DLIR-H E1. The images were acquired using the abdomen protocol.

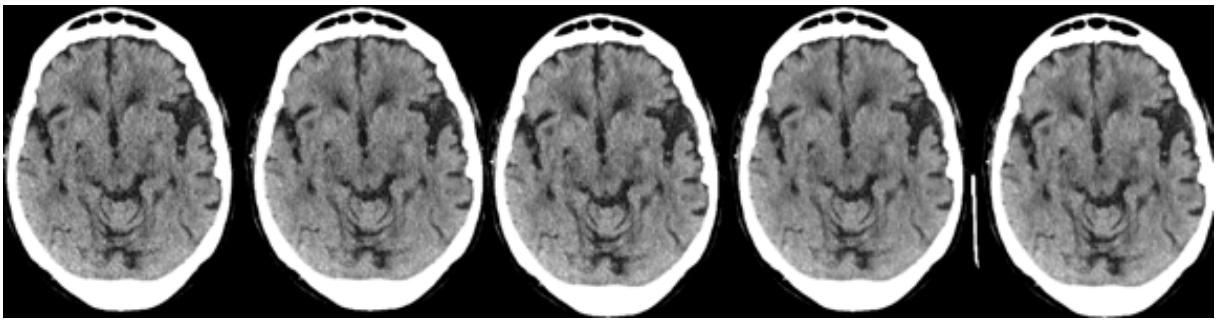

Figure 10: Axial 3 mm slices showing the evaluated reconstructions for the brain exam, (from left to right); ASIR-V 50%, DLIR-M EC1, DLIR-M EC2, DLIR-H EC1 and DLIR-H EC2. The images acquired using the brain protocol.

Visual Grading Characteristics study

Abdomen

Figure 11 shows the area under the curve (AUC) for the different reconstructions evaluated in the present study. The DLIR reconstructions were rated significantly higher (the 95% CI of the AUC was above 0.5) compared to the ASIR-V 40% for every image quality criterion, see Table 4. The highest scores were obtained by the reconstruction DLIR-H E1 3 mm.

The 0.625 mm slices were rated lower than the 3 mm slices for both DLIR-M and DLIR-H for every quality criterion.

When comparing the different criteria against each other, criterion 2 (visual reproduction of the differentiation of the right adrenal gland from adjacent structures) had the highest ratings for all DLIR reconstructions compared to ASIR-V 40%. The lowest ratings for DLIR compared to ASIR-V was obtained for criterion 4 (visual reproduction of the terminal ileum).

The VGC-curve for each criterion can be seen in Appendix under headings “Visual Grading Characteristics, Abdomen”.

When pooling the results from all criteria together, DLIR-H E1 3 mm got the highest ratings of all reconstructions with an AUC of 0.86 while DLIR-M E1 0.625 mm got the lowest ratings, with an AUC of 0.74. However, DLIR-M E1 0.625 mm was still rated higher than ASIR-V 40%, see Figure 12.

When comparing the two groups including patients of different sizes with each other, the DLIR reconstructions obtained higher ratings in the group containing larger patients (BMI ≥ 25) than in the group containing smaller patients (BMI < 25), see Figure 13.

Table 4: The area under the curve for the five image quality criteria for the abdomen for the evaluated reconstructions, where ASIR-V is the reference condition. The 95% CI of the AUC is shown inside the parenthesis.

	DLIR-M 3 mm	DLIR-M 0.625 mm	DLIR-H 3 mm	DLIR-H 0.625 mm
Criterion 1	0.81 (0.75-0.88)	0.74 (0.64-0.83)	0.85 (0.78-0.92)	0.80 (0.73-0.87)
Criterion 2	0.88 (0.83-0.93)	0.83 (0.75-0.89)	0.93 (0.88-0.97)	0.87 (0.81-0.92)
Criterion 3	0.76 (0.69-0.84)	0.71 (0.65-0.78)	0.82 (0.76-0.88)	0.78 (0.72-0.84)
Criterion 4	0.70 (0.62-0.78)	0.67 (0.60-0.74)	0.82 (0.75-0.88)	0.74 (0.66-0.80)
Criterion 5	0.88 (0.83-0.92)	0.75 (0.68-0.82)	0.92 (0.88-0.97)	0.84 (0.80-0.88)

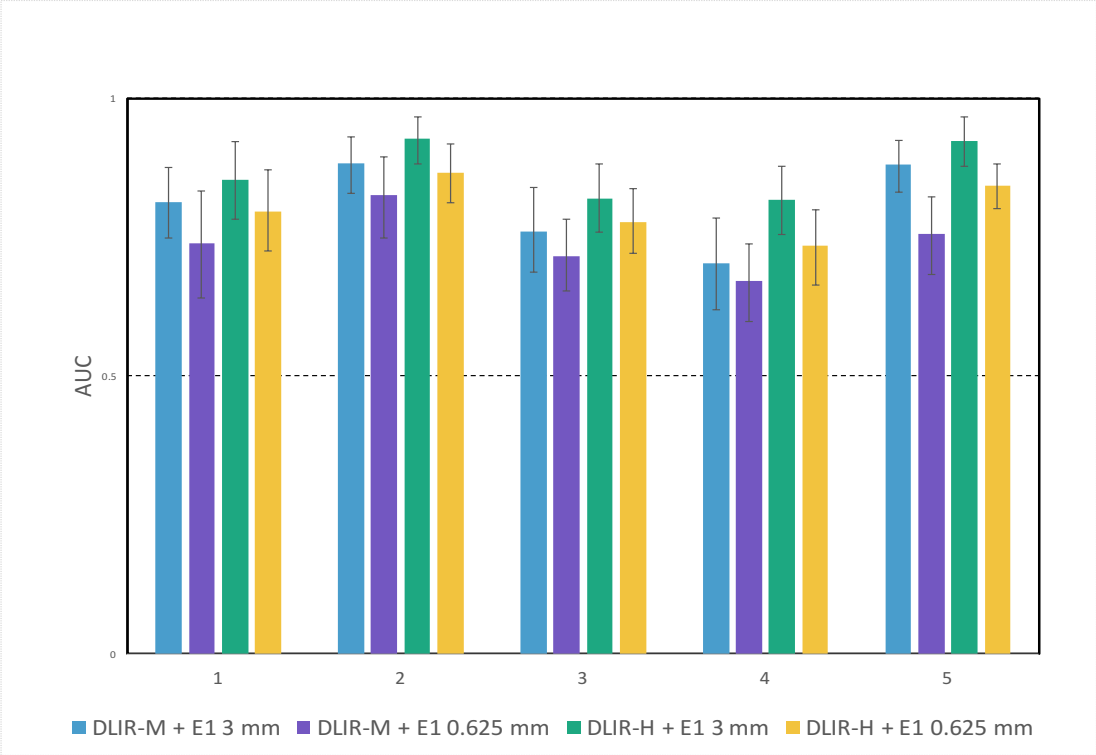

Figure 11: The area under the curve for the evaluated reconstructions for the abdominal exam, plotted for each image quality criterion. The 95% CI of the AUC is shown in the graph as error bars.

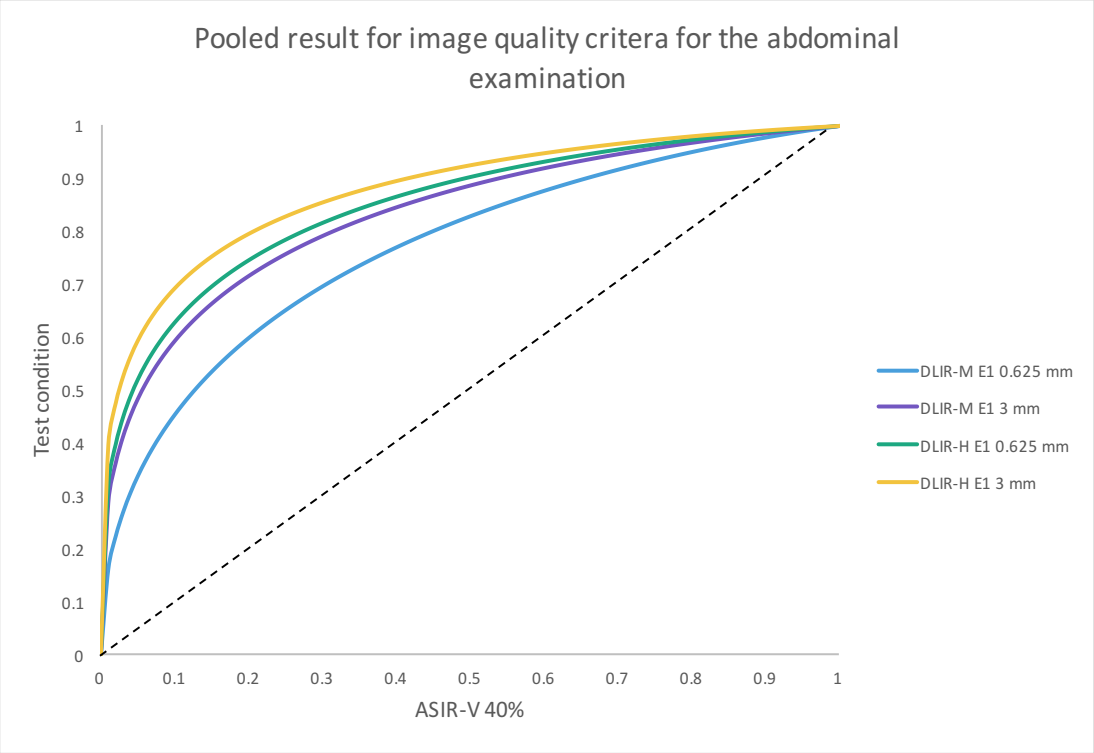

Figure 12: VGC curves from the analysis where all image quality criteria are pooled together for the abdominal exam.

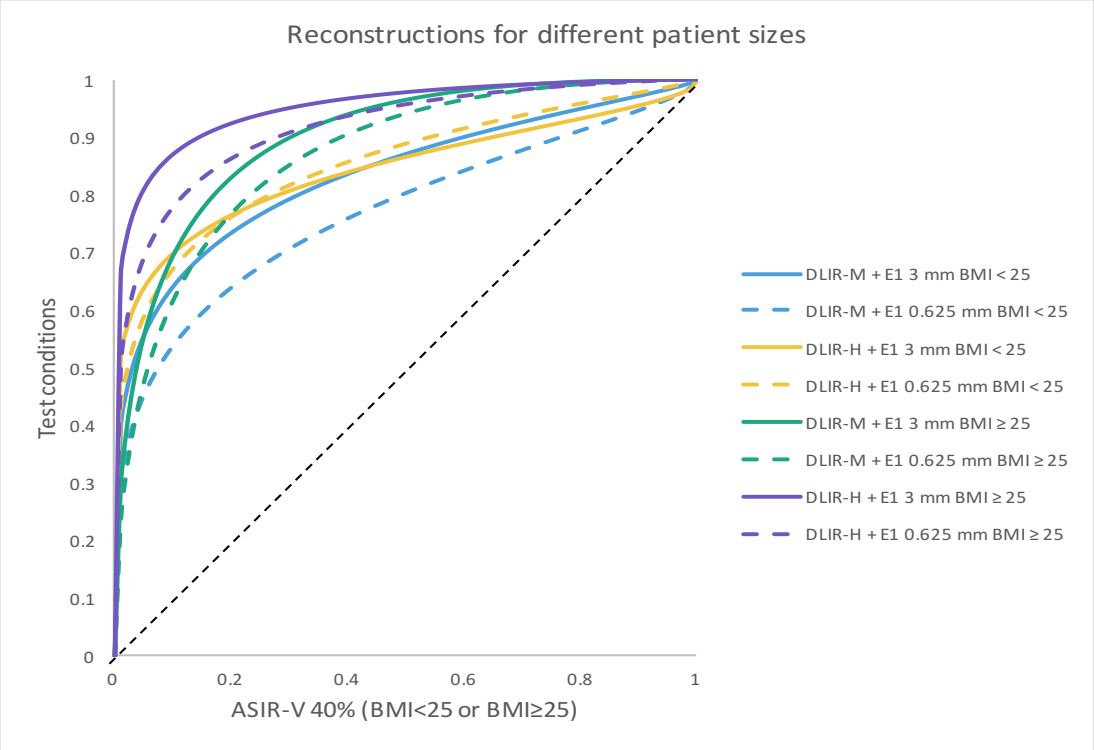

Figure 13: Pooled data for the two patient groups and their reconstructions where the solid lines show the reconstructions made with 3 mm slices filter and the dashed lines show the reconstructions made with 0.625 mm slices.

