TASK-BASED OPTIMIZATION of ADMINISTERED ACTIVITY for PEDIATRIC RENAL SPECT IMAGING

by

Ye Li

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

October 2020

Abstract

Like any real-world problem, the design of an imaging system always requires tradeoffs. For medical imaging modalities using ionization radiation, a major tradeoff is between diagnostic image quality (IQ) and risk to the patient from absorbed dose (AD). In nuclear medicine, reducing the AD requires reducing the administered activity (AA). Lower AA to the patient can reduce risk and adverse effects, but can also result in reduced diagnostic image quality. Thus, ultimately, it is desirable to use the lowest AA that gives sufficient image quality for accurate clinical diagnosis.

In this dissertation, we proposed and developed tools for a general framework for optimizing RD with task-based assessment of IQ. Here, IQ is defined as an objective measure of the user performing the diagnostic task that the images were acquired to answer. To investigate IQ as a function of renal defect detectability, we have developed a projection image database modeling imaging of ^{99m}Tc-DMSA, a renal function agent. The database uses a highly-realistic population of pediatric phantoms with anatomical and body morphological variations. Using the developed projection image database, we have explored patient factors that affect IQ and are currently in the process of determining relationships between IQ and AA in terms of these found factors. Our data have shown that factors that are more local to the target organ may be more robust than weight for estimating the AA needed to provide a constant IQ across a population of patients. In the case of renal imaging, we have discovered that girth is more robust than weight (currently used in clinical practice) in predicting AA needed to provide a desired IQ. In addition to exploring the patient factors, we also did some work on improving the task simulating capability for anthropomorphic model observer. We proposed a deep learning-based anthropomorphic model observer to fully and efficiently (in terms of both training data and computational cost) model the

clinical 3D detection task using multi-slice, multi-orientation image sets. The proposed model observer is important and could be readily adapted to model human observer performance on detection tasks for other imaging modalities such as PET, CT or MRI.

Primary Reader and Advisor: Eric C. Frey, Ph.D.

Secondary Reader: Yong Du, Ph.D.

ACKNOWLEDGEMENTS

There are many people who have contributed to the finish of my dissertation. In particular, I would like to thank Professor Eric Frey, my thesis advisor, who gave me the opportunity to work with him at the outset, arranged for my financial support, and continued to support me even during my toughest periods (due to my father's leave). He encouraged me to refine my research to higher and higher standards by providing steady, timely and insightful feedbacks. Thanks to Prof. Frey, I have finished the goals that I set for myself at the beginning of my Ph.D. program.

I would also like to thank the faculty and collogues at the Division of Medical Imaging Physics (DMIP). In particular, I would also like to thank Professor Yong Du for helping me get into the door of medical imaging research at the very beginning of my Ph.D. journey, and for proof-reading my dissertation at the end. I am also grateful to Profs. Ken Taguchi and Abhinav Jha, for their time devoted to the numerous individual meetings to help prepare for my Ph.D. oral exam. Through those discussions, I was deeply inspirited by their scientific rigor, insights and knowledge in the field which benefited me a lot. I am very thankful to Prof. Benjamin Tsui, whose advice had been very helpful not only in research, but also in how to become a great researcher. I appreciate the friendships and truly enjoyed working with every collogue at DMIP: Tao Feng (the questions), Nate Crookston (software bugs), George Fung (the advices), Jingyan Xu (the answers), Michael Ghaly (image simulation questions), Xin Li (chat), and Junyu Chen (the deep learning stuff). And thanks to Martin Stump for making the cluster always ready for me and the group to use, and to Debbie Race for helping with my administrative matters.

I would also like to express my gratitude to Profs. George Sgouros, Vishal Patel, and Archana Venkataraman, for serving in my dissertation committees, and to Profs. Jerry Prince,

Laurent Younes, Ken Taguchi, Alan Yuille, John Goutsias, and Arman Rahmim for serving in my Graduate Board Oral committee. In particular, I would like to thank Prof. Archana Venkataraman for providing helpful feedbacks to improve the dissertation.

I wish to acknowledge the funding agency that has supported me and my research. The research has been supported by the National Institute of Biomedical imaging and Bioengineering under grant number R01–EB013558.

Finally, I would like to give my deepest thanks to my parents and wife. It was through their hard work and sacrifice that I was able to embark on this challenge, and through their love and unconditioned support that I was able to complete it.

DEDICATION

Dedicated to my wife, my great mother, and father in heaven.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

GLOSSARY

2AFC	Two-alternative forced-choice
AD	Absorbed dose
AUC	Area under the curve
BKE	Background-known-exactly
BKS	Background-known-statistically
CAD	Computer-aided diagnosis
CDRF	Collimator detector response function
CHO	Channelized Hotelling observer
CNN	Convolutional neural network
CNR	Contrast-to-noise ratio
COV	Coefficient of variation
CV	Cross-validation
DeepAMO	Deep-learning -based anthropomorphic model observer
DI	Detectability index
DMSA	Dimercaptosuccinic acid
FBP	Filtered backprojection
FWHM	Full width at half maximum
IQ	Image quality
LEUHR	Low-energy-ultra high-resolution
MDN	Mixture Density Network

MLP	Multilayer perceptron
MRI	Magnetic resonance imaging
MTCLDO	Multi-template channelized linear discriminant observer
MVN	Multivariate normally
OS-EM	Ordered-subsets expectation-maximization
PK	Pharmacokinetic
PSF	Point spread function
PSNR	Peak signal-to-noise ratio
RMSE	Root mean squared error
ROC	Receiver operating characteristic
ROI	Region of interest
SKE	Signal-known-exactly
SKS	Signal-known-statistically
SNR	Signal-to-noise ratio
SSIM	Structural similarity index

Chapter 1

Introduction

1.1 Significance

Like any real-world problem, the design of an imaging system always requires optimizing tradeoffs for a given task. For medical imaging modalities using ionizing radiation, a major tradeoff is between diagnostic image quality (IQ) and the risk to the patient from absorbed radiation dose. In nuclear medicine imaging, reducing the radiation dose to the patient will always increase the Poisson noise in the image, which may result in decreased IQ, resulting in unreliable images and even diagnostic errors. However, reducing the radiation dose (RD) on the other hand will always decrease the risk of adverse effects to the patient. Thus, it is critically important to use the "just right" amount of RD for each individual patient that maximizes diagnostic benefits while maintaining minimum adverse effects to the patient. This need for children patients is more pressing as they are more vulnerable to radiation than adults.

In nuclear medicine, reducing RD is achieved through the reduction of the administered activity (AA). In current clinical practice, AA for pediatric molecular imaging is often based on the North American consensus guidelines (U.S.) and the European pediatric dosage card (Europe). Both of these dosing guidelines involve scaling the adult AA by patient weight, which subject to upper and lower constraints on the AA. However, these guidelines were developed based on expert consensus or rough estimations of IQ (estimated count rates) rather than rigorous, objective

measures of performance on the diagnostic task. Accurate quantification of IQ plays an important role in the IQ-RD tradeoff analysis. However, acquiring an accurate measure of IQ is not easy as it is not only dependent on AA but also many other factors such as the imaging system, patient body morphometry, reconstruction and compensation methods and post-reconstruction processing, etc. The overall goal of this research is to pin down the most significant factors that affect IO in renal nuclear medicine imaging and to develop a rigorous and comprehensive IQ-RD tradeoff analysis framework applicable to all medical imaging modalities using ionizing radiation. The results of this study will provide information for standards bodies to improve current dosing guidelines for pediatric molecular imaging that result in more consistent IQ and absorbed dose

1.2 Organization

This dissertation is organized as follows.

Chapter 1 states the significance and organization of this dissertation.

Chapter 2 reviews the background of this work, previous researches in this area, and the technologies that underline it. First, this chapter describes the clinical imaging procedure, including the imaging modality and the chemical tracer used, that this study is aimed to optimize. Secondly, the method of task-based image quality assessment is discussed and existing model observers are reviewed. Finally, a brief description about convolutional neural network is provided at the end of the chapter.

Chapter 3 is a peer-reviewed journal publication which describes the initial work of developing a SPECT projection image database for a population of pediatric patients. The chapter outlines the methods for image simulation that were used throughout this research [1].

Chapter 4 is a peer-reviewed journal publication that describes our findings about a new external body parameter, which has better performance than weight (currently in clinical use) on predicting the AA needed to provide a desired IQ in renal nuclear medicine imaging [2]. Results show that factors that are more local to the target organ may be more robust than weight for estimating the AA needed to provide a constant IQ across a population of patients. We found that in the case of renal imaging, girth is more robust than weight in predicting AA needed to provide a desired IQ.

Chapter 5 is a submitted manuscript that describes the design of a deep learning based anthropomorphic model observer (DeepAMO) that can be used for clinically realistic visual detection tasks performed on volume images. Results show that the proposed model observer has the potential to mimic human observer in performing defect detection task in a clinically realistic diagnostic setting.

Chapter 6 summarizes the novel findings of this dissertation, highlights the significance and importance of this work to medical imaging system optimization and suggests potential avenues of future research.

Chapter 2

Background

2.1 Nuclear medicine imaging

2.1.1 Introduction to nuclear medicine imaging

Nuclear medicine (NM) imaging is a branch of functional imaging that encompasses two main modalities – single-photon imaging, including planar scintigraphy and Single Photon Emission Computed Tomography (SPECT), and Positron Emission Tomography (PET) – which use small amounts of radioactive materials called radiotracers to provide in vivo imaging of functional and physiological processes of the human body [3]. It is distinguished from modalities such as X-ray planar radiography that principally depicts the body's structure (anatomy). A radiotracer is a chemical compound where one or more of the atoms is a radionuclide. By nature of its biological and biochemical properties, the radiotracers can be used to explore the mechanism of physiological processes in vivo, such as glucose metabolism, by tracing the gamma photons emitted through the decay of the radioisotope. In single-photon imaging, a gamma camera is used to detect the gamma photons and produces a set of 2D projection data of the 3D radiotracer distribution within the patient body. In SPECT, projections are acquired at several views around the body, and the resulting set of projection data is reconstructed to provide 3D volume image of the activity distribution in the body.

2.1.2 Radiation dose and image quality tradeoff in pediatric nuclear medicine imaging

For medical imaging with ionizing radiation, there is always a tradeoff between image quality (IQ) and risk to the patient from absorbed dose (AD). In the dose range relevant to most nuclear medicine studies (below 10 mSv), the patient risks are specifically referring to a low-probability risk of inducing caner (from stochastic effects) in the patient later in life. Children are thought to be at a higher risk of certain adverse effects from radiation exposure than adults owing to the enhanced radiosensitvity of their tissues and the longer time-period over which stochastic radiation effects may manifest [4]. Thus, it is particularly important to expose pediatric patients to as low a radiation dose as is commensurate with providing sufficient diagnostic information [5].

In nuclear medicine imaging, reducing the AD requires reducing the administered activity (AA). Lower AA results in increased Poisson noise (introduced in section 2.4.1) in the images or requires longer acquisition durations to maintain the noise level. Thus, finding the optimal AD, i.e., the one giving the lowest risk sufficient to provide acceptable diagnostic image quality, comes down to finding the lowest AA that gives sufficient IQ for clinical diagnosis or other relevant tasks.

The fundamental differences that separate pediatric nuclear medicine from adult nuclear medicine are that children generally have smaller organs and lesions [3] and are thought to be at a higher risk for adverse effects from radiation exposure than adults [4]. Therefore, special considerations are needed for imaging children. First, higher resolution images are needed in pediatric nuclear medicine in order to detect these smaller organs or lesions [3], as detectability of a lesion is fundamentally limited by the lesion size with respect to the resolution of the imaging system. Second, sedation is often required for children of young age and, especially for longer acquisitions. Longer acquisition durations increase the chance of patient motion, which can

degrade image quality. Short acquisition durations are thus desirable. Lastly, AA for pediatric nuclear medicine needs to be carefully optimized due to the potentially higher radiation-induced cancer risk of children. A detailed discussion about dose sensitivity in children is given in section 2.2.2.