Brain

For the brain examination, most of the DLIR were rated similar as the ASIR-V 50% reconstruction, but for some criteria the ratings were lower for DLIR than for the ASIR-V 50% reconstruction, see Figure 14 and Table 5. The DLIR-H was generally rated lower than DLIR-M. For criterion 2 (visual reproduction of the basal ganglia), DLIR-M EC2 were rated significantly higher (the 95% CI of the AUC was above 0.5) than ASIR-V 50%. For all criteria except criterion 4 (visual reproduction of the posterior fossa structures), DLIR-M EC2 was preferred, see Table 5. For criterion 4, images reconstructed using ASIR-V 50 % were rated higher than images reconstructed using DLIR. The difference was statistically significant (the 95% CI of the AUC was below 0.5) for both DLIR-H EC1 and DLIR-H EC2. The VGC-curve for each criterion can be seen in Appendix under headings "Visual Grading Characteristics, Brain".

In the pooled result, see Figure 15, there was a significant difference between DLIR-H EC1 and DLIR-H EC2 compared to ASIR-V 50%, where both reconstructions using DLIR-H were rated lower than ASIR-V 50%. The DLIR-M EC2 obtained the highest ratings of the different DLIR combinations with an AUC of 0.53 and DLIR-H EC1 had the lowest score, with an AUC of 0.43.

When comparing the two groups of patients with different age to each other, no major differences could be found, see Figure 16.

Table 5: The area under the curve for the five image quality criteria for the brain examination for the evaluated reconstructions, where ASIR-V is the reference condition. The 95% CI of the AUC is shown inside the parenthesis.

	DLIR-M EC1	DLIR-M EC2	DLIR-H EC1	DLIR-H EC2
Criterion 1	0.50 (0.42-0.62)	0.51 (0.42-0.59)	0.48 (0.40-0.55)	0.50 (0.42-0.59)
Criterion 2	0.55 (0.50-0.58)	0.60 (0.53-0.66)	0.49 (0.46-0.53)	0.51 (0.44-0.58)
Criterion 3	0.50 (0.42-0.57)	0.53 (0.47-0.60)	0.45 (0.38-0.52)	0.44 (0.36-0.52)
Criterion 4	0.47 (0.39-0.54)	0.45 (0.38-0.52)	0.33 (0.26-0.40)	0.38 (0.31-0.44)
Criterion 5	0.51 (0.44-0.59)	0.53 (0.47-0.58)	0.36 (0.29-0.44)	0.43 (0.35-0.50)

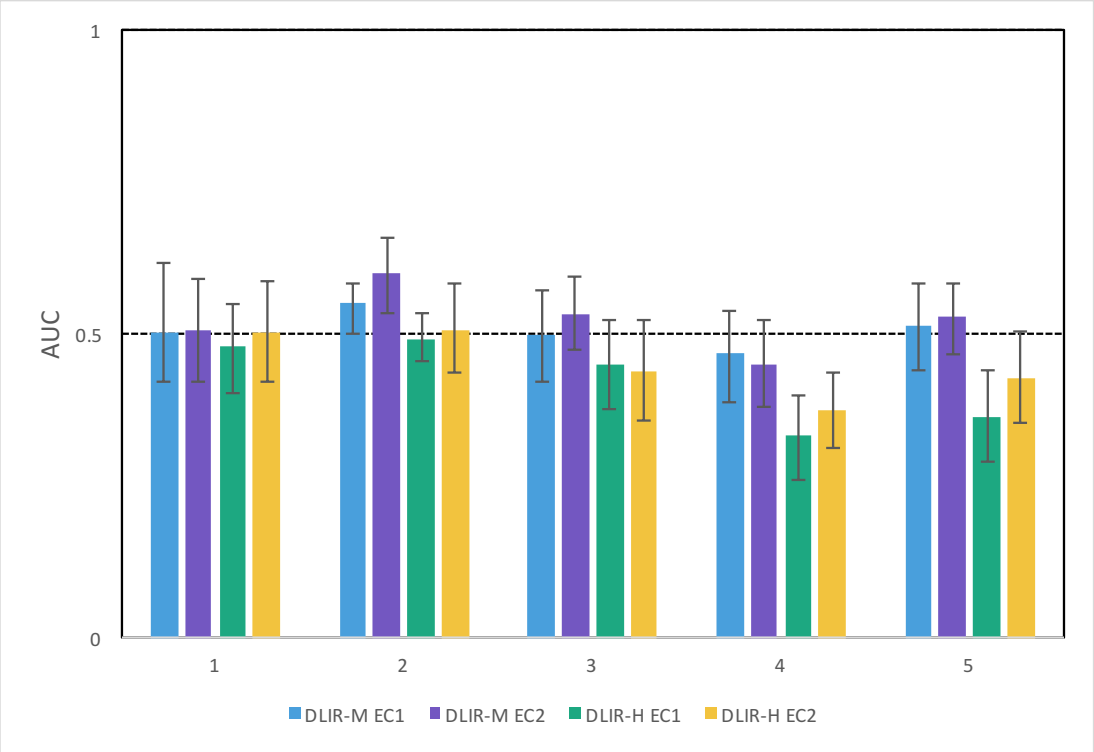

Figure 14: The area under the curve for the evaluated reconstructions for the brain exam, plotted for each image quality criterion. The 95% CI of the AUC is shown in the graph as error bars.

Figure 15: VGC curves from the analysis where all image quality criteria are pooled together for the brain exam.

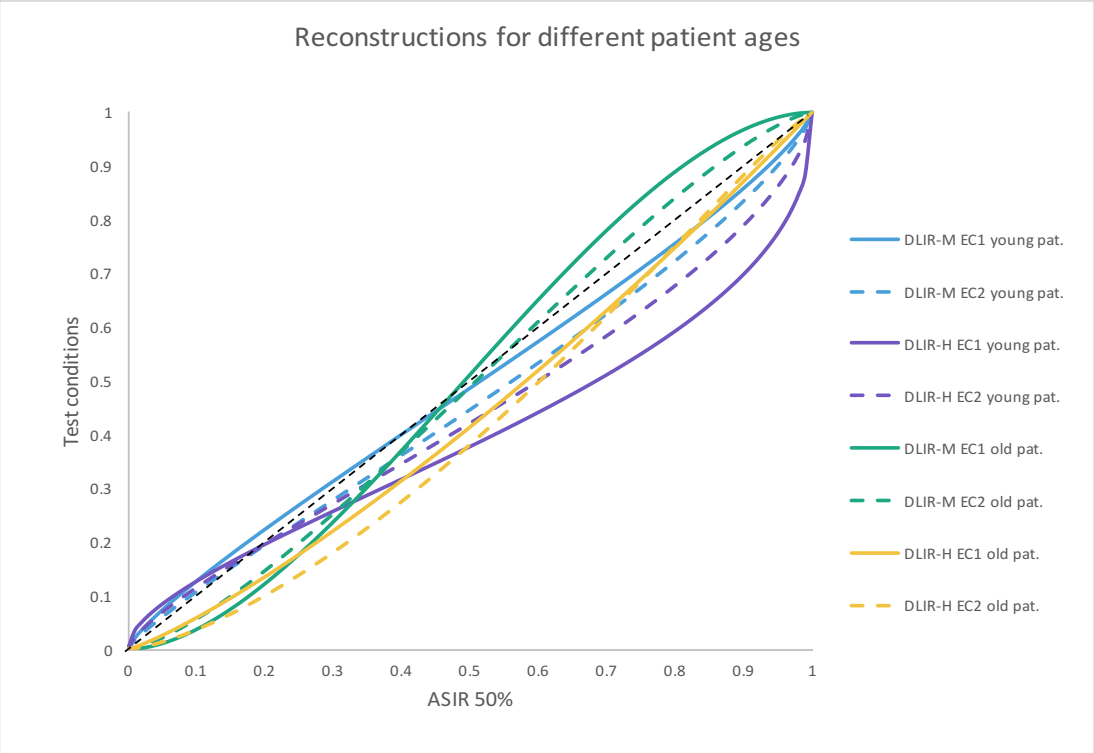

Figure 16: Pooled data for the two patient groups and their reconstructions where the solid lines show the reconstructions made with EC1 filter and the dashed lines show the reconstructions made with EC2 filter.

Phantom studies

Catphan

Abdomen

For the examination performed using the abdomen protocol, the Hounsfield Units (HU) had a slight, but not statistically significant, variation between the different reconstructions, see Table 6.

Table 6: The measured HU for different materials for NI 14, the standard deviation is shown inside the parenthesis.

	Recon.	Water	Polystyrene	Teflon	Air	Homogenous
NI 14	ASIR-V 40%	-10.6 (30.3)	-25 (30)	806 (142)	-848 (264)	99 (18)
	DLIR-M E1	-10.9 (40.9)	-28 (34)	823 (146)	-864 (276)	99 (20)
	DLIR-H E1	-11.2 (39.3)	-28 (33)	823 (146)	-864 (276)	99 (17)

The different reconstructions had no obvious effect on spatial resolution. For all reconstructions, six line-pairs per mm could be distinguished see Figure 17.

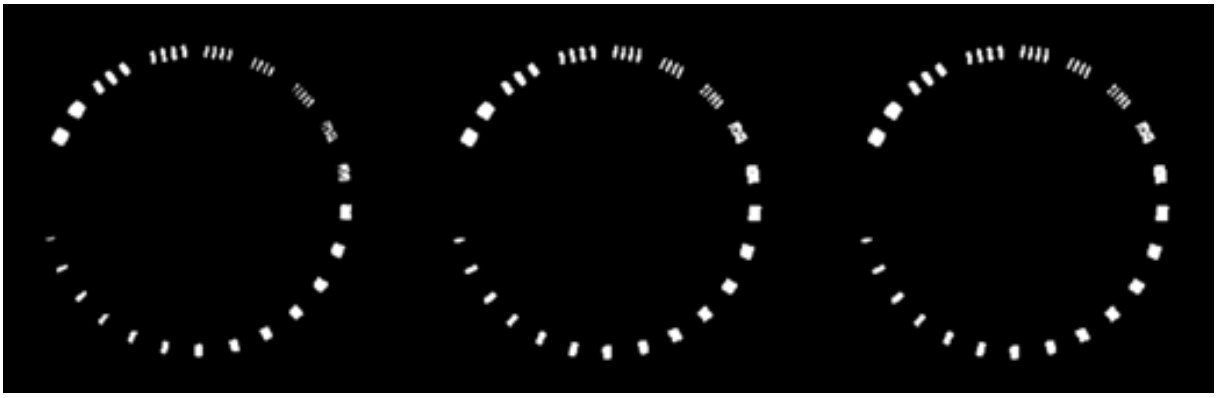

Figure 17: Spatial resolution for the reconstructions; (left to right) ASIR-V 40%, DLIR-M E1 and DLIR-H E1.

The low contrast objects were slightly easier to distinguish in the DLIR-H E1 image compared to the ASIR-V 40% image, see Figure 18. However, because of the relatively high NI used in the abdomen protocol, i.e. NI 14, the images were relatively noisy which limits the possibilities to distinguish low contrast objects from the background.

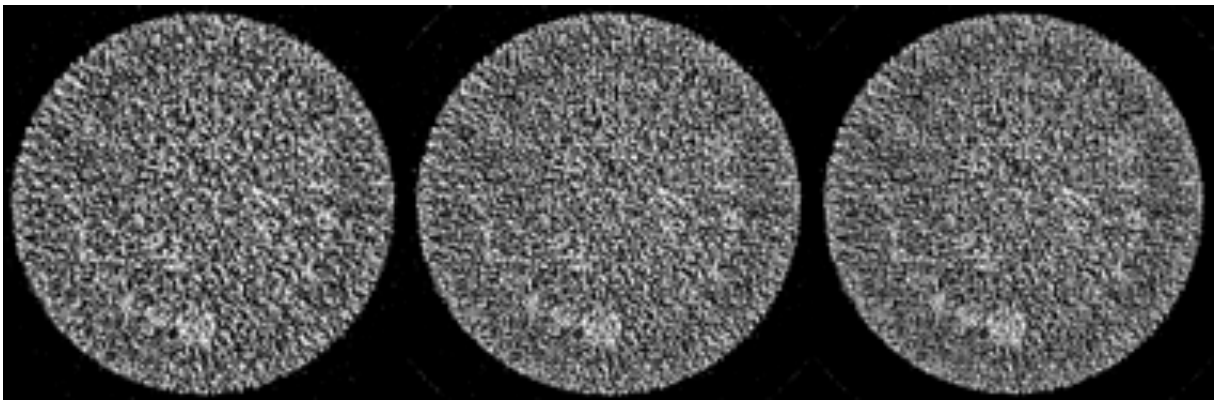

Figure 18: Low contrast for the reconstructions; (left to right) ASIR-V 40%, DLIR-M E1 and DLIR-H E1, all with NI 14.

Brain

For the examination performed using the brain protocol, the HU had a slight, but not statistically significant, variation between the reconstructions, see Table 7.

Table 7: The measured HU for different materials for NI 3.3, the standard deviation is shown inside the parenthesis.

	Recon.	Water	Polystyrene	Teflon	Air	Homogenous.
NI 3.3	ASIR-V 50%	-3.3 (8.2)	-29 (27)	743 (245)	-720 (392)	95 (5)
	DLIR-M EC1	-2.6 (7.3)	-30 (27)	745 (250)	-722 (398)	95 (5)
	DLIR-M EC2	-2.5 (7.3)	-30 (26)	745 (250)	-722 (398)	95 (5)
	DLIR-H EC1	-2.6 (7.3)	-30 (26)	745 (250)	-722 (398)	95 (4)
	DLIR-H EC2	-2.6 (7.3)	-30 (26)	745 (250)	-722 (398)	95 (4)

The different reconstructions had no obvious effect on spatial resolution. For all reconstructions, six line-pairs per mm could be distinguished, see Figure 19.

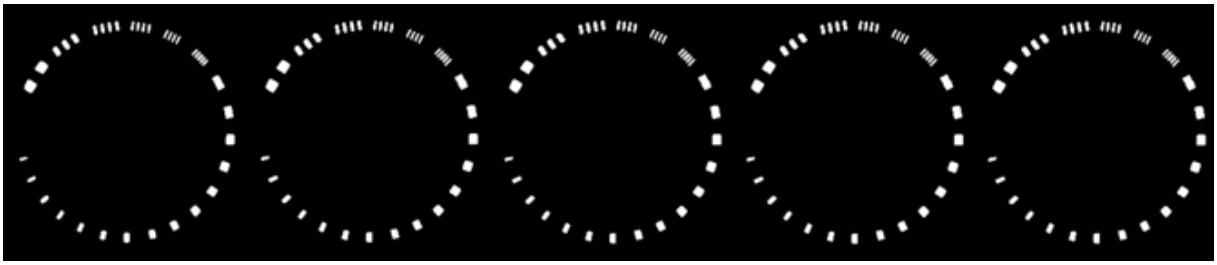

Figure 19: Spatial resolution for the reconstructions; (left to right) ASIR-V 50%, DLIR-M EC1, DLIR-M EC2, DLIR-H EC1 and DLIR-H EC2, all with NI 3.3

The low contrast resolution for all reconstructions can be seen in Figure 20. The low contrast objects were slightly easier to distinguish in the images reconstructed using DLIR than in the images reconstructed using ASIR-V 50%.

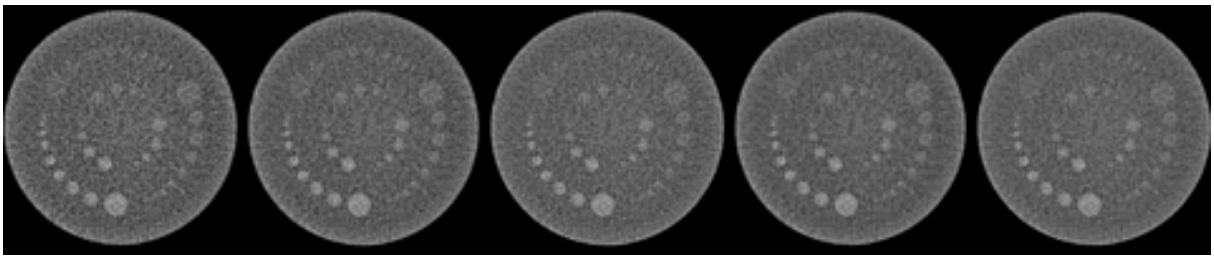

Figure 20: Low contrast for the reconstructions; (left to right) ASIR-V 50%, DLIR-M EC1, DLIR-M EC2, DLIR-H EC1 and DLIR-H EC2, all with NI 3.3

Anthropomorphic phantom

Abdomen

For BMI 19, the $CTDI_{vol}$ for the standard scan (NI 14) was measured to 4.5 mGy, and for the scan performed at the 50% dose level (NI 18) the $CTDI_{vol}$ was 2.5 mGy, see Table 8.

For BMI 40, the $CTDI_{vol}$ for the standard scan (NI 14) was measured to 24.8 mGy. When increasing the NI to 18, the dose was lowered by approximately 50% to 11.3 mGy, see Table 8.

In Figure 21 and 22, images reconstructed using ASIR-V 40% at original dose level, i.e. NI 14, and images reconstructed using DLIR-H + E1 at 50% reduced dose, i.e. NI 18, for BMI 19 and BMI 40, respectively. Even at a dose level of 50% of original dose, i.e. NI 18, the noise content in the DLIR image seems to be lower than in the images acquired at the original dose level and reconstructed using ASIR-V 40 %.

Table 8: $CTDI_{vol}$ and DLP measured in the two ROI's for the anthropomorphic phantom, both for BMI 19 and for BMI 40.

	NI	$CTDI_{vol}$ [mGy]	DLP [mGy*cm]
BMI 19			
ASIR-V 40%	**14**	**4.47**	**228.41**
DLIR H E1	18	2.52	129.02
BMI 40			
ASIR-V 40%	**14**	**24.77**	**1264.98**
DLIR H E1	18	11.32	579.28

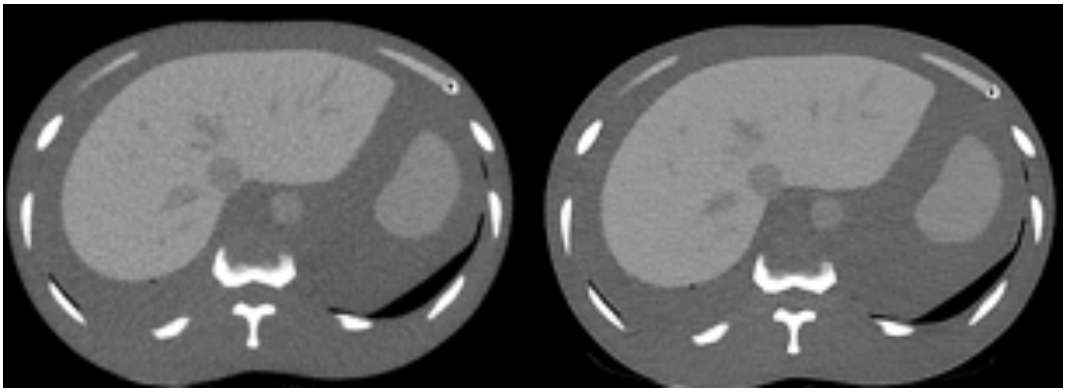

Figure 21: Images of the anthropomorphic phantom with BMI 19, reconstructed using ASIR-V 40% at original dose level (NI 14) (left) and DLIR-H E1 at reduced dose level (NI 18) (right).

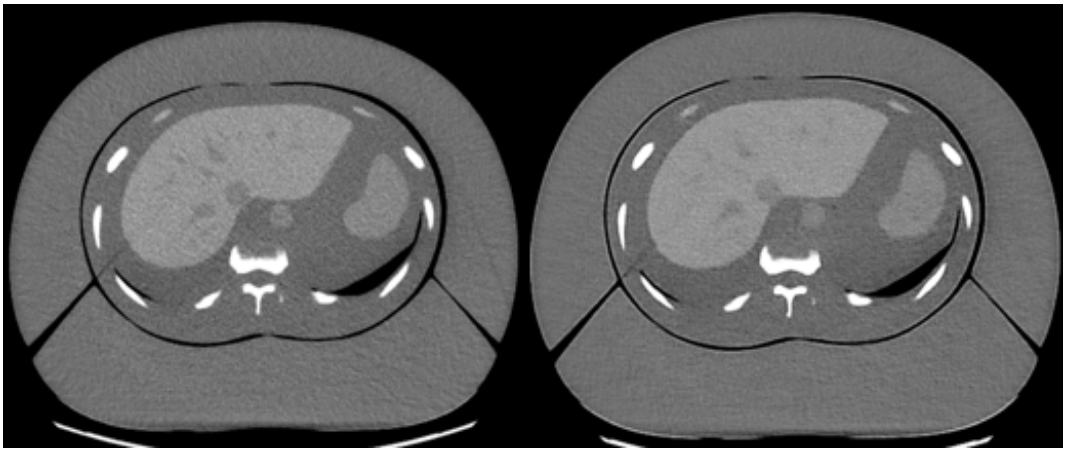

Figure 22: Images of the anthropomorphic phantom with BMI 40, reconstructed using ASIR-V 40% at original dose level (NI 14) (left) and DLIR-H E1 at reduced dose level (NI 18) (right).

Water phantom

Abdomen

The mean standard deviation measured (SD) for the water phantom images can be seen in Table 9, for the reconstructions made with the E1-filter and for the reconstructions made without the E1-filter. When the E1-filter was applied, the mean SD for DLIR-M was higher than the mean SD for ASIR-V 40%, both for NI 14 and NI 18. However, when the E1-filter

was removed, the mean SD in the images reconstructed using DLIR was decreased and the mean SD for ASIR-V 40% was the highest.

Table 9: The mean SD in the water phantom images for the abdominal protocol, for the reconstructions made with and without the E1-filter on the reconstructions made with DLIR.

Reconstruction	Noise Index	Mean SD with E1	Mean SD without E1
ASIR-V 40%	14	16.8	16.8
DLIR-M	14	18.2	15.0
DLIR-H	14	15.0	12.5
ASIR-V 40%	18	20.3	20.3
DLIR-M	18	22.6	18.7
DLIR-H	18	18.5	15.4

The NPS for the different reconstructions evaluated in the present study are shown in Figure 23. The highest magnitude of noise was found in the DLIR-M E1 images, while the DLIR-H E1 images had the lowest magnitude of noise. Comparing the NPS for NI 14 and NI 18, the presence of noise was, as expected, higher for the reconstructions with NI 18.

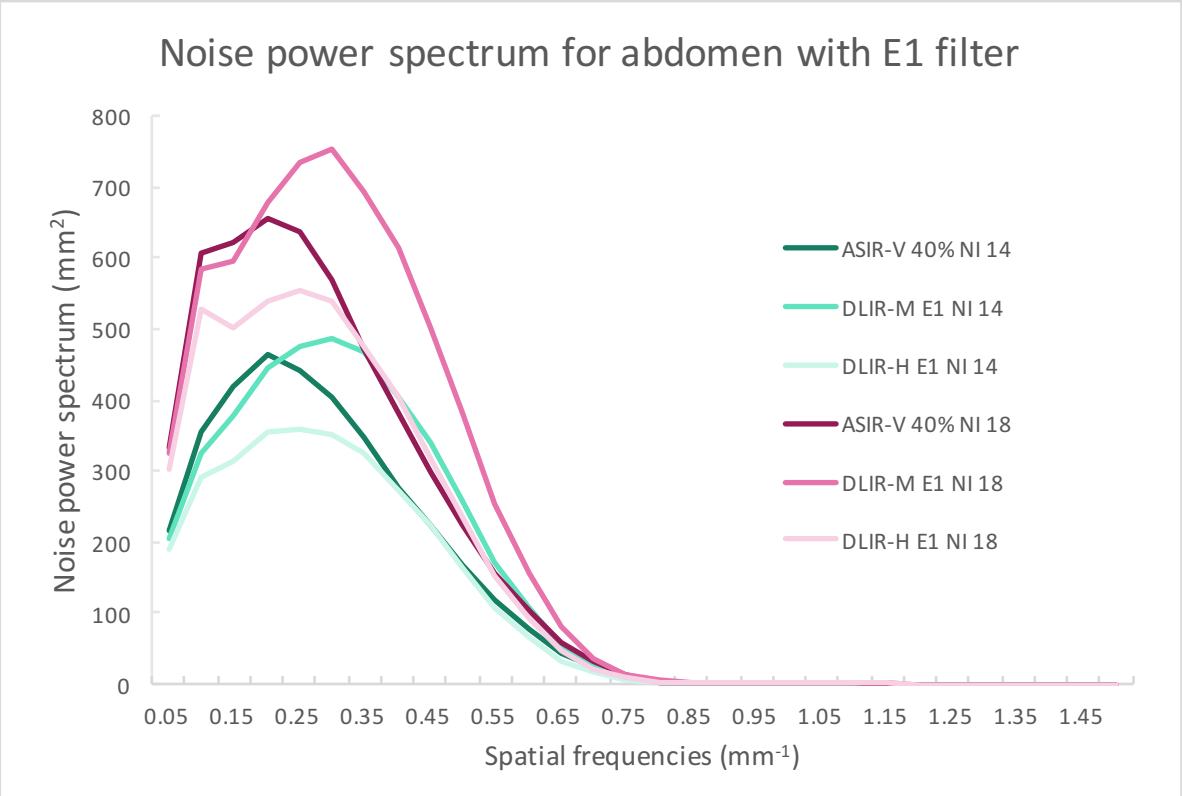

Figure 23: Noise power spectrum for the abdominal examination with NI 14 and NI 18 using the E1-filter on the DLIR.