2.2 Radiopharmaceutical dosing in nuclear medicine imaging

2.2.1 Relationship between administered activity and radiation dose

In nuclear medicine imaging, the total AD to a patient is proportional to the amount of administered activity injected to the patient. Here, the absorbed dose is defined as the mean energy deposited to the target tissue (or region) per unit tissue mass. According to the Medical Internal Radiation Dose (MIRD) schema [6], the absorbed dose to a target region of interest is computed as follows:

$$D(r_T, T_D) = \sum_{r_S} \tilde{A}(r_s, T_D) S(r_T \leftarrow r_s), \quad (2.1)$$

where $D(r_T, T_D)$ is the mean absorbed dose to a target region of interest r_T over a dose integration period T_D; $\tilde{A}(r_s, T_D)$ is the time integral of activity in the source region r_s; and $S(r_T \leftarrow r_s)$ is the so called "S-value" in the field of dosimetry, which represents the absorbed dose per unit time-integrated activity. The S-value depends on the average energy and abundance of the particle or particles emitted during the decay, the fraction of the energy that is absorbed in the target region, and the mass of the target region.

2.2.2 Dose sensitivity in children

Children are often thought to be more sensitive to adverse effects from radiation exposure than adults [7]. This is mainly because children have: (1) more tissues with high mitotic rates, which are more vulnerable than tissues with lower mitotic rates to radiation [8], and (2) longer post-exposure lifespans to manifest these stochastic radiation effects [5]. In infancy and early childhood, these considerations become even more pressing, as cells are growing (undergoing high rates of division) and differentiating into mature cells, and thus are more vulnerable to ionizing radiation [9]. Although the cells attempt to repair themselves when they are damaged (mostly in the form of DNA breaks), very rarely, however, mistakes do happen in the DNA repair process, resulting in genetic abnormalities (mutations) [10, 11]. Therefore, there has been significant interest in the nuclear medicine community in establishing universally accepted and optimized dosing guidelines for pediatric nuclear medicine studies.

2.2.3 Current dosing guidelines for pediatric nuclear medicine imaging and limitations

To address the dosing of pediatric patients, the European Association of Nuclear Medicine (EANM) and Society of Nuclear Medicine and Molecular Imaging (SNMMI) have published, respectively, the European pediatric dosage card and the North American consensus guidelines for pediatric administered activity (AA) [12, 13]. However, these guidelines were developed either based on a consensus of best practices or a simple estimation of image quality instead of a rigorous evaluation of diagnostic image quality relative to AA. A comprehensive introduction to the dosing

guidelines and their respective limitations is provided in Chapter 3 (terminology, such as count rate, effective dose, etc., is introduced in section 2.3).

2.3 Renal functional imaging with ^{99m}Tc-DMSA SPECT

2.3.1 Clinical problem

^{99m}Tc-DMSA is the agent of choice for renal cortical imaging by planar, pinhole scintigraphy, or by SPECT [3]. The DMSA tracer is principally concentrated (1 hour or more after injection) in the proximal convoluted tubules of the kidneys, which is ideal for detecting cortical functional defects in pyelonephritis, infarction, scarring, duplication, and fetal lobations [3]. Fig. 2.1 shows an example of SPECT image in pyelonephritis. The dim area, as indicated by the arrow, shows the non-functional regions of the cortex.

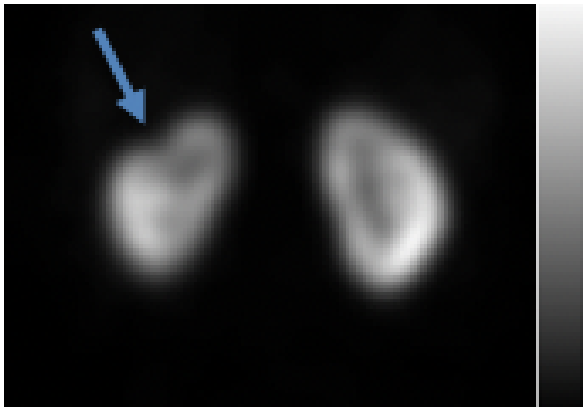

Figure 2.1. Example of a clinical SPECT image in pyelonephritis of a 16-year-old reconstructed using two iterations of eight subsets of the OS-EM reconstruction with detector response compensation followed by a Gaussian filter with a FWHM of 0.5 mm.

2.3.2 Use of DMSA in diagnosing acute pyelonephritis

2.3.2.1 ^{99m}Tc-DMSA Characteristics

^{99m}Tc is one of the most commonly used medical radioisotopes in nuclear medicine imaging. It has a half-life of 6 hours [14] (93.7% of it decays in 24 hours) and emits pure gamma rays with a single photon energy of 140 keV. This energy is high enough that the photons leave the body, but low enough that the photons can be detected and collimated with relative ease. ^{99m}Tc labeled Dimercaptosuccinic acid (DMSA) was first developed by Lin et al. [15] as a renal imaging agent to replace the ^{197}Hg-labeled chlormerodrin because of the poor imaging characteristics of ^{197}Hg, and the toxicity of mercury [16]. DMSA has a high absolute renal concentration, about twice that of the other ^{99m}Tc labeled compounds in humans, approaching the concentration of labeled chlormerodrin [17]. The physical characteristics of ^{99m}Tc and the mercurial-like kinetics of the chelator make this compound a unique agent for imaging the renal parenchyma in patients of all ages [18]. Moreover, DMSA has a high uptake in the renal cortex, with about 50% remaining there at 1 hour, resulting in a high gamma flux, and is thus ideal for imaging [19]. ^{99m}Tc-DMSA is the most commonly used agent for renal cortical imaging in planar scintigraphy and SPECT [20].

2.3.2.2 Biokinetic behavior of ^{99m}Tc-DMSA

^{99m}Tc-DMSA is administered intravenously with a usual dose in adults of 0.05 mCi/kg (1.85 MBq/kg) [20]. After intravenous injection, this agent is 90 % bound to plasma proteins, and only a small amount (0-5%) is associated with red cells [19]. The clearance of ^{99m}Tc-DMSA in the blood follows a single exponential with a mean half-life of 56 minutes and with 6-9% of the administered dose present in the blood at 14 hours after injection [20]. The blood clearance of

^{99m}Tc-DMSA is very slow compared to other renal agents. The renal uptake of ^{99m}Tc-DMSA is approximately 40-50 % of the injected dose at 4 hours post-injection [19]. Most accumulated tracer is found in the proximal convoluted tubules, with small amounts elsewhere in the kidneys [21]. Although most ^{99m}Tc-DMSA is retained in the renal parenchyma, cumulative urinary excretion has been reported to be 6 % at 1 hour, 1-12 % at 2 hours, and 25 % at 14 hours [20].

According to the International Commission in Radiological Protection (ICRP) model [22], an intravenous injection of ^{99m}Tc-DMSA gives rise to an initial distribution in the extracellular fluid. About half of the material entering the extracellular fluid is deposited in the renal cortex and is retained there for a long time, and a further fraction is temporarily retained in the liver and spleen [22]. Excretion of ^{99m}Tc-DMSA is exclusively via the kidneys and could take up to 2 days [23].

2.3.3 Diagnostic task with DMSA SPECT

Often, the associated diagnostic task for DMSA renal imaging is to detect renal parenchymal defects or cortical functional defects. Thus, we have modeled the clinical task as a defect detection task. As mentioned in section 2.4.3.3, there are some practical limitations of the current model observers. To overcome these limitations, we have developed a deep learning-based anthropomorphic model observer to fully simulate clinical detection task for DMSA renal SPECT imaging. The new model observer will be introduced in Chapter 5. The next section provides a general background about deep convolutional neural networks and their applications as model observers in task-based image quality evaluation.

2.4 Image quality in nuclear medicine imaging

In nuclear medicine imaging, there are two fundamental methodologies for evaluating image quality [24]. The first is by means of physical characteristics that can be quantitatively measured for the image or imaging system. The second is by means of task-based image quality evaluation such as human observer studies (a detailed introduction is provided in section 2.4.3)

2.4.1 Physical characteristics of image quality

In order to understand the surrogate measures of image quality used in Chapter 3, it is essential to introduce the physical characteristics of image quality for nuclear medicine images.

In principle and among other factors, the quality of nuclear medicine images is mainly characterized by three factors: (1) spatial resolution (sharpness), (2) noise (variations in the image due to random effects such as quantum noise), and (3) contrast (difference in image intensity between areas of the imaged object). Other factors such as artifacts, non-uniformity or distortions, and patient or organ motion can also affect image quality but will be largely neglected in the following discussion. Although resolution, noise, and contrast describe three different aspects of image quality, they cannot be treated as completely independent parameters: improvement in one is frequently obtained at the expense of deteriorating the others [24]. For example, in nuclear medicine, reduced image noise can be obtained by the use of a higher sensitivity collimator. However, there is an inverse relationship between sensitivity and resolution, thus reducing noise via the use of a high sensitivity collimator results in poorer spatial resolution. Poor spatial resolution will result in poorer contrast of small objects.

Spatial resolution refers to the ability of an imaging system to separate fine details in the image [25]. There are two main components that contribute to the lack of details or sharpness in

nuclear medicine: geometric resolution and intrinsic resolution (detailed introduction is provided in section 2.4.3.1). In theory, the higher the spatial resolution of the imaging system the shaper the image it can produce. However, there is a tradeoff between spatial resolution and noise. For example, improved collimator resolution results in decreased collimator efficiency, and, hence, decreased counting rates and increased image statistical noise for the same acquisition duration and administered activity.

In nuclear medicine, noise most often refers statistical fluctuations in the recorded counts that result from the random nature of radiation decay and photon counting statistics. These fluctuations can be described using the Poisson distribution:

$$P(N = n|m) = \frac{m^n e^{-m}}{n!}, \tag{2.2}$$

where m is the mean number of detected counts in a projection bin and n is the number recorded counts for one particular acquisition. The fact that the recorded counts can be different from the mean is referred to as Poisson noise. In projection data, the noise is uncorrelated Poisson noise whose variance $Var[N]$ equals its expectation $E[N]$. In general, the only way to reduce the amount of Poisson noise is to increase the mean number of counts. This can be seen by considering the coefficient of variation (COV), defined as the standard deviation divided by the mean, which describes the relative level of noise in a projection bin and is given by

$$COV = \frac{\sqrt{m}}{m} = \frac{1}{\sqrt{m}}, \tag{2.3}$$

which shows that Poisson noise, while growing in absolute terms with the signal, is relatively smaller at higher count levels.

Reconstruction and other image processing can alter the magnitude of the noise fluctuations and change the noise texture (by introducing correlations) depending on the algorithm used. Noise is a very important aspect of nuclear medicine. When the size of an object is substantially larger than the limiting spatial resolution of the imaging system, noise can still impair detectability, especially when noise fluctuations are large compared to the contrast of the object of interest.

Image contrast refers to the differences in counts or intensity in the object of interest compared to the background. In nuclear medicine, this difference is caused by the different levels of radioactive uptake in the patient [24]. For example, if R_b is the number of counts recorded in a background area and R_s is the number of counts recorded over a signal area (i.e., a lesion), the contrast of the signal is defined as [24]

$$C_s = \frac{R_s - R_b}{R_b}. \tag{2.4}$$

There are several factors that affect the contrast including intrinsic object uptake, scattering, and septal penetration. Among these factors, intrinsic object uptake is the major component that affects image contrast, and is largely determined by the radiopharmaceutical and patient biokinetics. The other two factors affect the image contrast primarily by adding counts to the background. The degraded image contrast with the added background R_0 can be expressed as

$$\begin{aligned} C_{s'} &= \frac{(R_s + R_0) - (R_b + R_0)}{R_b + R_0} \\ &= C_s \frac{1}{1 + \frac{R_0}{R_b}}. \end{aligned} \tag{2.5}$$

It can be seen in the above equation that the larger the additional factor R_0/R_b the more decrease would be seen in the contrast. Also, contrast is related to noise through the contrast-to-noise ratio (CNR), which is a critical parameter for detectability. The CNR is defined as the contrast divided by the noise. One way to characterize noise is the COV of the recorded counts in the same area where the contrast is measured. A high noise level will have a large denominator (COV) in the CNR and thus a smaller CNR, resulting in reduced detectability and less accurate diagnosis. In nuclear medicine, image contrast can be degraded by physical factors involved in image formation such as the effects of scattered photons from surrounding tissues, and septal penetration and scatter. A detailed introduction to the causes and effects of photon scatter and septal penetration is given in section 2.4.2.1and 2.4.2.2, respectively.

Although these physical measures set the fundamental limits for image quality in nuclear medicine imaging, they may not directly reflect the performance of an observer on a clinical task performed with those images. Clinically relevant image quality should be assessed with respect to the task that is to be performed [26-32].

2.4.2 Image quality in SPECT

2.4.2.1 SPECT image formation process

SPECT images originate from measurements of gamma photons emitted from the radiotracers distributed within the patient body. These photons are recorded by a gamma camera that is rotated around the patient to form multiple 2D images (also called projections), from different projection views. Then these projections are reconstructed to form a 3D image (of the radiotracer distribution of the body) using a tomographic reconstruction algorithm.

The process of projection acquisition is represented mathematically as follows

$$\mathbf{g} = \mathbf{Hf} + \boldsymbol{n}, \tag{2.6}$$

where $\mathbf{f} = [f_1, f_2, \dots, f_N]^{\boldsymbol{T}} \in \mathbb{R}^{\mathrm{N}\times 1}$ is the voxelized object being imaged (the continuous 3D radiotracer distribution of the body), $\mathbf{g} = [g_1, g_2, \dots, g_M]^{\boldsymbol{T}} \in \mathbb{R}^{\mathrm{M}\times 1}$ is the projection image, and $\mathbf{H} \in \mathbb{R}^{\mathrm{M}\times\mathrm{N}}$ is the imaging operator, which maps the activity distribution to the measured projections, and $\mathbf{n} \in \mathbb{R}^{\mathrm{M}\times 1}$ is the Poisson noise (introduced in section 2.4.1) resulting from the random nature of radioactive decay and interactions of the emitted photons with the patient and detection system. Specifically, **H** is a matrix characterizing all the image degrading factors in the image formation process which includes the attenuation and scatter in patient and the collimator-detector blurring (introduced in section 2.4.3), with $\mathbf{H}_{ij}$ representing the probability of a photon emitted in the image voxel j to be detected in the projection bin i. Pictorially, the image formation process is illustrated in Fig. 2.2.

The relative noise in the projections increases as a result of attenuation in the body due to decreased count rate (discussed in section 2.4.1), at a fixed AA and acquisition time. Attenuation is caused by the interaction of photons within the body (e.g., the photoelectric effect and Compton scattering), and leads to a depth-dependent reduction in the number of primary (unscattered) photon counts detected in the projection image. The amount of attenuation is dependent on the composition (e.g., atomic number and density), the energy of the photons, and the thickness of the absorber.

In theory and among other factors, the spatial resolution in SPECT (both axial and in-plane) is determined largely by the collimator resolution (the most important component of which is the geometric component introduced in section 2.4.2.2) and the intrinsic resolution of the gamma camera [24]. The further away the camera is from the patient, the worse the resolution. Thus, it is

desirable to use a body contour orbit that places the camera as close to the patient as possible at each projection view to acquire high-resolution image of the patient.

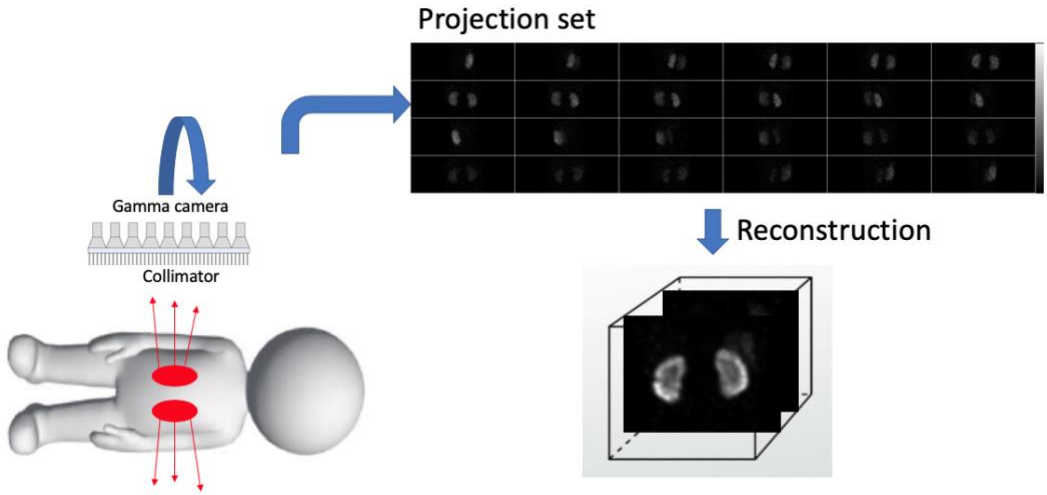

Figure 2.2. Illustration of the image formation process of SPECT

2.4.2.2 Physical factors of image quality in SPECT

In section 2.4.1, we discussed the physical measures that affect image quality in nuclear medicine imaging. In this section, we focus on discussing the factors that affect these physical measures in SPECT renal imaging in particular.

In the projection domain, the PSF (introduced in section 2.4.1) that describes the spatial resolution of a source in air is simply the collimator detector response function (CDRF), which is the image generated from a point source of activity. In the presence of a medium (i.e., patient), the PSF is also affected by the attenuation and scatter in the medium in addition to the CDRF. Since the CDRF various with position, the PSF is spatially varying. In a patient, the PSF is affected by scatter and attenuation, and is thus patient-dependent. In the reconstruction domain, the camera

orbit and the reconstruction algorithm can also affect the PSF in addition to the CDRF. The CDRF includes the intrinsic, geometric, septal scatter, and septal penetration response components. Since the geometric component is distance-dependent, the CDRF is spatially varying. The intrinsic response is due to the uncertainty of position estimation in the camera's detector system and the effects of scattering in the detector crystal. The geometric response accounts for all the photons from the point source that travels through the collimator holes without any interactions with the collimator septa and are detected in the acquisition energy window. The septal scatter and penetration components account for all the other photons that interact with or pass through the septa, respectively. For the 140 keV photons emitted by the agents used in renal imaging, septal penetration and scatter have relatively small effects on the CDRF.

The relative noise level in SPECT, as measured by the COV (equation 2.3), is inversely related to the number of detected photons received by the gamma camera. The number of detected photons is largely determined by the activity in the target organ, which in turn depends on the following factors: (1) fraction of the AA taken up in the target organ; (2) the size of the organ; and (3) the amount of attenuating medium between the organ and collimator of the scanner. Attenuation refers to the loss of photons emitted from a source as they travel toward the detector due to interactions with the body. Attenuation results in a depth-dependent reduction in the number of detected primary photon counts as compared to the same source in the air. The amount of attenuation depends on the energy of the photons and the composition and thickness of the absorber.

Scatter is another major effect that affects image quality (mostly contrast) during SPECT image formation. Scatter refers to scatter interactions (mostly Compton) in the patient. A scattered photon is a photon detected after it has undergone these scatter interactions. These scatter

interactions result in a change in the direction of the photons, and thus a loss in correlation between photon direction and the position of emission. These scattered photons can pass through the collimator and be detected, providing false position information and distortion of the estimated activity distribution. Scatter can cause the image to lose contrast by adding a low-frequency background of the image. To reduce the number of scattered photons counted in the projection image, energy discrimination is used to reject scattered photons. In this method, an acquisition energy is set around the energy window, centered on the energy of the gamma photon being imaged, and photons incident with energies outside this range are not counted. The scatter rejection from this method is imperfect due to the imperfect energy resolution of gamma cameras.

2.4.2.3 Radiopharmaceutical

The intrinsic contrast of a target object is determined by the activity uptake concentration of the target object relative to its surrounding tissue. The intrinsic contrast defines the upper limit of contrast that the imaging system can obtain for the target object, in a noise-free scenario with a perfect system PSF. When the target object is a defect, the intrinsic contrast of the defect affects the detectability of the defect, which would in turn affect image quality.

2.4.2.4 Patient factors

Besides system parameters, patient factors also have an important effect on image quality in SPECT. There are mainly three factors: (1) patient body morphometry local to the target organ; (2) patient target organ uptake; and (3) defect size versus imaging system resolution.

Patient body morphometry can have a direct impact on defect detectability. As an example, we found that patients with large girth at the location of the kidney can have lower image quality, as measured by defect detectability, than those with small girth [2]. The difference in image quality was due to a combination of three factors: the large girth patient resulted in fewer photons escaping the body, required a larger average camera orbit radius resulting in poorer spatial resolution, and produced a higher scatter-to-primary ratio, resulting in higher noise, poorer resolution, and poorer contrast, respectively, for the patient with a larger girth.

In addition to patient body morphometry, the amount of tracer uptake specific to an individual patient can also affect defect detectability, as previously explained in section 2.3.3.3.

Lastly, the defect size of the patient, which is unknown before imaging, can also affect defect detectability. To see details of a small defect with reasonable contrast, the spatial resolution must be better than the object size. Thus, it is generally preferred to use high-resolution collimators to image pediatric patients as children generally have smaller organs, and smaller defects are more clinically significant than adults.

2.4.3 Task-based image quality

2.4.3.1 Introduction to task-based image quality assessment

As described above, the quality of a medical image can be measured in terms of physical characteristics of the image, such as image contrast, spatial resolution, and noise [33] using various physical metrics. Alternatively, fidelity-based measures such as root mean squared error (RMSE), peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM), which evaluate image quality in terms of similarity of the image with respect to the imaged object, have also been widely

used in the medical imaging community. Fidelity-based measures are appealing because they are relatively easy to compute, have straightforward physical interpretations, and can provide objective quantitative measures of image quality. However, neither the physical nor fidelity-based measures are directly related to performance on the diagnostic task that will be performed with the images and thus may not be clinically relevant [31]. To be clinically relevant, image quality should be assessed with respect to a specific task that will be performed with the images [26-32], i.e., detect a tumor or estimate tumor volume. Assessing image quality objectively in terms of performance on a specific clinical task is called task-based image quality assessment.

Typically, the task is performed by an observer, and the figure of merit for image quality is the performance of the observer on the task. In the vast majority of clinical tasks, the observers are humans, and thus the observers used in the assessment should be drawn from the population of people performing the task, i.e., for medical images, a radiologist or nuclear medicine physician.

2.4.3.2 Human observers

Humans serve as observers or expert readers in the vast majority of medical imaging applications in task-based image quality assessment studies [31]. Human observers are the most relevant in assessment of the images used by human observers to perform a task [31]. However, in practice, the use of human observers (and especially physicians) is practically extremely challenging and expensive, especially in large-scale developmental research studies. Furthermore, human observers exhibit a significant amount of intra-observer and inter-observer variability in performance [33]. Thus, models of human observers (anthropomorphic model observers) have been widely used as surrogates for human observers. A great deal of effort has gone into the development of anthropomorphic model observers that predict human observer performance [34-

37].

In the context of a clinical defect detection task, the human observer is a radiologist or a nuclear medicine physician. However, the task to be performed in a human-observer study is slightly different from the routine clinical diagnostic task, i.e., classifying the patient as abnormal or diseased based on the image. For defect detection tasks, two-alternative forced-choice (2AFC) and continuous rating scale methods have been widely used in human-observer studies to allow measurement of a figure-of-merit for the observer's performance [38, 39]. In the majority of the human-observer studies, the input to the human observer is either a single image (i.e., a short-axis slice), a stack of slices from a specific orientation, or a set of images from 3 orthogonal (e.g., transaxial, sagittal, and coronal) orientations. The output from the human observer is a rating value, which represents the confidence that the observer thinks that there is a defect present in the image. Fig. 2.3 shows an example of a human-observer study display window using a continuous rating scale for a defect detection task.