The NPS for the reconstructions performed without the E1 filter is shown in Figure 24. When removing the E1 filter, both DLIR reconstructions had a lower noise magnitude than ASIR-V 40%. This was observed for both NI 14 and NI 18.

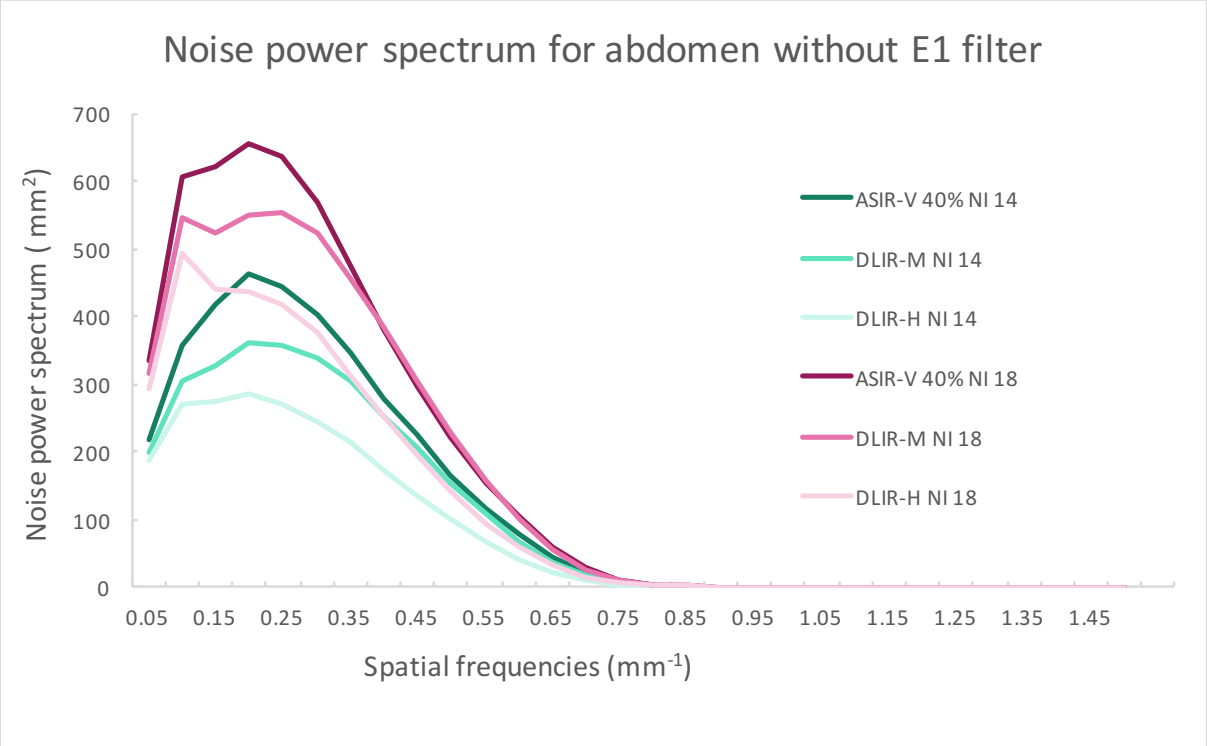

Figure 24: Noise power spectrum for the abdominal examination with NI 14 and NI 18 without the E1-filter on the DLIR.

A visual example of the difference in image quality between the DLIR-M images reconstructed with and without the E1 filter, for the anthropomorphic phantom with BMI 40, is shown in Figure 25. The image reconstructed using the E1 filter is perceived as noisier than the image reconstructed without the E1 filter.

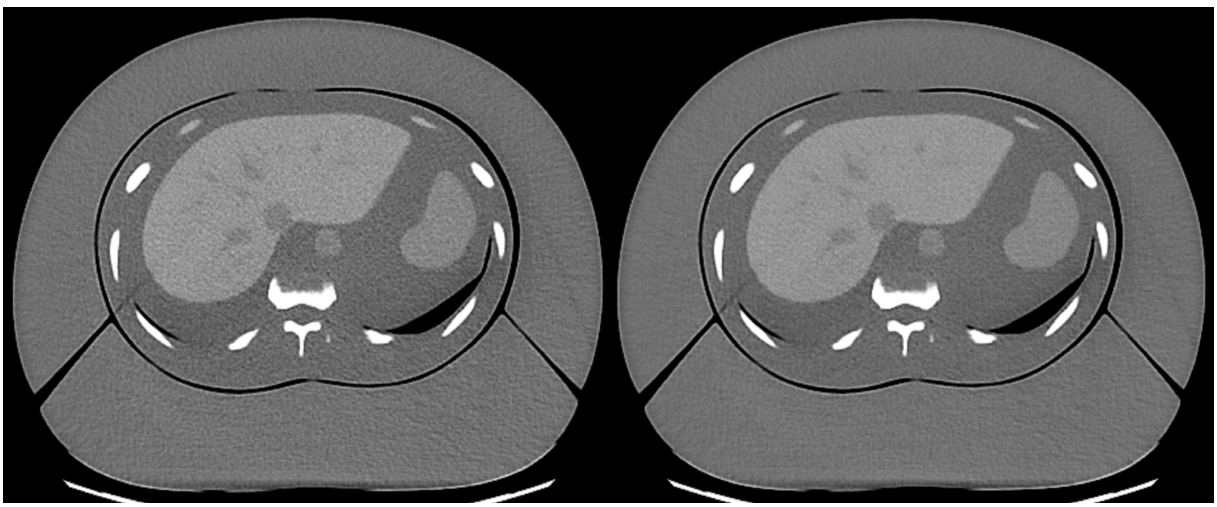

Figure 25: Axial slices of the abdomen from the anthropomorphic phantom with BMI 40 showing the DLIR-M reconstruction with E1 filter (left) and without E1 filter (right).

Brain

The measured mean SD in the NPS for the brain protocol can be seen in Table 10. For ASIR-V 50% and DLIR-M EC1, the mean SD was the same. For DLIR-H EC1, the mean SD was lower compared to the other two reconstructions.

Table 10: The mean SD in the water phantom images for the brain protocol.

Reconstruction	Noise Index	Mean SD
ASIR-V 50%	3.3	3.4
DLIR-M EC1	3.3	3.4
DLIR-H EC1	3.3	2.8

The NPS for the brain protocol is shown in Figure 26. The highest magnitude of noise is found in the ASIR-V 50% reconstruction and the lowest magnitude is found in DLIR-H EC1. This is similar to the NPS for the abdominal protocol without the E1 filter where both DLIR reconstructions have a lower noise magnitude than the ASIR-V reconstruction.

When comparing the NPS for images reconstructed with the EC1 and EC2 filter, no difference in NPS was found, why only EC1 is shown in Figure 26.

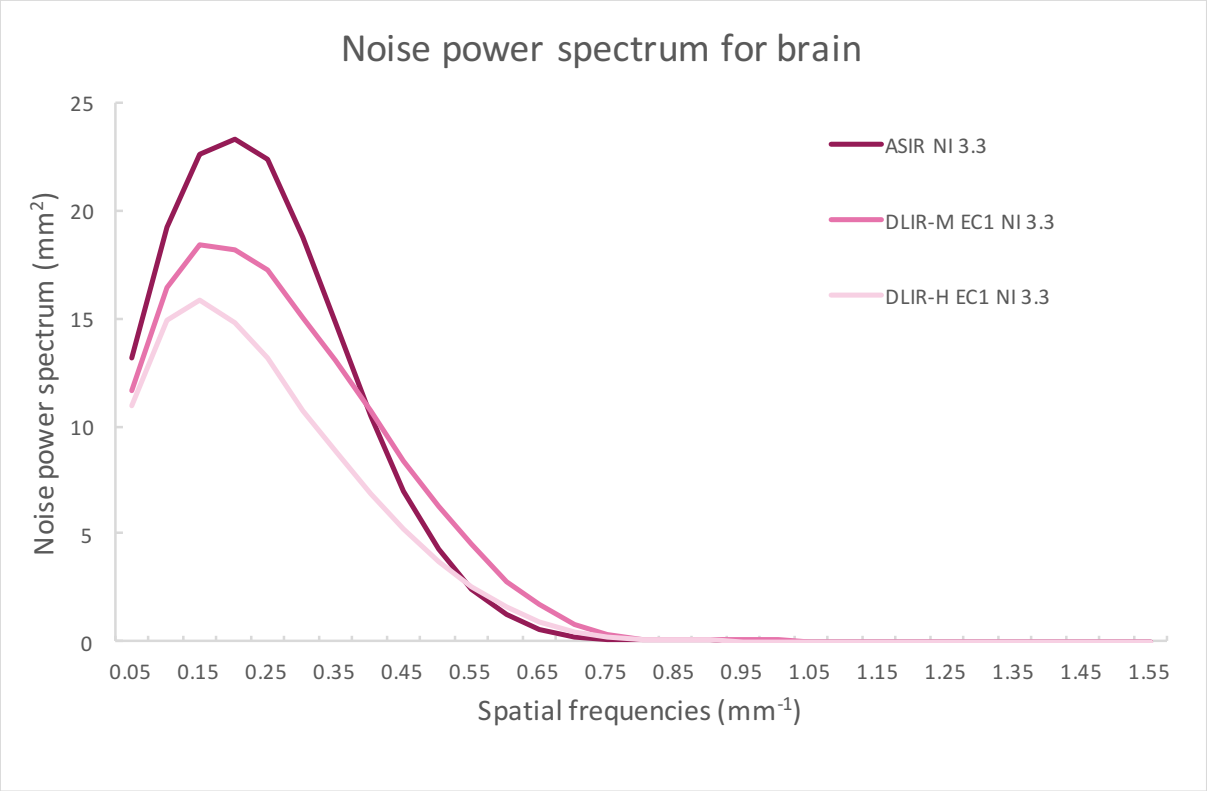

Figure 26: Noise power spectrum for the brain examination with NI 3.3.

Discussion

In the VGC study including the abdominal examinations, all DLIR reconstructions were rated significantly higher than the ASIR-V 40% reconstruction for all image quality criteria. The

DLIR-H 3 mm reconstruction was preferred by the radiologists. For the brain examination, most of the DLIR reconstructions were rated similar to the ASIR-V 50% reconstruction.

The Catphan study showed that the reconstructions did not affect the spatial resolution or the HU linearity. Low contrast objects were slightly easier to distinguish int the images reconstructed using DLIR than in the images reconstructed using ASIR-V. The evaluation of standard deviation and NPS in the water phantom showed that images reconstructed using ASIR-V had a higher standard deviation and higher noise magnitude than images reconstructed using DLIR. Finally, the evaluation of the images obtained using the anthropomorphic phantom showed that it might be possible to reduce the dose for abdomen examinations with approximately 50% when DLIR-H E1 is used instead of ASIR-V for image reconstruction without a significant reduction in image quality.

Evaluation of reconstructions

Abdomen

The results from the VGC study including abdominal examinations showed that images reconstructed using DLIR were rated higher for the larger patients than for the smaller patients. Larger patients will give rise to higher amounts of scattered photons than a smaller patient. This is due to the fact that a larger patient has more tissue that the photons can interact with. The scattered photons that reach the detector will result in a reduced contrast in the image. Hence, images from the larger patient generally have a lower image contrast than images from a smaller patient [27]. In the present study it was shown that low contrast objects were slightly easier to perceive in images reconstructed using DLIR than in images reconstructed using with ASIR-V. This might partly explain why DLIR was rated higher for larger patients then for smaller patients.

The 0.625 mm slices had lower scores than the 3 mm slices in most cases, for both DLIR-H and DLIR-M. A reason for this could be that the signal to noise ratio decreases when the slice thickness, and thereby the voxel size, is reduced. Therefore, images reconstructed with thinner slice thickness are noisier and thereby harder to evaluate. However, a thinner slice thickness might facilitate the possibility to find anatomy that otherwise would be hidden because of partial volume effect.