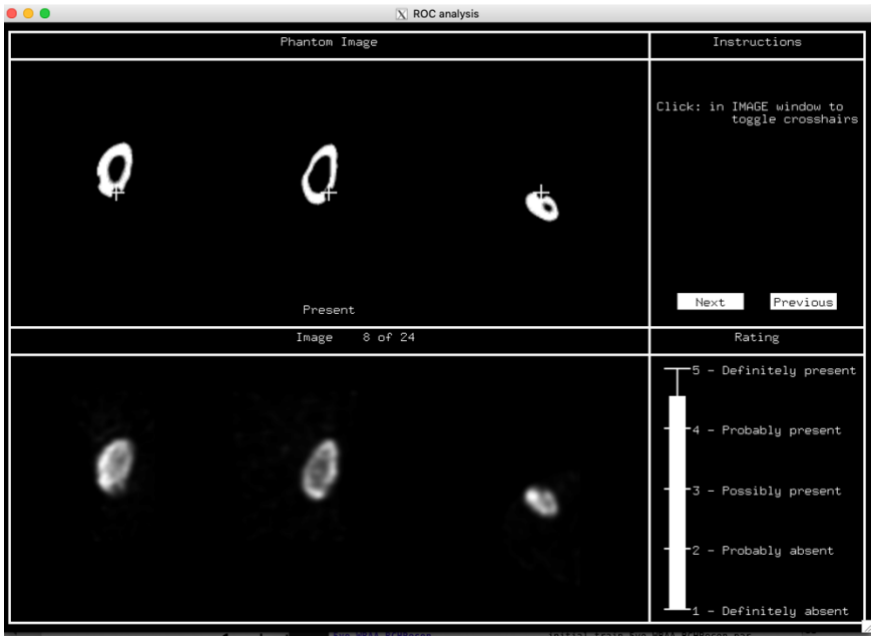

Figure 2.3. Example of a human-observer study display window. The image is displayed in the bottom lower corner; the instructions are in the top right corner; the continuous rating scale is in the bottom right corner. Cross-hair indicates a possible center of a defect to the observer. The study was for detecting renal cortical damage with DMSA SPECT images.

2.4.3.3 Model observers

Of the existing anthropomorphic observer models, the channelized Hotelling observer (CHO) has been the most widely used as a substitute for human observers in signal-location-known tasks in nuclear medicine imaging research [40]. The essential component that distinguishes a CHO from a Hotelling observer is the introduction of the concept of frequency channels. The channels are introduced for dimensionality reduction or to make the HO better model human observers [31]. There is widely accepted psychophysical evidence that, when visually processing an image, humans are sensitive only to the total power in a series of frequency bands or channels rather than to individual frequencies (infinitesimally small frequency bands) [41]. Thus, the entire frequency content of an image within a given frequency band or channel can be reduced to a single

output channel value. This process is often called channelization, which involves multiplying (in frequency domain) or taking dot products (in spatial domain) of the input image with a series of channel template images (shown in Fig. 2.4). The resulting scalar can be considered as the energy contained in the frequency channel. A total of N frequency channels results in an N-dimensional vector, often referred to as a feature vector [33]. For example,

$$\boldsymbol{v}_i = \boldsymbol{u}_i^t \boldsymbol{g}, \tag{2.7}$$

where $\boldsymbol{g}$ is the original input image, $\boldsymbol{u}_i$ is the ith channel template image, and $\boldsymbol{v}_i$ is the ith channel response. Stacking the channel responses together results in a feature vector $\boldsymbol{v}$,

$$\boldsymbol{v} = (v_1, v_2, v_3, v_4, \ldots v_L), \tag{2.8}$$

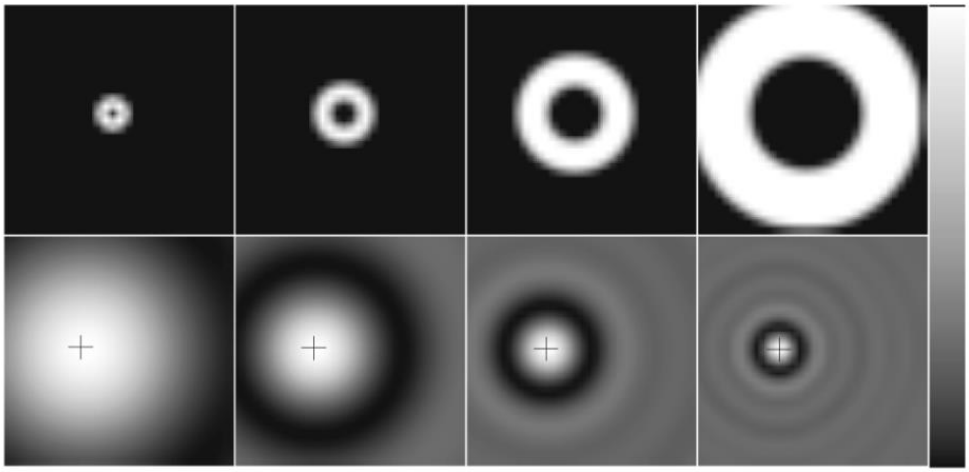

Figure 2.4. Images of 4 anthropomorphic difference-of-mesa [42] channels in the frequency-domain (top row) and their corresponding shifted spatial domain template images (bottom row). The cross-hair indicates the center of the template, which must be aligned with the center of the suspected defect location when taking the dot product.

For a binary detection task, the CHO test statistic λ is computed by taking the dot product of the CHO template and the channelized data vector $\boldsymbol{v}$ [43]:

$$\lambda = \boldsymbol{w}^t \boldsymbol{v}, \tag{2.9}$$

where $\boldsymbol{w}$ is the CHO template vector and the superscript t denotes the transpose operation. The CHO template is generated using the 1st and 2nd order statistics of the channelized data vector and is given by [43]:

$$\boldsymbol{w}^t = (\langle \boldsymbol{v}_1 \rangle - \langle \boldsymbol{v}_2 \rangle)^t \boldsymbol{K}_g^{-1}, \tag{2.10}$$

$$\boldsymbol{K}_g = P_1 \boldsymbol{C}_1 + P_2 \boldsymbol{C}_2, \tag{2.11}$$

where $\langle \boldsymbol{v}_1 \rangle$ and $\boldsymbol{C}_1$ are the mean vector and covariance matrix of the channelized data vector $\boldsymbol{v}$ under the hypothesis H_i ($i = 1,2$). P_1 and P_2 are the occurrence probabilities (prevalences) for the two classes.

The CHO has been shown to correlate well with human observer performance on signal-known-exactly/background-known-exactly (SKE/BKE) tasks [44, 45], SKE-background known statistically (BKS) (e.g., lumpy backgrounds) tasks [46], and SKE-realistic anatomical backgrounds tasks [38, 41, 47] for a variety of types of nuclear medicine imaging. However, in those tasks the observer is only asked to decide whether the defect is present or not at a specified location. A more clinically realistic detection task is the signal-known-statistically (SKS)/BKS task, where variability can be present in both the signal and background. Here, signal variability is present in the form of variations in signal/defect shape, size, orientation, or topology/texture or combinations of the above. Background variability can come from two sources: quantum noise and anatomical variability. Modeling the latter is important in order to model clinical task where patients can vary greatly in size, shape, uptake, etc. It is important to model these image features, especially in studies such as virtual clinical trials, in order to accurately model performance on images from patient populations. For these clinically more realistic SKE/BKS and SKS/BKS tasks, there is evidence that rankings or ranking trends of human observers and the CHO are correlated

for different noise levels [47, 48], reconstruction methods and phantom populations [49], imaging systems [50], compensation methods, and post-filter cutoff frequencies[51]. Scanning forms of the CHO can be applied for the clinically more realistic SKS/BKS tasks to analyze each location within a particular region of interest (ROI) as a potential defect site [52]. However, the location-specific nature of the defect profile for the scanning CHOs will be a problem if there are extensive search areas and defect profile variability with respect to location is relatively high [53]. From an implementation perspective, for SKS tasks with a large search region and a high signal variability, the use of scanning CHO can be computationally demanding (even with the use of channels). Specifically, the scanning CHO requires computing covariance matrices numerically for every single pixel within the search region, which can be a problem for input image that has an extensively large search region such as multi-slice, multi-orientation image sets. Furthermore, scanning observers do not model the human's process of confirming a defect in slices across multiple orientations. For these reasons, the previous attempts to use scanning observers on multi-orientation, multi-slice images have focused on reducing the search region. The main techniques include using of a front-end search process [53] to obtain a subset of the original search location set (reduce the number of slices that analyzed by the scanning observer) and simplifying the defect confirmation process by simulating a simpler SKE/BKE detection task, etc. [54, 55]

In addition to the above limitations, existing model observers often predict rankings but not the absolute performance of human observers [56-59]. For imaging system optimization or comparison studies, this can be sufficient, but for other applications, such as selecting imaging time, administered activity, or radiation dose, prediction of absolute performance measures is required [32]. Obtaining absolute agreement for these model observers typically is done with the addition of observer internal noise [56]. The calibration process is a parameter search exercise

where the goal is to find the value of an internal noise parameter that matches performance between the model and human observers. Note that the calibration process is often performed for one specific combination of signal (shape, size, and orientation) and noise level, and it is unclear the degree to which the calibration generalizes to other situations.

In an attempt to resolve the limitations described above, we proposed a novel deep learning-based anthropomorphic model observer (DeepAMO) in Chapter 5. The proposed model observer can evaluate multi-orientation, multi-slice image sets to model the clinical diagnostic process of a radiologist or nuclear medicine physician in a clinically realistic 3D defect detection task. The DeepAMO was evaluated on an SKS/BKS tasks using a realistic anatomical background with variation in organ uptake and defect position (and thus orientation and shape). We also proposed a novel calibration method that 'learns' the underlying distribution of the human observer rating values (including the internal noise) using a Mixture Density Network. In the next section, we will introduce the fundamentals of convolutional neural networks (a deep learning algorithm) as well as review some of the current model observers that are based on convolutional neural networks.

2.5 Review of the current model observer based on Convolutional neural network

In this section, we will first give a brief introduction to convolutional neural networks (CNNs) from the perspective of object detection, which is a classic computer vision problem that is closely related to the defect detection problem of interest to this dissertation. There, we will introduce CNNs in the context of object detection, which includes the problem formulation, the

backpropagation algorithm, and the functionalities of some essential layers that have been widely used in object detection. Finally, we will review two recently published CNN-based model observer publications aimed at reproducing a human observer's defect detection performance, summarize the limitations in them, and finally state the aims of the proposed model observer in Chapter 5.

2.5.1 Introduction to convolutional neural network

A CNN is a deep learning algorithm that has been widely used in compute vision for recognizing and classifying features in images [60-63]. It is a multi-layer neural network originally designed to analyze visual inputs and perform tasks such as object detection, image recognition and classification. With successful experimental results and wide applications in computer vision, the use of CNN has become increasingly popular in the medical imaging community, particularly in medical image analysis, computer-aided diagnosis, radiotherapy, and task-based image quality evaluation.

2.5.1.1 Object detection with convolutional neural network

2.5.1.1.1 Loss function

The loss function works as the steering wheel for a neural network by defining the objective function and boundary for the task of the network. The loss function provides a measure of the difference between the output of the neural network for a given input and the desired (true) output. Depending on the application of the neural network, the loss function can be very different. The object detection problem that is most relevant to this dissertation is a one-class object detection problem, of which the goal is to detect an object's presence in an image. For this task, the binary cross-entropy loss (binary classification) has been widely used [64]:

$$L_{BCE}(\hat{y}_i, y_i) = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i), \tag{2.12}$$

where y_i is the ground truth label or target value (-1 or +1) for the ith image and $\hat{y}_i$ is the predicted label (a probability) for the ith image. N is the output size which is the number of images in the model output.

During training, the parameters (weights) of a network are updated by the update terms, which are the negative derivatives of the loss with respect to the weights times a small change δ, referred to as the learning rate. The algorithm used to calculate the gradient of a loss function with respect to the weights (local parameters) is called backpropagation, short for "backward propagation of errors". The backpropagation algorithm is introduced formally in section 2.5.1.2.2.

2.5.1.1.2 Architecture of a CNN

The architecture of a network can be understood as a way to achieve the objective defined by the loss function. In this section, we will introduce the problem formulation of the one-class object detection problem.

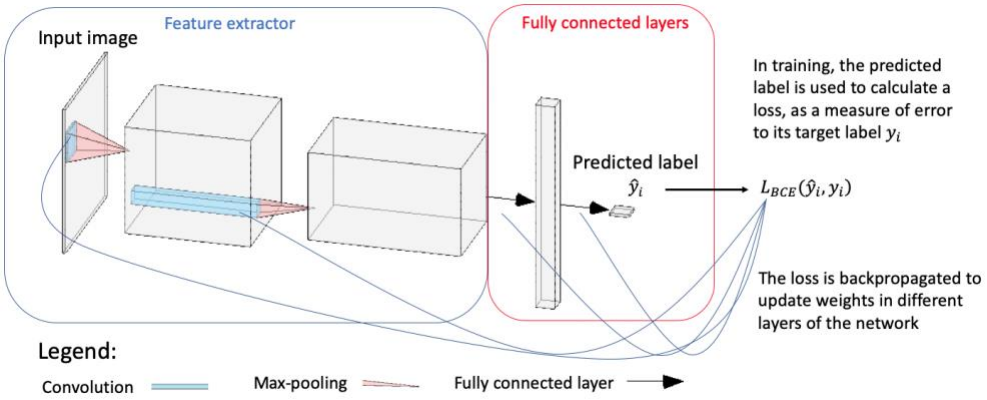

Figure 2.5. Illustration of the problem formulation for a one-class object detection problem

For a 2D object detection problem, the input to the network is a 2D image data, and the output is a predicted label for the input image. Typically, there are two main parts to a CNN designed for performing such task: (1) a feature extractor that is based on convolution responsible for producing various features of the image for analysis, and (2) a fully connected layer that uses the output of the feature extractor to select the best label for the image. A pictorial illustration of the problem formulation is shown in Fig. 2.5.

In the next section, we will introduce the basics of the modern CNN for object detection and provide a high-level view of why these networks have been some of the most influential innovations in the field of compute vision and medical image analysis.

2.5.1.2 Basics of the modern CNN for object detection

To understand how a modern CNN learns to detect an object in an image, we first need to understand backpropagation, the most widely used algorithm for training a neural network. In the following section, we will introduce backpropagation in the context of multi-layer perceptron, the precursor of the modern CNN, the reasons for use of convolution in analyzing visual data, and the basic layers in a modern CNN.

2.5.1.2.1 Multilayer perceptron

The idea of a multilayer perceptron is to address the limitations of a single-layer perceptron, namely, it can only classify linearly separable data into binary classes $(1, -1)$ [65]. A single-layer perceptron is a feed-forward network based on a threshold transfer function and has the structure as shown in Fig. 2.6.

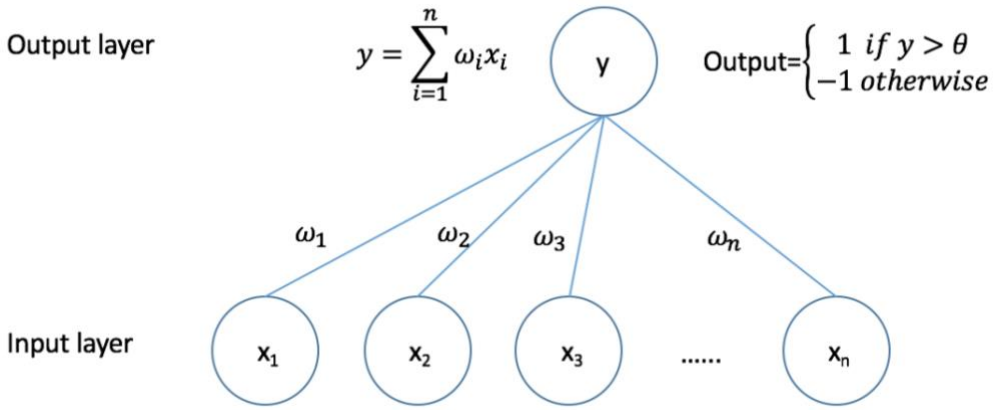

Figure 2.6. Illustration of a single-layer perceptron

A multilayer perceptron (MLP) is built on top of single-layer perceptron. In an MLP, the outputs from one layer are used as inputs to the next layer. Therefore, many layers can be specified to model complex non-linear relations between the inputs and outputs. The capacity of the MLP is related to the number of hidden units within it. More hidden layers (any layers in between the output and input layer) mean more parameters and thus greater capacity of the MLP; however, more training data are also needed at the same time. A sample MLP is given in Fig. 2.7.

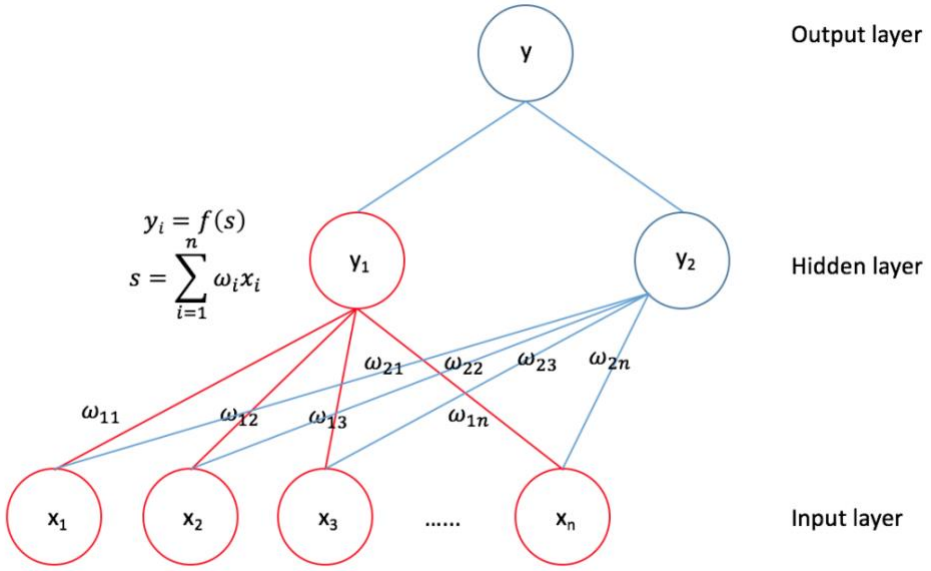

Figure 2.7. Illustration of a multilayer perceptron

The red nodes in the figure above represent the original part from the single layer perceptron that is shown in Fig. 2.5. The essential components that make an MLP differ from a single-layer perceptron are: (1) a soft thresholding function after each summation (linear combination of inputs), and (2) hidden layers. In theory, any complex non-linear relationship can be modeled by an MLP with enough hidden layers [66]. Thus, an MLP is often preferred over single-layer perceptron in modeling more sophisticated data, such as linear inseparable data, due to its ability to capture complex non-linearity.

2.5.1.2.2 Backpropagation

Backpropagation refers to application of the chain rule many times to calculate the gradient of a loss function with respect to the weights in a network. Fig. 2.8 shows a simple network containing only three inputs, two operations, and a single output. To understand backpropagation, we need to first answer the following question: how much change there is on the final result if the input is changed by an amount of δ. That is, in the example shown in Figure 2.8, how would change in a affect f, which is the final result of the network. To answer this question, we need to calculate the partial derivative with respect to that particular input, which is $\partial f/\partial a = 5$ in the example. This simply means that an increase in a would increase f by an amount equal to 5 δ (δ here denotes the change in a itself). In general, a positive gradient would positively influence the loss (the final result) and a negative gradient would negatively influence the loss, by the amount that is equal to the gradient multiplied by Δ. In a real neural network or a large computational circuit (imagine a very large number of operations and inputs), we can think a as one of the weights ω and b as one of the inputs x such as the ones shown in Figure 2.8. In the update equation, if we want to decrease the loss, we just need to update the weight by a tiny bit in the opposite direction

of its local partial gradient, i.e., decrease ω from -2 to -3. Doing so would give a smaller value of the loss function, f.

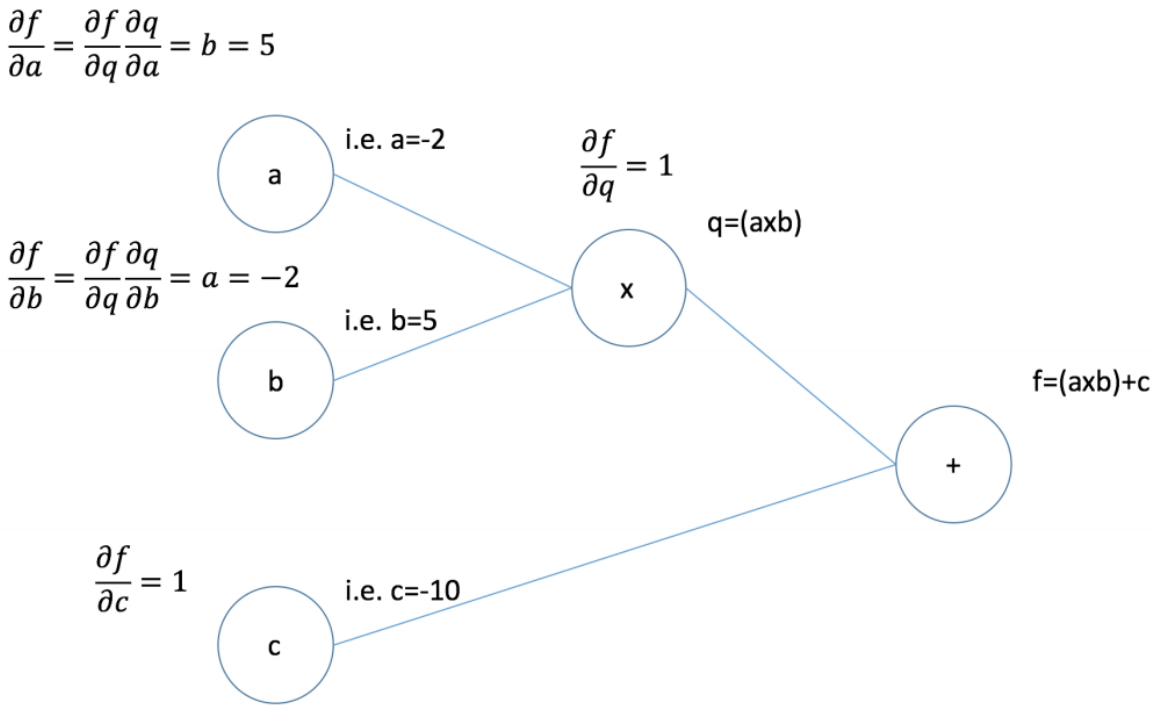

Figure 2.8. A pictorial illustration of backpropagation

To train an MLP, we have to estimate the weights of the perceptron. First, we need to calculate the loss, which acts as an indication of the error between the output of the network and the true output value for a corresponding sent of inputs.

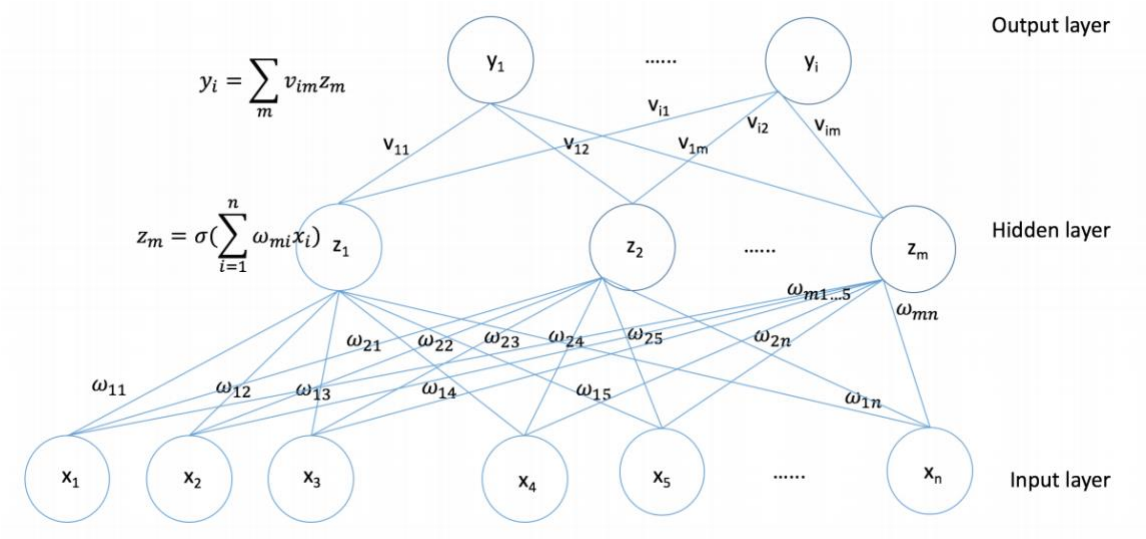

Figure 2.9. A sample MLP for demonstration of backpropagation

For the MLP in Fig. 2.9, the loss can be defined as

$$E[\omega, v] = \sum_i (y_i - \sum_m v_{im} \sigma \left(\sum_n \omega_{mn} x_n \right))^2. \tag{2.13}$$