Even though the images reconstructed using a slice thickness of 0.625 mm slices obtained lower ratings than images reconstructed using a slice thickness of 3 mm, the 0.625 mm slices were still rated significantly higher than the images reconstructed using ASIR-V 40% 3 mm slice thickness for all image quality criteria, which can be seen in Figure 11. This fact, together with the fact that images reconstructed using ASIR-V 40% with 3 mm slice thickness previously have been considered to have sufficient clinical image quality, indicates an opportunity to significantly reduce the radiation dose used for abdominal examinations when DLIR is used instead of ASIR-V 40% for image reconstruction.

Brain

The results from the VGC study including brain examinations showed, for most criteria, no statistically significant differences between images reconstructed using DLIR and images reconstructed using ASIR-V 50%. In some cases, images reconstructed using ASIR-V 50% were rated slightly higher than images reconstructed using DLIR. This might be because the radiologists are used to review images reconstructed with ASIR-V.

The posterior fossa is hard to review because of the density of the skull bones. The bones induce streak artifacts in the images and makes the images harder to interpret. Although not statistically significant for all DLIR, this criterion was the only criteria were all the DLIR were rated lower than ASIR-V 50%. Perhaps the posterior fossa is slightly better visualized with ASIR-V 50%.

There were no obvious differences in ratings between the younger and older patients. One observer made a comment that there might be a bias for criterion 3 (visual reproduction of the cerebrospinal fluid space around the mesencephalon), where the cerebrospinal fluid becomes easier to distinguish in older patients' due to atrophy. However, this potential bias could not be identified in the results from the study.

Today, DLIR is only applicable with a standard kernel. For the abdominal examination, which clinically is performed with a standard kernel, this did not matter. However, for the brain examination, which is clinically performed using the *Soft* kernel, it is important to take this fact into account. In the present study ASIR-V 50% images reconstructed with *Soft* kernel was compared to DLIR images reconstructed with *Stnd* kernel. This means that not only are the radiologists reviewing images reconstructed using two completely different methods, but also images reconstructed with different kernels. The brain radiologists are used to review images with the *Soft* kernel, and therefore there might be a bias in the assessment. For a fairer comparison of DLIR and ASIR-V, a new assessment should be made in the future when the *Soft* kernel can be applied also with DLIR, so that images reconstructed with ASIR-V and *Soft* kernel can be compared with images reconstructed using DLIR and *Soft* kernel.

Noise properties

In theory, DLIR should affect the noise frequencies and lower the noise in the images compared to ASIR-V [6]. The same result was found both for NI 14 and NI 18 in the present study, see Figure 24 and Table 9. However, when post-processing the DLIR images using the E1-filter, the noise magnitude in both DLIR-M + E1 and DLIR-H + E1 increased. In this situation, the noise magnitude in the DLIR-M + E1 images was higher than the noise magnitude in the ASIR-V 40% images. The E1 filter is an edge enhancing filter, which also enhances the noise in the image. When this is applied on the reconstruction, the noise which was suppressed by the reconstruction algorithm, will be enhanced again. This is visible when comparing DLIR images reconstructed both with and without the E1 filter, see Figure 25.

Even though the noise magnitude is increased when image post-processing using the E1 filter is performed, the edge enhancement obtained could increase the visual delineation of anatomical structures in the images and thereby also facilitate the detection of pathology. Therefore, it might still be more valuable for the overall image quality to apply the E1 filter on the images after reconstructing the images using DLIR.

In the images of the anthropomorphic phantom at the different dose levels, the noise content seems to be similar between ASIR-V 40% (NI 14) and DLIR-H E1 (NI 18), see Figures 21 and 22. However, the NPS and mean SD from the water phantom showed that there is a slightly higher noise content in the DLIR-H E1 (NI 18) images compared to ASIR-V 40% (NI 14) images when using the E1-filter, see Figure 23 and Table 9. When removing the E1-filter,

the NPS and mean SD for DLIR-H E1 (NI 18) was decreased to the same level as ASIR-V 40% (NI 14), see Figure 24 and Table 9.

Limits of the study

It is hard to evaluate image quality in an objective way since an image is perceived differently for different individuals. Visual grading is an effective method of evaluating image quality when comparing two conditions, in this case two reconstruction methods. To add objectivity to the study, three phantom studies were done to quantify the image quality.

In the VGC study, three radiologists reviewed each examination, which is a small number of observers. The ratings of each reconstruction are a personal observation from each observer. Having more observers would have made the results more reliable since it then would have been easier draw general conclusions of the results. Since it is a VGC study, the result depends on how the three radiologists perceived the reconstructions and the different image quality criteria. By increasing the number of observers, the result would become closer to the truth. However, finding a large group of specialist radiologist that would be able to participate in the study is difficult due to limited resources available. The abdominal and neuro specialist radiologists that participated in this study have many years of experience in their field. Therefore, their views on the reconstructions are of great value for the evaluation.

Each radiologist reviewed five reconstruction series for 20 patients, adding up to a total of 100 series. Having a larger group of patients also would have improved the validity of the results from the present study, but more patients to review would also have led to a longer review time for each observer. In the present study, the number of included patients therefore needed to be weighed against the number of different reconstructions evaluated.
In this study, five reconstructions for each examination were evaluated, including the reconstruction used today at Sahlgrenska University Hospital. The reconstruction and filter combinations were chosen by two experienced radiologists. There are many other combinations of reconstructions and filters available, but it would be too time consuming to evaluate them all. Therefore, the radiologist made the selection based on which filters are used clinically today. For a more thorough evaluation, future studies should include also other reconstruction settings. Also, it is not obvious that the results from a VGC study can be translated to the clinical situation where, for example, detection of pathology is the task. The visual grading concept is based on the theory that visibility of anatomical structures can be translated to visibility of pathology [17] [28]. This is not always true, as studies based on visual grading and detection of pathology in some cases have been shown to give different results regarding image quality [29]. Therefore, future studies should also include the task of detection of pathology to ensure that the image quality obtained in images reconstructed using DLIR is sufficient also for detection of pathology.

Conclusion

This study showed that DLIR is valuable when reviewing abdominal images, as the images reconstructed using DLIR were rated significantly higher than images reconstructed using ASIR-V 40% by all physicians in the VGC-study. However, for brain examinations, no major

difference between the images reconstructed using DLIR and the images reconstructed using ASIR-V 50% was found.

The phantom studies showed that there were no major visual differences in spatial resolution, low contrast detectability and linearity of HU between images reconstructed using DLIR and ASIR-V. It was however shown that the magnitude of noise was lower for the images reconstructed with DLIR than the images reconstructed with ASIR-V. Notable, when post-processing the images using the E1-filter, there is an increase in the magnitude of noise.

The overall result showed that there might be a possibility to reduce the radiation dose for abdominal examinations when reconstructing the images using DLIR instead of ASIR-V.

References

[1] Radiation dose in X-ray and CT Exams. RadiologyInfo. Updated February 1, 2021. https://www.radiologyinfo.org/en/info/safety-xray

[2] Alara. United States Nuclear Regulatory Commission. Updated March 9, 2021. https://www.nrc.gov/reading-rm/basic-ref/glossary/alara.html

[3] Schofield R, King L, Tayal U, et al. Image reconstruction: Part 1 - understanding filtered back projection, noise and image acquisition. *J Cardiovasc Comput Tomogr*. 2020;14(3):219-225. doi:10.1016/j.jcct.2019.04.008

[4] Franck C, Zhang G, Deak P, Zanca F. Preserving image texture while reducing radiation dose with a deep learning image reconstruction algorithm in chest CT: A phantom study. *Phys Med*. 2021;81:86-93. doi:10.1016/j.ejmp.2020.12.005

[5] Euler A, Solomon J, Marin D, Nelson RC, Samei E. A Third-Generation Adaptive Statistical Iterative Reconstruction Technique: Phantom Study of Image Noise, Spatial Resolution, Lesion Detectability, and Dose Reduction Potential. *AJR Am J Roentgenol*. 2018;210(6):1301-1308. doi:10.2214/AJR.17.19102

[6] A new era of image reconstruction: TrueFidelity. White Paper. GE Healthcare. 2019

[7] Racine D, Becce F, Viry A, et al. Task-based characterization of a deep learning image reconstruction and comparison with filtered back-projection and a partial model-based iterative reconstruction in abdominal CT: A phantom study. *Phys Med*. 2020;76:28-37. doi:10.1016/j.ejmp.2020.06.004

[8] Goldman LW. Principles of CT and CT technology. *J Nucl Med Technol*. 2007;35(3):115-130. doi:10.2967/jnmt.107.042978

[9] Hsieh J. *Computed Tomography: Principles, design, artifacts and recent advances.* 2nd edition. John Wiley & Sons; 2009.

[10] Stiller W. Basics of iterative reconstruction methods in computed tomography: A vendor-independent overview. *Eur J Radiol*. 2018;109:147-154. doi:10.1016/j.ejrad.2018.10.025

[11] Gonzalez RC, Woods R. *Digital Image Processing*. 4th edition. Pearson Education Limited; 2018. p. 153-197, 368-386, 943-948, 953-954.

[12] Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. *Insights Imaging*. 2018;9(4):611-629. doi:10.1007/s13244-018-0639-9

[13] Rydén T, Van Essen M, Marin I, Svensson J, Bernhardt P. Deep-Learning Generation of Synthetic Intermediate Projections Improves ^{177}Lu SPECT Images Reconstructed with Sparsely Acquired Projections. *J Nucl Med*. 2021;62(4):528-535. doi:10.2967/jnumed.120.245548

[14] Cruickshank A, Bell D. Kernel (image reconstruction for CT). Radiopaedia. Updated April 19, 2020. Accessed on November 14, 2021. https://doi.org/10.53347/rID-72906

[15] Rock P, Bell D. Noise power spectrum. Radiopaedia. Updated July 20, 2021. Accessed on November 14, 2021. https://doi.org/10.53347/rID-82093

[16] Båth M. Linear-systems analysis. Lecture presented at: Canvas, October 20, 2020.

[17] Båth M, Månsson LG. Visual grading characteristics (VGC) analysis: a non-parametric rank-invariant statistical method for image quality evaluation. *Br J Radiol*. 2007;80(951):169-176. doi:10.1259/bjr/35012658

[18] Hansson J, Månsson LG, Båth M. Evaluation of VGC Analyzer by comparison with gold standard ROC software and analysis of simulated visual grading data. *Radiat Prot Dosimetry*. 2021;195(3-4):378-390. doi:10.1093/rpd/ncab066

[19] Hansson J. *Development of Methods and Strategies for Optimisation of X-ray Examinations.* Doctoral Thesis Sahlgrenska Academy, Gothenburg; 2009.

[20] Båth M, Hansson J. VGC Analyzer: A software for statistical analysis of fully crossed multiple-reader multiple-case visual grading characteristics studies. *Radiat Prot Dosimetry*. 2016;169(1-4):46-53. doi:10.1093/rpd/ncv542

[21] Svalkvist A, Svensson S, Hagberg T, Båth M. VIEWDEX 3.0-Recent development of a software application facilitating assessment of image quality and observer performance. *Radiat Prot Dosimetry*. 2021;195(3-4):372-377. doi:10.1093/rpd/ncab014

[22] van Pinxteren W. Systemets uppbyggnad Fördjupningskurs Revolution CT/APEX. Lecture presented at: Sahlgrenska University Hospital, September 28, 2021, Gothenburg.

[23] van Pinxteren W. Factors Affecting Dose & Image Quality in CT. Lecture presented at: Sahlgrenska University Hospital, September 16, 2021, Gothenburg.

[24] REVOLUTION APEX: The new way to your best image. Product data sheet. GE Healthcare. 2019

[25] A Guide to CT Radiation Dose Management. Blueprint. GE Healthcare. 2012.

[26] Boedeker KL, McNitt-Gray MF. Application of the noise power spectrum in modern diagnostic MDCT: part II. Noise power spectra and signal to noise. *Phys Med Biol*. 2007;52(14):4047-4061. doi:10.1088/0031-9155/52/14/003

[27] Modica MJ, Kanal KM, Gunn ML. The obese emergency patient: imaging challenges and solutions. *Radiographics*. 2011;31(3):811-823. doi:10.1148/rg.313105138

[28] Jensen K, Andersen HK, Smedby Ö, et al. Quantitative Measurements Versus Receiver Operating Characteristics and Visual Grading Regression in CT Images Reconstructed with Iterative Reconstruction: A Phantom Study. *Acad Radiol*. 2018;25(4):509-518. doi:10.1016/j.acra.2017.10.020

[29] Mirzai M, Meltzer C, Vikgren J, et al. The Effect of Dose Reduction on Overall Image Quality in Clinical Chest Tomosynthesis. *Acad Radiol*. 2021;28(10):e289-e296. doi:10.1016/j.acra.2020.05.041

Appendix

Evaluated anatomy

Abdomen

In Figure A and Figure B, the evaluated anatomy for the abdominal examination are shown.