As explained above, the update terms are the negative derivatives of the loss with respect to the local parameters (weights) times a small change δ, referred to as the learning rate:

$$\Delta \omega_{mn} = -\frac{\partial E}{\partial \omega_{mn}} \times \delta, \tag{2.14}$$

which is computed by the chain rule, and

$$\Delta v_{im} = -\frac{\partial E}{\partial v_{im}} \times \delta, \tag{2.15}$$

which is computed directly as they are weights of the last layer. By defining $z_m = \sigma(\sum_n \omega_{mn} x_n)$, we can write

$$E = \sum_i (y_i - \sum_m v_{im} z_m)^2, \tag{2.16}$$

and, its derivative as:

$$\frac{\partial E}{\partial \omega_{mn}} = \frac{\partial E}{\partial z_m} \frac{\partial z_m}{\partial \omega_{mn}}, \tag{2.17}$$

Taking the derivative of E with respect to z_m gives

$$\frac{\partial E}{\partial z_m} = 2\sum_i (y_i - \sum_m v_{im} z_m) v_{im}, \tag{2.18}$$

If we assume σ is a sigmoid function whose derivative is $\sigma(t)\big(1-\sigma(t)\big)$, then we have

$$\frac{\partial z_m}{\partial \omega_{mn}} = x_n \sigma\left(\sum_n \omega_{mn} x_n\right)\left(1 - \sigma\left(\sum_n \omega_{mn} x_n\right)\right), \tag{2.19}$$

And, finally, we have

$$\frac{\partial E}{\partial \omega_{mn}} = \frac{\partial E}{\partial z_m}\frac{\partial z_m}{\partial \omega_{mn}} \tag{2.20}$$
$$= 2\sum_i (y_i - \sum_m v_{im} z_m) v_{im}\, x_n \sigma(\sum_n \omega_{mn} x_n)(1 - \sigma(\sum_n \omega_{mn} x_n)),$$

where $\sum_i (y_i - \sum_m v_{im} z_m)$ is the error calculated at the output layer.

2.5.1.2.3 From MLP to modern CNN

For an MLP, the inputs are always 1D vectors. However, an image is a 2D vector and the structural information among the neighboring pixels or voxels does represent a great deal of information provided by the image. Vectorizing the image to a large 1D vector results in an oversized matrix of input weights. Consider a 2D image of size 23 × 23 (shown in Fig. 2.10) for which we would have 529 input nodes. If the hidden layer has 200 nodes, the size of the matrix of input weights would be 529 × 200 = 105,800. This is just the first layer, and as we increase the number of layers, the matrix size increases even more rapidly. Furthermore, vectorization inevitably destroys the spatial structural information in the image.

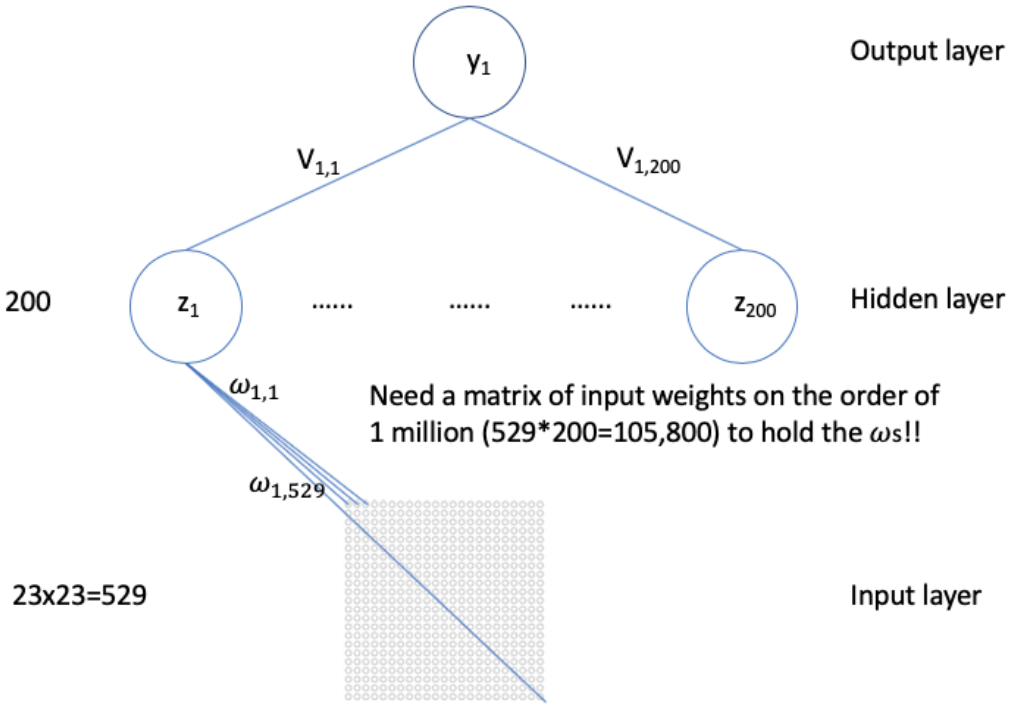

Figure 2.10. Illustration of an MLP on 2D image data

As early as 1987, researchers had started to explore the use of convolutional neural networks (CNN) to overcome both of these disadvantages. In 1998, the first work on modern CNNs was introduced by Yann LeCun for handwriting recognition [67]. In that paper, LeCun demonstrated that a CNN was able to aggregate simpler features into progressively more complex features, which could then be successfully used for handwritten digit recognition.

The fundamental difference between a CNN and an MLP is the addition of 2D convolution. 2D convolution has multiple unique advantages when it comes to processing 2D images. First, it can replace the computationally expensive matrix multiplications required by an MLP as learning a set of convolutional filters (each of 3x3) is much more tractable than learning a large matrix of millions of parameters [63]. Second, the 2D convolution filters can provide local connectivity (on the order of the size of the filter used) and weight-sharing (the same filter applied across the image) [68]. Third, the 2D convolution can naturally account for 2D spatial structural information in the

image [68]. As a result of the combination of these advantages and the great leap computational performance provided by GPUs, the CNN has enjoyed a huge surge in use after the AlexNet achieved state-of-the-art performance labeling of pictures in the ImageNet challenge [60].

2.5.1.2.4 Convolution layer

The most fundamental operation in a CNN is convolution. The role of a convolution layer is to detect local features at different locations in the input image, producing feature maps [68]. A convolution layer is essentially a set of learnable kernels or filters. A feature map is an output obtained by applying a filter in the convolution layer to an image. The image can be the input image or another feature map resulting from a preceding convolution layer. To calculate a set of feature maps for a convolution layer l, we need the feature maps in the preceding layer $l-1$ and the filters in the current convolutional layer. Mathematically, the feature maps resulting from the convolution layer l are given by [68]:

$$\boldsymbol{A}_j^{(l)} = f\left(\sum_{i=1}^{M^{(l-1)}} \boldsymbol{A}_i^{(l-1)} * k_{ij}^{(l)} + b_j^{(l)}\right), \tag{2.21}$$

where $M^{(l-1)}$ is the number of feature maps in the layer $l-1$, $*$ denotes a convolution in the spatial domain, $b_j^{(l)}$ is a bias parameter, and $f(\cdot)$ is a nonlinear activation function. The gradients or the derivative of a loss function with respect to the filter weights at a particular layer are computed by backpropagation as described in section 2.5.1.2.2. However, since the same filter kernel (i.e., set of weights) is applied multiple times at different locations in the image or feature map, the total derivative of the loss function with respect to the filter weights becomes a total gradient summed over the gradients computed at all these locations using that filter [68]. The

derivative of the loss with respect to the filter determines how much change in the filter weights will be needed in each iteration of training. In the following sections, we will introduce two other building blocks of CNN, namely, nonlinearity and pooling.

2.5.1.2.5 Nonlinearity

A CNN is usually composed of a series of convolutions intersected by nonlinearity operations. Nonlinearity operation in CNN is like the soft thresholding function in an MLP (introduced in section 2.5.1.1). It is essential as cascading a series of linear systems (like convolution) results in another linear system. By introducing the nonlinearities in between the layers, the model can be more expressive than a linear model [68]. Some of the most widely used nonlinearity functions include sigmod ($\sigma(x) = \frac{1}{1+e^{-x}}$), tanh ($tanh(x) = \frac{1}{1+e^{-x}}$), and ReLU ($ReLU(x) = max(0, x)$).

2.5.1.2.6 Pooling layer

A pooling layer is another building block of a CNN. The purpose of a pooling layer is two-fold: (1) reduce the spatial dimensionality of the feature maps, and (2) provide a small degree of spatial invariance. One limitation about convolution layer is that they can only produce feature maps that record the precise position of the features in the input image. A small shift in the position of the features in the input image will thus result in a different feature map. A pooling layer solves this problem by providing a lower resolution of the feature map which still contains the important structural information in the feature map, but without the fine details that may not be useful to the task. At the meanwhile, the number of parameters and amount of computation need to be learned

in the later layers are drastically reduced. Among the different types of pooling operations, the most commonly used pooling is max-pooling. A max-pooling layer is essentially a $n \times n$ max filter, where each region the filter covers is replaced by its max value within the region [68]. A pictorial illustration of a max-pooling layer is in Fig. 2.11.

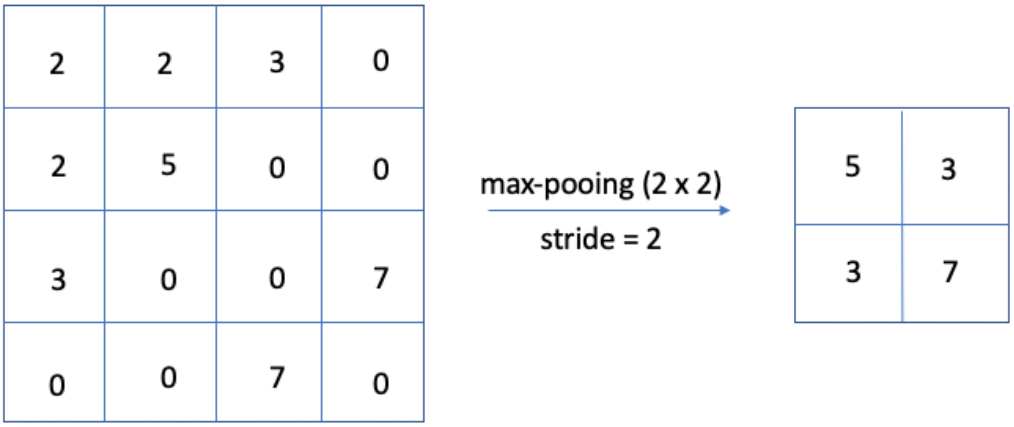

Figure 2.11. A pictorial illustration of a max-pooling layer with a filter size of 2x2 and stride of 2.

2.5.1.2.7 Summary

The basic layers and the backpropagation algorithm discussed above cover the most essential components in a modern CNN for object detection, which is the most relevant to our goal of modeling the defect detection task in a model observer (the surrogate of a human observer). In the next section, we will review two recently published works that use CNN to model human observer in performing defect detection tasks with CT images.

2.5.2 Review of CNN-based model observer

Recent developments in deep learning have opened up a door to new opportunities in the field of task-based image quality assessment. Several recent studies have explored the use of CNNs

as model observers. In [69] and [70], the authors demonstrated good agreement between CNN-based model observers and human observers on single-slice 2D detection tasks in both simulated and clinical mammography but concluded that a large amount of training data is needed. In [71], CNN-based model observers achieved similar performance to human observers on a uniform background in CT phantom images. In a more recent work [72], a CNN was trained to approximate the ideal observer, using a computer-simulated uniform background with correlated noise.

Although these above-mentioned studies demonstrated good agreement between CNN-based model observers and human observers, those observers were not designed to reproduce human observer task performance. To model human observer task performance, a calibration process is often needed to model inter- and intra-variability of the human [56]. Intra-observer variability refers to the fact a human observer will produce, in general, in different rating values in different reading trials. Inter-observer variability refers the variation in rating values for the same image read by different human observers. Only recently have CNN model observers been proposed to model human observer performance on 2D defect detection tasks.