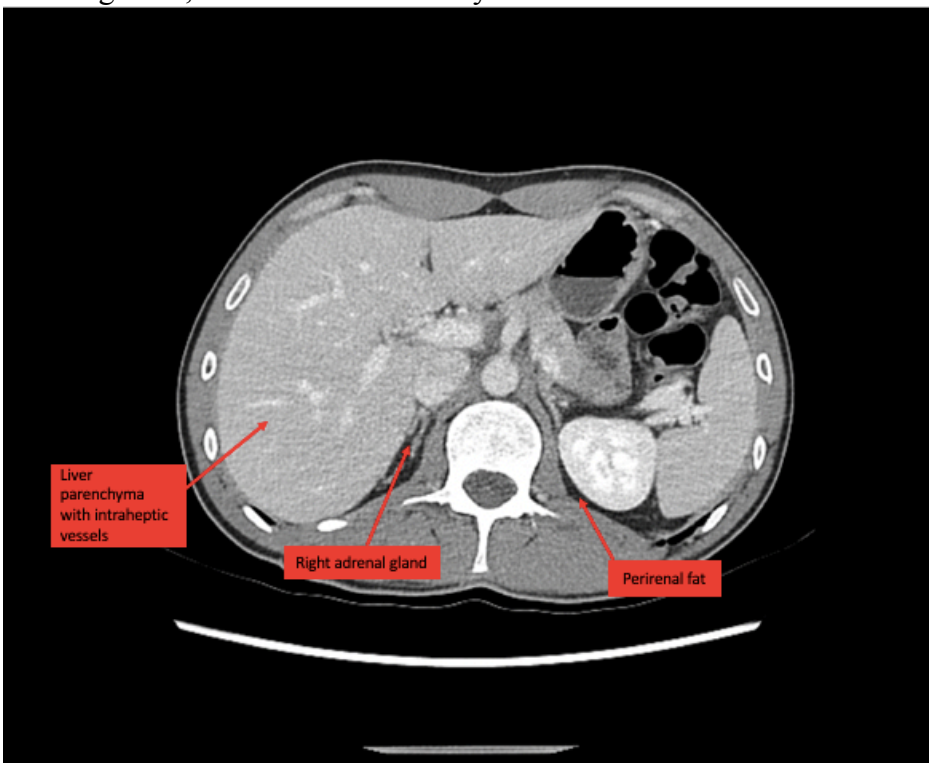

Figure A

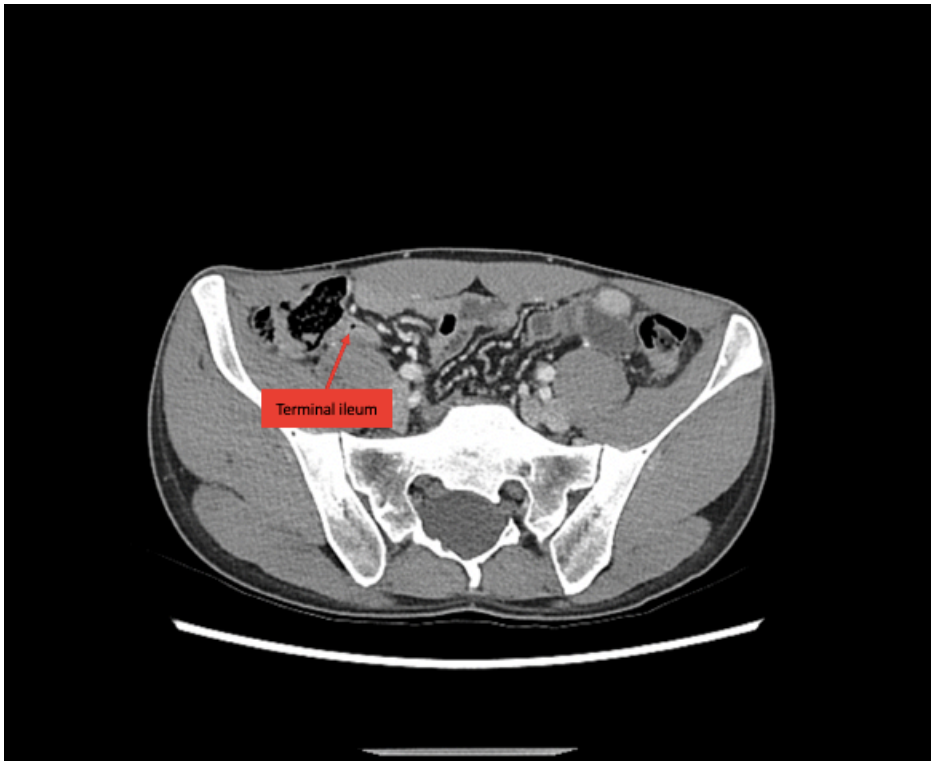

Figure B

Brain

In Figures C-E, the evaluated anatomy for the brain examination are shown.

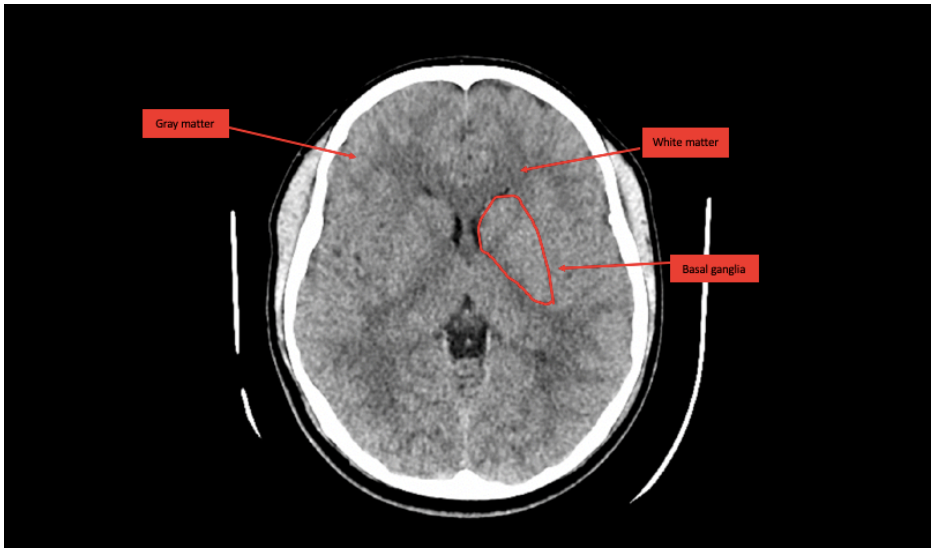

Figure C

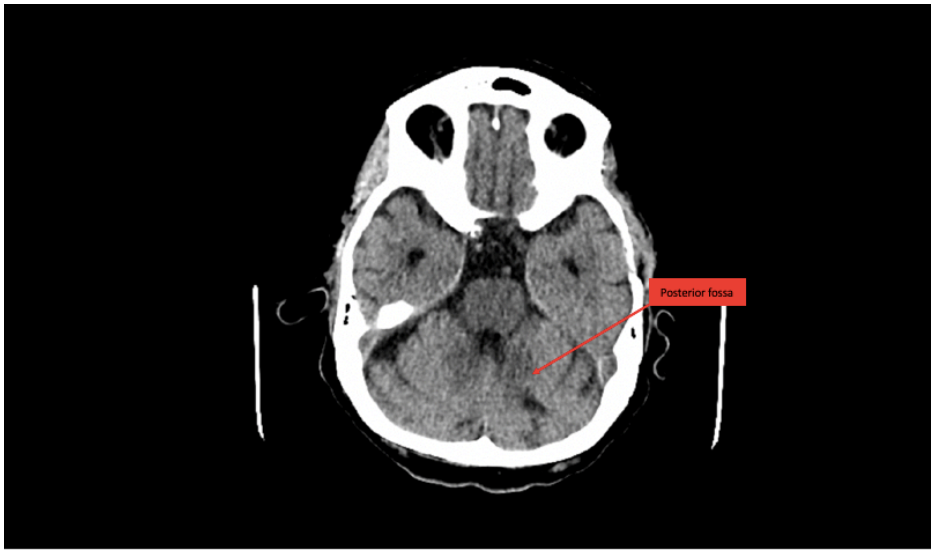

Figure D

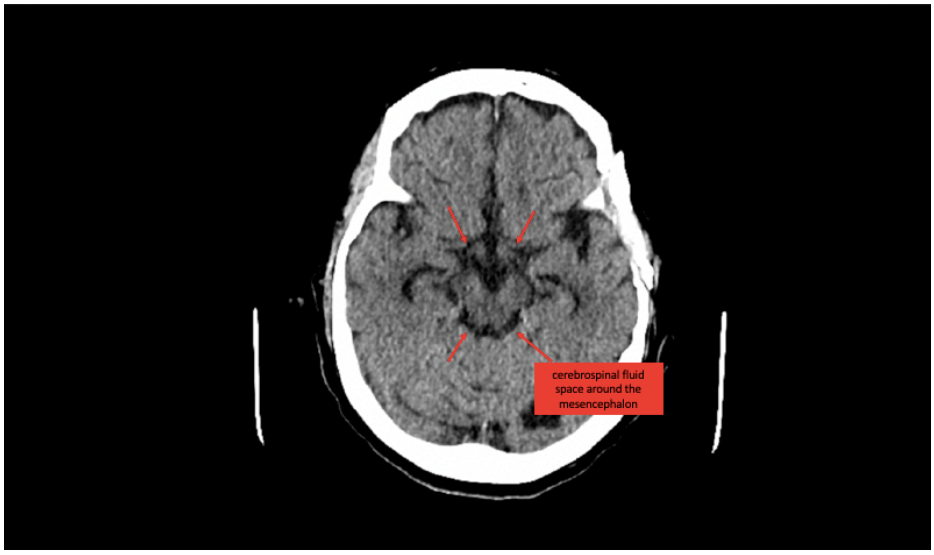

Figure E

Visual Grading Characteristics

Abdomen

In Figures F- J, the VGC curves for each image quality criteria from the evaluation of the abdominal examinations are shown.

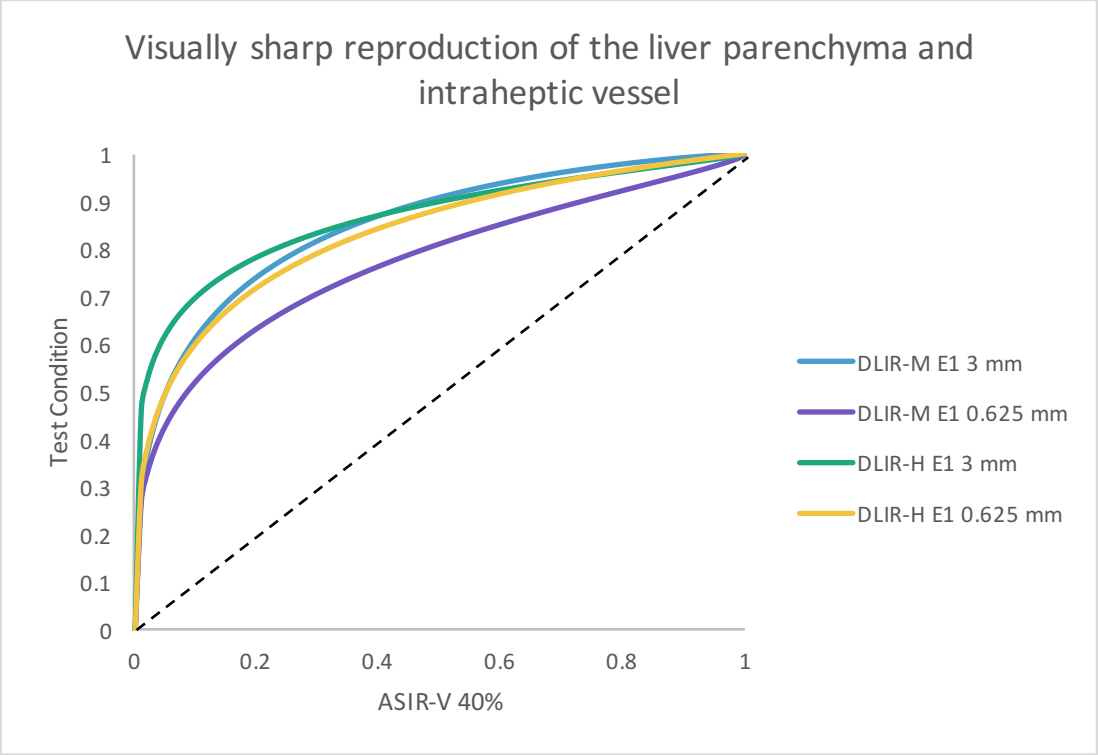

Figure F

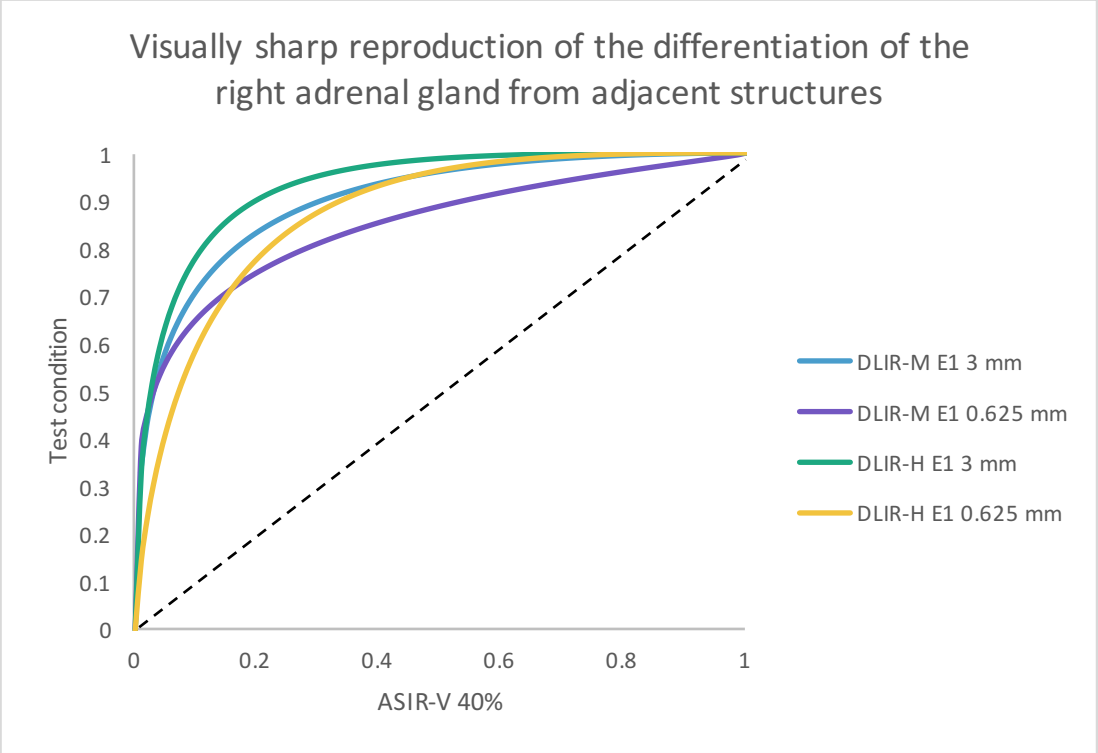

Figure G

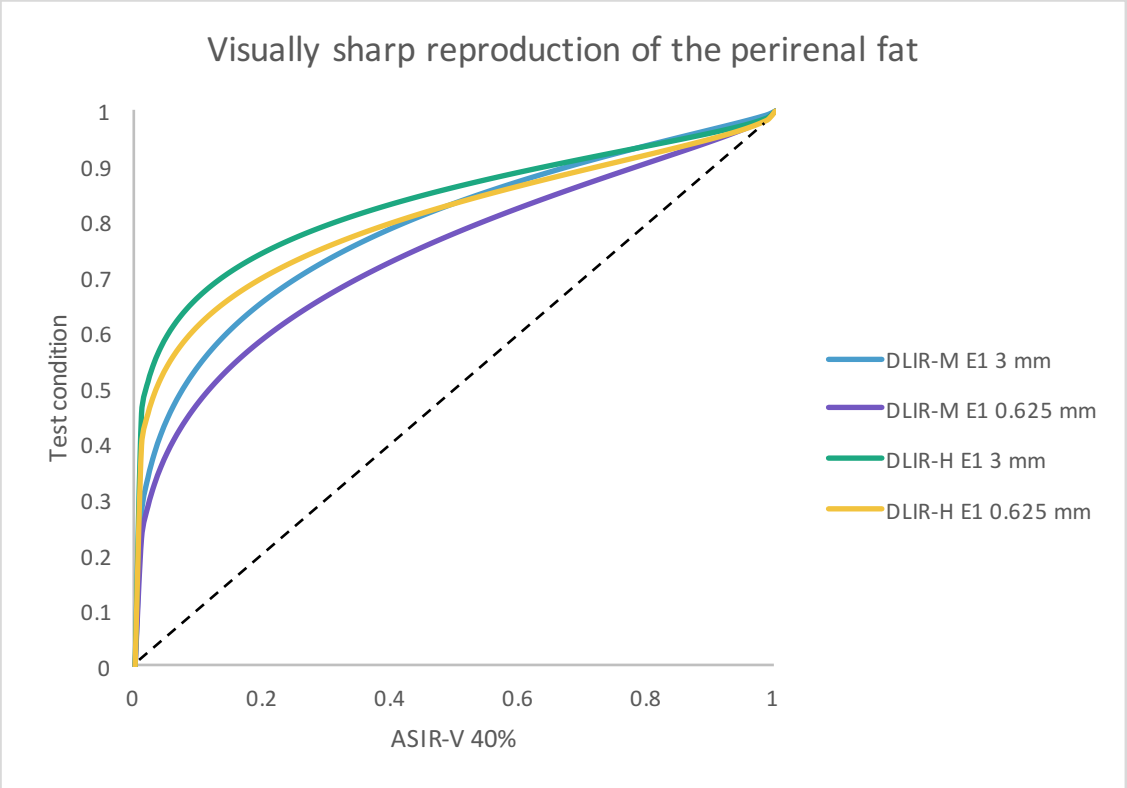

Figure H

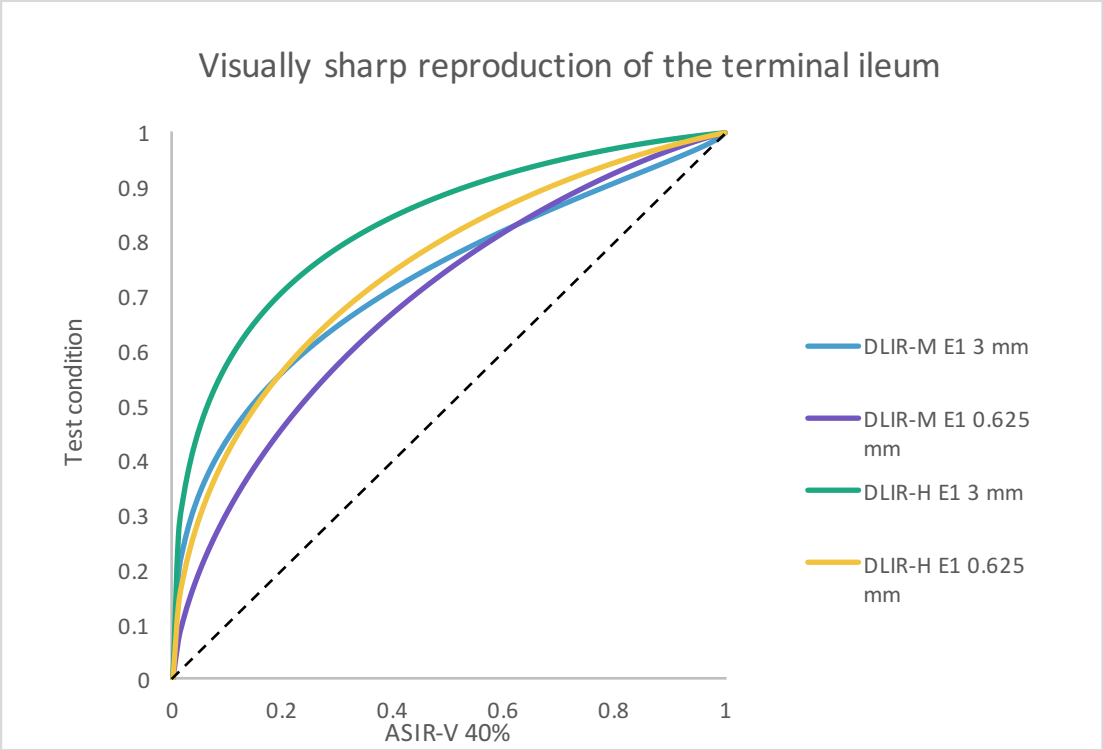

Figure I

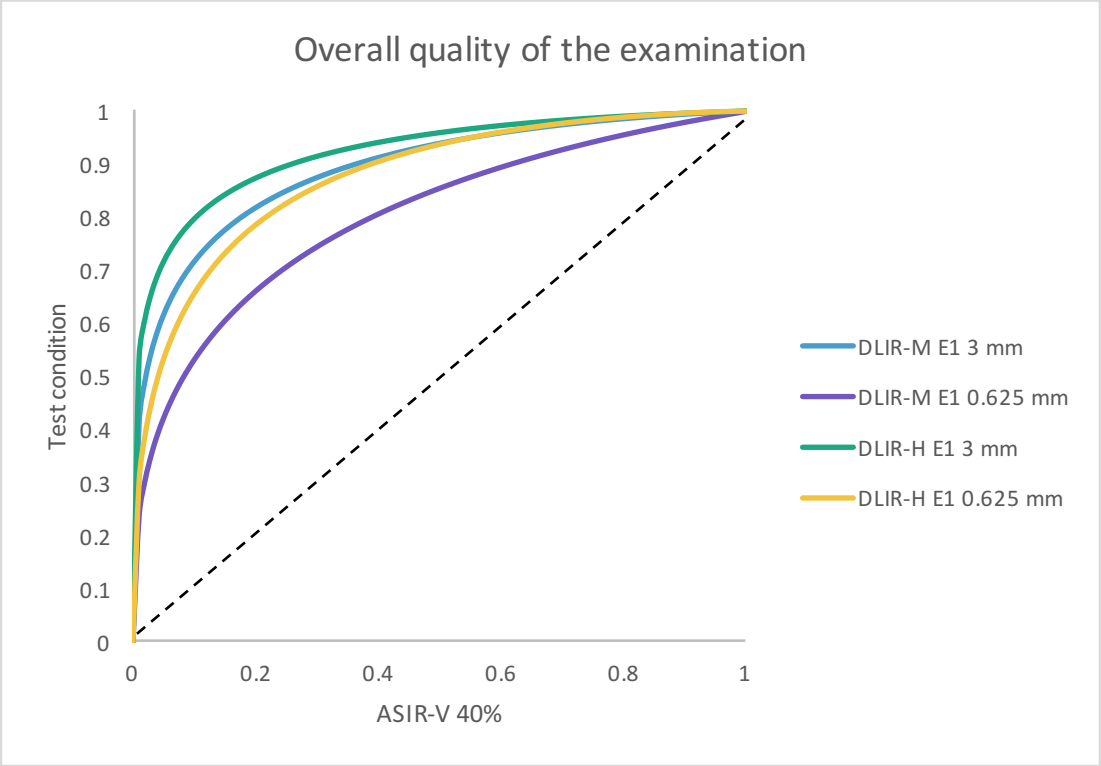

Figure J

Brain

In Figures K- O, the VGC curves for each image quality criteria from the evaluation of the brain examinations are shown.