In [71], an MLP and a CNN were proposed to predict the performance of a human observer on a liver lesion detection task using single-slice, single-orientation CT images and were compared to a CHO (with Gabor channels and internal channel noise). The MLP consisted of an input layer and an output layer with a nonlinear activation (SoftMax) function. A 2D image was vectorized and fed as input to the MLP and the SoftMax function normalized the output values of all units ($k = 1, \ldots, K$; human observer's rating value ranging from 1 to K) of the output layer and returns the likelihood that the image is of class k. The CNN was composed of two convolutional layers, each followed by a max-pooling layer and a flattening layer followed by two fully connected layers with the last one having an output size $1 \times K$. The human observer rating values were used to train

the MLP and CNN. The results of the work showed that the MLP and CNN correlated well (very close to the performance of the CHO with internal noise) with the results from a human observer for different x-ray exposure levels (multiple models were trained) and lesion sizes. However, the authors pointed out that they had a relatively large amount of training data for the MLP-based model observer and also their results were generated on a relatively simple task using CT phantom images with a uniform background. Thus, further evaluation on more challenging and realistic tasks is needed.

More recently, a deep-learning-based model observer (DL-MO) was proposed to model human observer performance on a lung nodule detection and localization task on multi-slice, single-orientation 2D CT images [73]. The work was based on an underlying assumption that there exists similarity between the CNN and the human visual system. So, they proposed to use a pre-trained CNN (trained using natural images) as a deep feature extractor as an initial stage applied to the input image. In order to reduce the dimension of the feature map, the extracted feature maps (from a pre-selected layer) were subsequently fed to a feature-engineering model to generate the test statistic for an input image. Specifically, their proposed framework included four major components: a pretrained CNN, a partial least square regression discriminant analysis (PLS-DA) model, an internal noise component, and a nodule searching process. A sliding window strategy was first used on the input image (single-orientation 2D slices) to extract local image patches that were used as the inputs to a pretrained CNN (ResNet50 [74]). The CNN was pretrained on a natural image dataset, and the output from an intermediate layer (pre-selected) of the CNN was used as input to the PLS-DA model to generate a test statistic, λ_0, for the input image patch. A spatial distribution of the test statistics (heat map) was obtained by scanning through all potential nodule locations. The nodule search process was then applied to identify the location of the voxel in the

original input image that coincided with the maximal value of the test statistic in the heat map, i.e., the most-likely location of lung nodules. Finally, an internal noise component was added to the maximal test statistic to model the variation of human reader performance, i.e., $\lambda = \lambda_0 + \alpha N(0, \lambda_{0,bkg})$, where λ denotes the final test statistic and α is the weighting factor that is to be found out in the calibration process. The work demonstrated strong correlation and agreement between the proposed DL-MO and human observers for a low-contrast liver lesion detection task in patient liver background. However, the author stated that one of the limitations of the work is that there are two free parameters (the CNN layer used for feature extraction and the number of PLS components) that need to be properly determined to achieve reasonable performance for this method.

In summary, the CNN-based model observer seems to have a promising use in the optimization of medical imaging systems and acquisition methods. However, the biggest limitations for the current observer models, which include both the CNN-based models and the traditional models, is the inability to handle 3D data in a rigorous way, or more specifically, to model the human scanning-and-confirming process in a faithful way. Most of those model observers were designed for analyzing single-orientation 2D slices. By contrast, many clinical tasks require the interpretation of 3D datasets, which requires the reader to scan and confirm defect(s) using multiple slices in multiple orientations. Thus, it remains a challenge to fully model a clinically realistic 3D defect detection task, using multi-orientation, multi-slice image sets.

Chapter 3

A projection image database to investigate factors affecting image quality in weight-based dosing: application to pediatric renal SPECT

3.1 Introduction

In nuclear medicine imaging, the product of acquisition duration and administered activity (AA) determines the level of quantum noise present in the image. Quantum noise can have a direct impact on diagnostic image quality, and, for the purposes of maximizing image quality, reducing AA, or reducing acquisition duration, it is desirable to study the relationship between these factors.

Over the past decade, there has been an increased interest in reducing patient radiation exposure in diagnostic imaging studies that use ionizing radiation. Therefore, there has been significant interest in the nuclear medicine community in establishing universally accepted and optimized dosing guidelines for pediatric nuclear medicine studies. The European Association of Nuclear Medicine (EANM) and Society of Nuclear Medicine and Molecular Imaging (SNMMI) have, respectively, published the European pediatric dosage card and the North American consensus guidelines for pediatric AA [12, 13]. The goal of these guidelines is to provide a balance between radiation risk and image quality. However, these guidelines were developed either based on a consensus of best practices or a simple estimate of image quality and not on a rigorous evaluation of diagnostic image quality relative to AA.

A second concern in pediatric imaging is the acquisition duration. Sedation is often required, especially for longer acquisitions. Longer acquisition durations increase the chance of patient motion, which can degrade image quality. Shorter acquisition durations are thus desirable.

All else being equal, reducing the product of AA times acquisition duration will increase the Poisson noise in the image. However, the effect of changes in quantum image noise on diagnostic performance are complicated [47]. Similarly, decreasing quantum noise in the images requires increasing AA, acquisition duration, or both. Increasing the AA above that needed to provide acceptable image quality violates the principle of as low exposure as reasonably possible (ALARA). Consequently, appropriate guidelines for pediatric AAs are of significant interest [75]. Similarly, increasing the acquisition duration in pediatric patients to compensate for reduced AA may not be acceptable. Thus, understanding the tradeoff between image quality and the product of AA and acquisition duration is an important problem.

In 2008, the Dosimetry and Pediatrics Committees of the EANM published the first version of the EANM pediatric dosage card to better standardize the AAs in pediatric nuclear medicine procedures. The dosage card was based on data from a publication by Jacobs et al. [76]. In that study, count rates and effective doses were computed as a function of body weight for 10 radionuclides and 95 radiopharmaceuticals, respectively, using 7 hermaphrodite anthropomorphic computational phantoms [77]. Count rate was used as the only surrogate for image quality; a discussion of the details and limitations of that aspect of that work are provided in the discussion section.

A second effort at standardization of pediatric dosages was the 2010 North America Consensus Pediatric Dosing Guidelines [78]. The AAs recommended in that report were slightly lower for infants and small children as compared to the EANM guidelines, compensating for the

higher radiation risk in early childhood. Those guidelines were based on a combination of experience and retrospective analysis of clinical data, taking into account the patient's weight and count rate density per unit area or volume, and using these as the surrogates for radiation risk and count rates as the surrogate for image quality.

In 2011, Sgouros et al. proposed a rigorous method to balance diagnostic image quality with cancer risk using ^{99m}Tc-DSMA as an example [5]. The study showed that weight alone may not be sufficient for optimally scaling AA in children. In that study, nonuniform rational B-spline (NURBS)-based anatomic phantoms, realistic organ uptakes and models of the image formation process, and task-based measures of image quality were used to objectively compare image quality of ^{99m}Tc-DMSA SPECT images. Two 10-year-old females of the same weight but different heights, respectively representing short-stout and tall-thin patients, were used in that study. Several different AAs (25%, 50%, 75%, 100%, 125%, and 150%), defect locations, and lesion severities with different target-to-background activity concentration ratios were simulated to represent clinical imaging. Channelized Hoteling observer methodology was used in a receiver-operating-characteristic (ROC) analysis of lesion detectability to study the relationship between AA and the area under the ROC curve (AUC). The results of the study showed that the same AUC could be obtained for the tall-thin phantom with approximately half the AA as for the short-stout phantom. [5].

In this present study, we have built upon the Sgouros et al. work by developing a realistic pediatric phantom population including variations in age, gender, kidney size, and height. We have also proposed a novel method that produces contrast-matched, clinically-relevant defects in all of the phantoms across different ages, gender, body morphometries, and kidney sizes. The combination of these methods allows application of task-based image quality methods to

rigorously assess current dosing guidelines in terms of their effectiveness for equalizing image quality across patients with different age and body morphometry.

Toward this end, we simulated realistic projections of the pediatric patient population in preparation for future detailed investigations of the tradeoffs between image quality, the product of AA and acquisition duration, patient weight and height, and reconstruction method for ^{99m}Tc-DMSA renal imaging. Using this realistic phantom population and projection database, we investigated the effects of scatter, count density, and radius of rotation as a function of patient morphometry. These studies provide insight into the changes in these surrogate indices for factors affecting image quality and how they change with patient weight and body morphometry and the limitations of weight-based scaling of AA. We also performed a model observer study to investigate further the impact of patient weight on image quality to study the validity of weight-based dose scaling for ^{99m}Tc- DMSA imaging.

3.2 Methods

3.2.1 Population of realistic digital phantoms

The database of projection data for this study was generated using the Advanced Laboratory for Radiation Dosimetry Studies (ALRADS) UF NHANES-based phantom series [79]. The phantom population realistically models pediatric heights, weights, organ sizes and anatomies for both genders at five ages. The phantoms were adjusted to model variations in height and organ size prior to voxelization. The ages modeled were newborn, 1-, 5-, 10-, and 15-years old. For each age, we modeled the 50th percentile weight and 10th, 50th, and 90th percentile heights, simulating patients having the average weight at each age with varying body habitus. The 10th, 50th and 90th

height percentile phantoms are referred to as short-stout, reference, and tall-thin patients, respectively

Figure 3.1. Sample coronal slices of the body reminder, cortex, medulla, pelvis, liver and spleen (from left to right) of a newborn 50th height percentile male phantom.

For each age and height percentile, we modeled three kidney masses: -15%, average, and +15%, where average is the International Commission on Radiological Protection (ICRP) standard mass for a patient with the corresponding age and the percentiles are the change relative to this standard mass. The variations in kidney mass model variations in patient kidney size; for newborn patients, the dosimetric impact of these sizes on risk has been previously studied [79].

In addition to anatomic variability, we simulated variations in uptake in 6 tissues: cortex, medulla, pelvis, spleen, liver, and body reminder (the remaining soft tissues of the phantom). Fig. 3.1 shows sample coronal slices of these different objects (organs and renal sub-structures) in a newborn phantom of average height (50th percentile height). Projections of each object were generated separately assuming a uniform activity distribution. The individual projections could then be scaled and summed to represent the count level that would be obtained in projections for an arbitrary AA, acquisition duration, or set of relative uptakes. By individually generating and scaling these projections, we were able to adjust the uptakes in each individual object to simulate uptake variability.

Each phantom was digitized prior to simulation into 0.1 cm cubic voxels and truncated in the axial direction to exclude regions more than 5 cm below the bottom or above the top of the kidneys. This truncation was done in order to reduce simulation time and data storage requirements.

3.2.2 Organ uptake model

Uptake in the kidneys was estimated using data from a single imaging time point, which varied slightly across patients, from datasets of 47 patients with ages ranging from 1 to 16 years acquired at the Boston Children's Hospital (BCH). We did not attempt to develop an age-specific pharmacokinetics model from this data, and considered the data from all patients as a single mixed-age population sample for estimating uptake of activities in the kidneys. CT scans of these patients were not available for attenuation compensation as they were not acquired as part of the patient's clinical study. Instead, attenuation maps were estimated based on automated intensity thresholding of images reconstructed from scatter windows. The data were reconstructed using 5 iterations with 8 subsets per iteration of an ordered-subsets expectation-maximization (OS-EM) reconstruction method that included attenuation, scatter and collimator-detector response compensation. Reconstructed images were converted to units of activity concentration using the measured camera sensitivity. The kidneys were segmented automatically using intensity thresholding, and the reasonability of the kidney VOIs and body contours were reviewed manually. The percent of the decay-corrected AA in the kidneys in these VOIs is referred to as the kidney uptake fraction. In addition, we used thresholding to segment the kidney cortex and pelvis/medulla regions. From these we computed the ratio of activity concentrations (sum of activity values divided by volume of the VOI) in the cortex to the medulla/pelvis. This ratio is referred to as the cortex-to-medulla plus pelvis activity concentration ratio. The results obtained from the above procedure are summarized in Table 3.1, and were used as estimates for percent tracer uptakes in the patient's kidneys.

The fractional uptakes in the spleen and liver at the imaging time were obtained from Evans et al. [80]; the values used were 4.3% and 1.7%, respectively. We validated these percentages against the real patient data from BCH and found that the uptake variations for these organs were small across the patient datasets and small compared to the uptake in the kidneys. Therefore, we used constant uptake percentages for liver and spleen in the simulations.