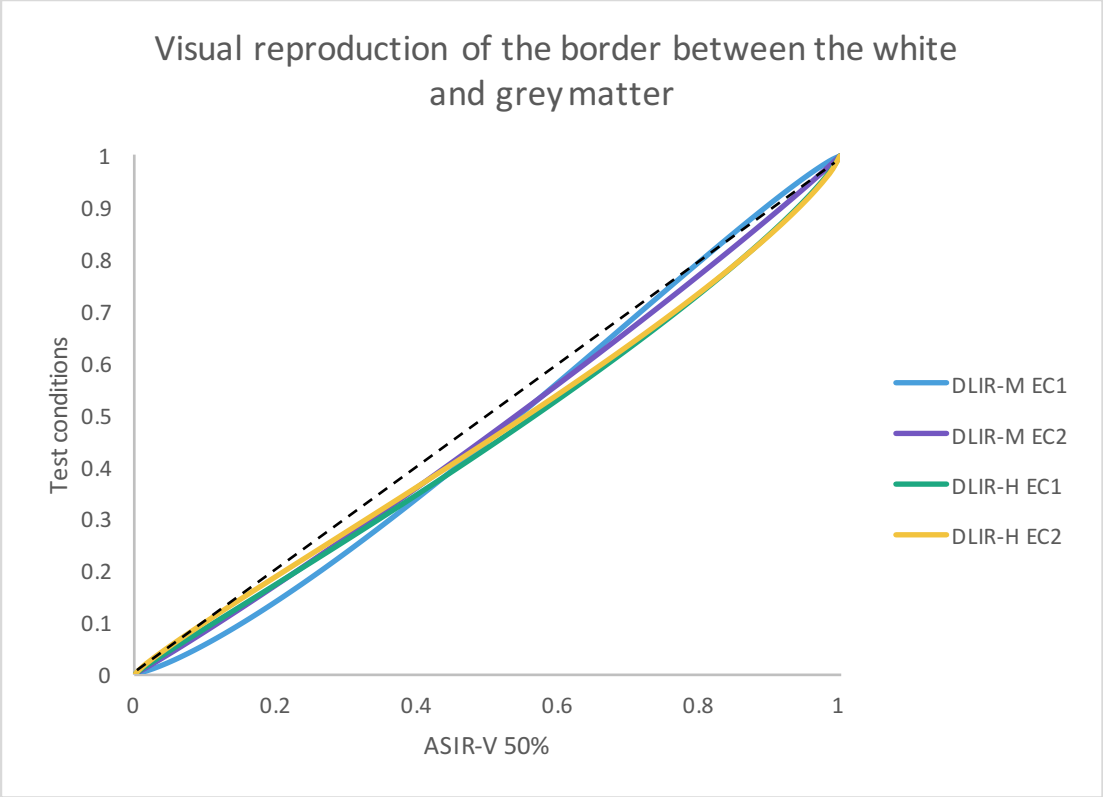

Figure K

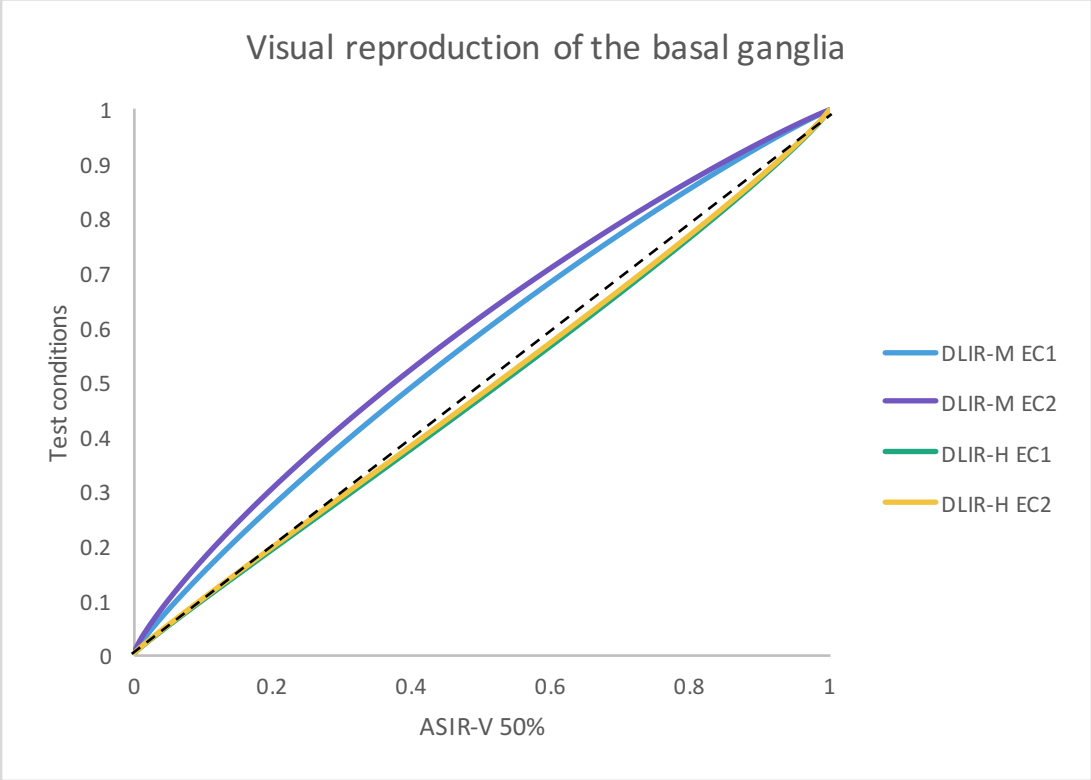

Figure L

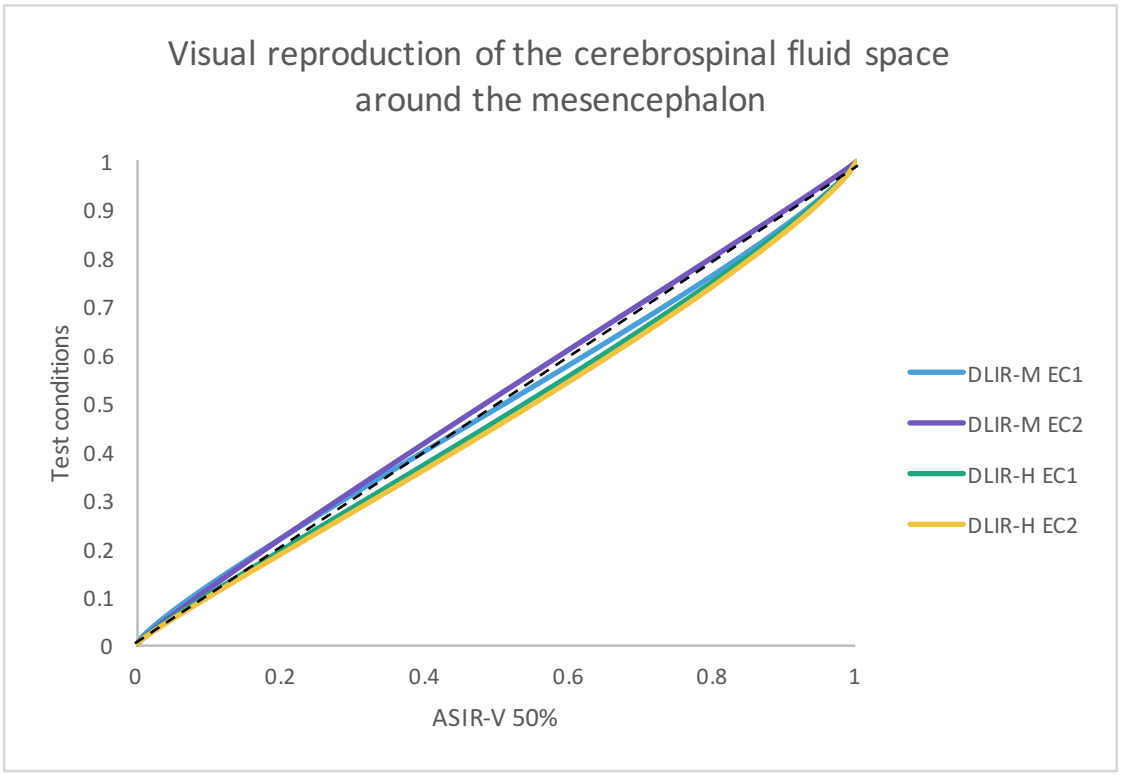

Figure M

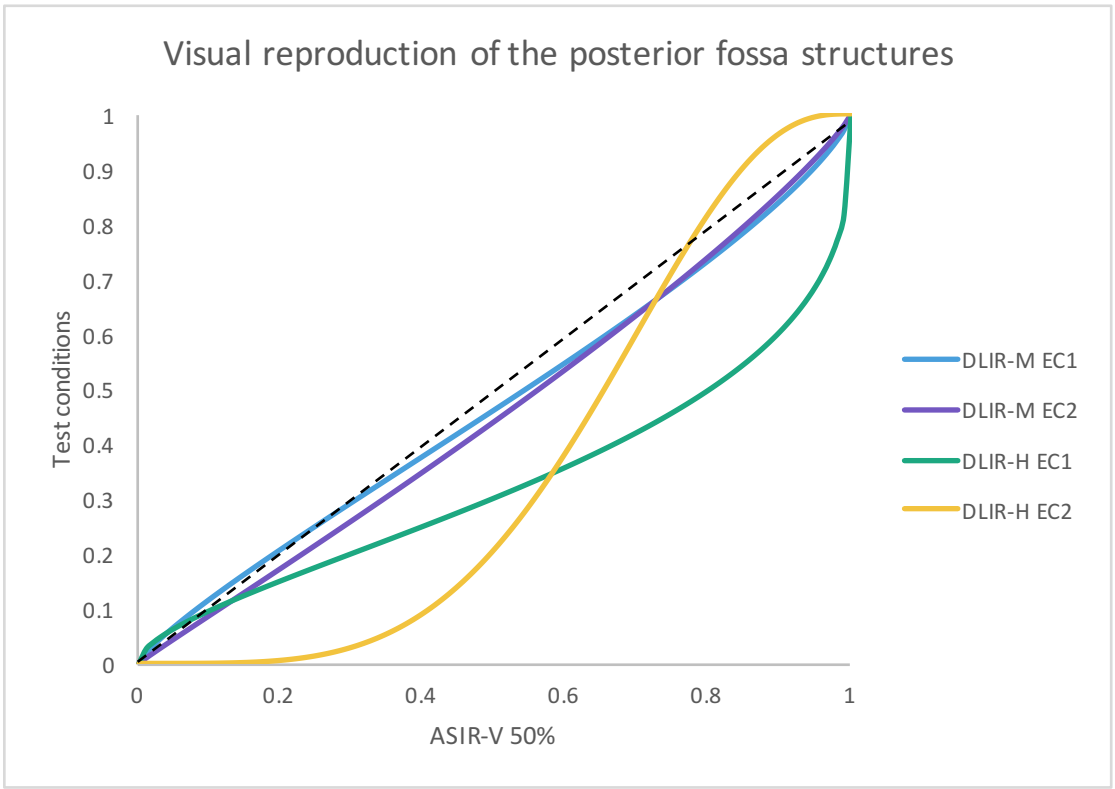

Figure N

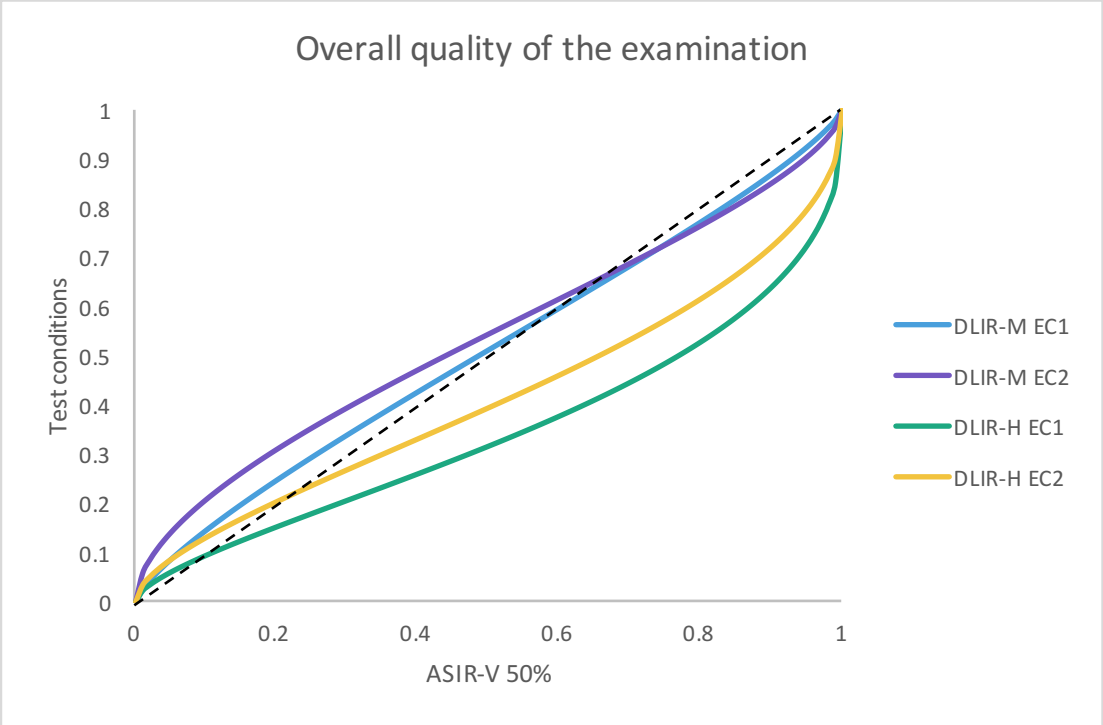
Overall quality of the examination
Test conditions
ASIR-V 50%
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8
DLIR-M EC1
DLIR-M EC2
DLIR-H EC1
DLIR-H EC2

Figure O