To model the differences in uptake of the fine structures inside the kidney (renal cortex, medulla, and pelvis), we quantified the relative uptakes inside these renal structures using the data from the aforementioned 47 patient images. The relative uptake values were estimated using the reconstructed images described above. We used threshold-based segmentation to separate the cortex from the medulla plus pelvis and created two separate VOIs for the two entities. These VOIs and the activity values inside them were used to estimate activity concentration of each entity. The mean and standard deviation of the cortex-to-medulla plus pelvis activity concentration ratios were calculated and are summarized in Table 3.1.

The resolution in the images was not sufficient to estimate accurately the activity concentration ratio between medulla and pelvis. Thus, we based the activity concentration ratio in these two structures on input from our clinical collaborators. Images were generated with a variety of concentration ratios; images having a medulla-to-pelvis concentration ratio of 1:1 were deemed most realistic, and that ratio was thus used in the study.

3.2.3 Organ uptake variations

We modeled random variations in the uptake of the kidneys as a whole and in the cortex relative to the medulla plus pelvis using truncated Normal distributions. The values of the minimum, maximum, mean and standard deviation of these distributions were obtained from the

47 patients described above, and are given in Table 1. For each phantom anatomy, we randomly sampled 384 values each of the fraction of injected activity in the entire kidney and the cortex-to-medulla plus pelvis activity concentration ratio. From these data combinations, the weight-based AA, and the kidney volume, we calculated the activity concentrations in the cortex, medulla and pelvis for each of the 384 uptake realizations.

3.2.4 Projection data simulation

The projections were simulated using an analytic projection code that models attenuation, spatially varying detector-to-collimator response [81], and object-dependent scatter [82]. This code has been extensively validated for imaging of a variety of radionuclides by comparison to Monte Carlo simulations and experimental projection measurements. We modeled a low-energy, ultra high-resolution (LEUHR) collimator, a 360° body-contouring orbit, 120 projection views, a 15% wide energy window centered at 140 keV, an energy resolution of 9% at 140 keV, and a 0.1 cm projection bin size. After each simulation, the projections were collapsed isotropically by a factor of 2 to simulate a 0.2 cm projection bin size. Attenuation maps used in the projection operation were constructed by assigning the attenuation coefficient of the organ in the phantom containing the voxel center to the entire voxel. The attenuation coefficients were evaluated at 140 keV for the material composition of each organ based on ICRP organ composition data [83]. Fig. 3.2 shows sample transaxial images of attenuation distributions that illustrate variations in body habitus of the population.

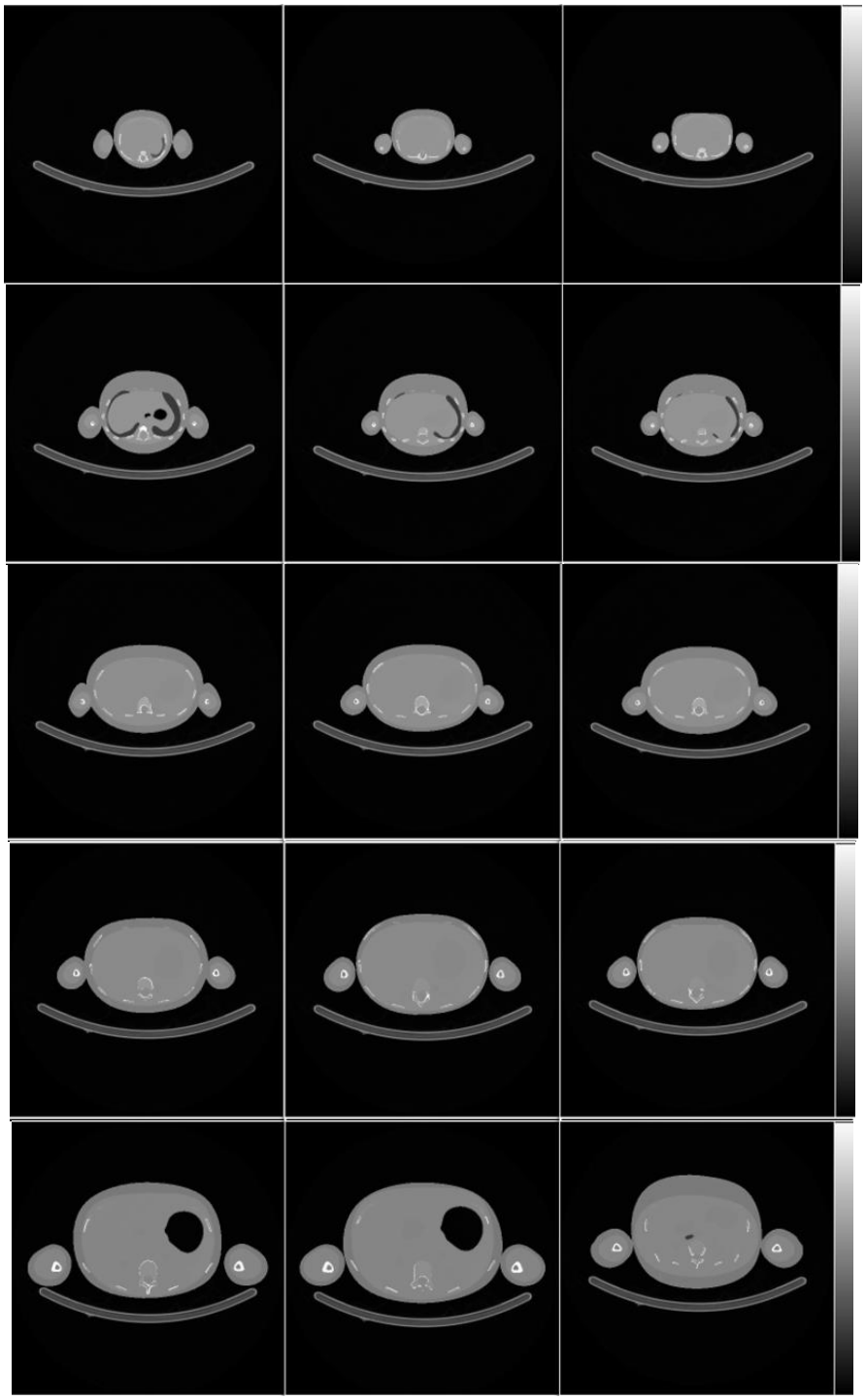

Figure 3.2. Sample transaxial images of the attenuation distribution for the (left to right) 10th, 50th, and 90th height percentile versions of the male phantom for ages (top to bottom) 0 (newborn), 1, 5, 10, and 15 years showing variations in body habitus.

Figure 3.3. Noise-free projection images of the kidney cortex, medulla, spleen, liver, pelvis, and body remainder for a male, reference-height, newborn phantom.

Table 3.1. Summary of population parameters

	Kidney Uptake Fraction	Cortex-to-Medulla + Pelvis Act. Conc. Ratio
Maximum	0.393	2.00
Minimum	0.329	1.36
Sample mean	0.361	1.68
Sample standard deviation	0.025	0.25

Projections were generated using the above methods individually for the kidney cortex, medulla, pelvis, liver, spleen, and the body reminder, by assigning unit intensity to the phantom voxels in each of these regions. The organ projections were then scaled by the randomly-sampled uptake scaling factors needed to obtain the desired activity concentrations. They were then summed and scaled by the camera sensitivity to obtain the raw projections per unit injected activity per acquisition duration. These raw projections were next scaled by the acquisition duration and desired AA to give the mean projections in units of counts. Using these as input to a Poisson random noise generator gave the noisy projections.

3.2.5 Simulated projection data with variation in injected activity

Six count levels were simulated: 25%, 50%, 75%, 100%, 125%, and 150%. Here, a count-level indicates the fraction of AA relative to the AA obtained from the 2010 North American Consensus Dosing Guidelines. Note that the suggested minimum and maximum AAs in these guidelines were only enforced for the clinical (100%) count level. Fig. 3.3 and 3.4 show a sample set of noise-free projections of the organs for the newborn and sample noisy projection images from the various count levels.

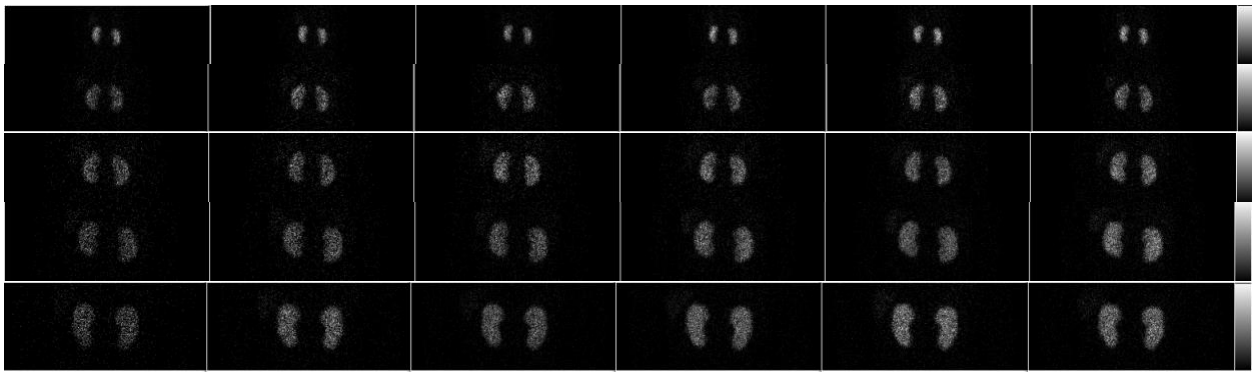

Figure 3.4. Sample noisy posterior projection images from the various count levels. From top to bottom, shows kidneys for the 0, 1, 5, 10, and 15-year-old phantoms. From left to right, the simulated count levels were 25%, 50%, 75%, 100%, 125%, and 150% of those of the 2010 North American Consensus Dosing Guidelines.

Table 3.2. Comparison of total counts in clinical and simulated projections

Phantom age	Corresponding patient age	Total counts in patient image	Counts in simulated image	Percent difference
1	1.2	328821	373482	-12.718
5	5	543850	540371	0.642
10	9	831381	711803	15.498
15	16	1100752	1215463	-9.905

3.2.6 Validation of simulated projection image

To validate simulation process, we chose patient images from each of the five ages. The patient and simulated projections were reconstructed using 2 iterations of 8 subsets of the OS-EM iterative reconstruction with detector-to-collimator response compensation only, followed by filtering with a 5-mm FWHM Gaussian filter. Sample reconstructed image slices are shown in Fig. 3.5. Note that the kidney model in the phantom does not model the detailed structure of the medulla and pelvis. We computed the total of the reconstructed voxel values in volumes of images containing the kidneys. Table 3.2 shows a comparison of the counts in the simulated and patient images for the various ages.

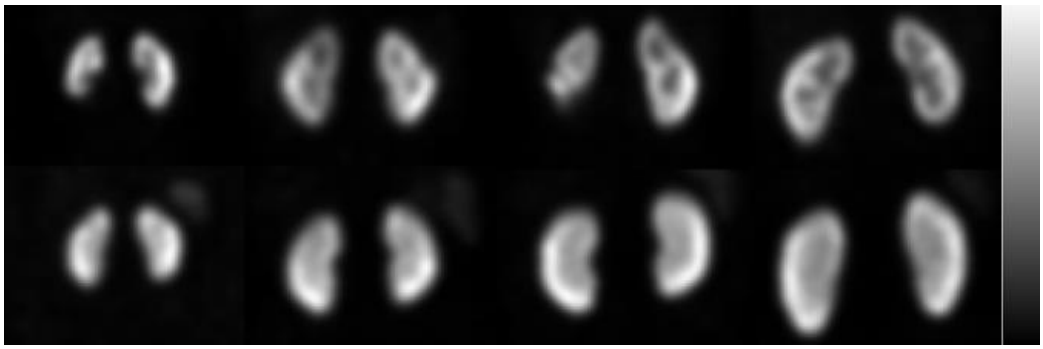

Figure 3.5. From left to right, the top row shows patient images from 1.2, 5, 9, and 16-year-olds reconstructed using 2 iterations of 8 subsets of the OS-EM reconstruction with detector response compensation followed by a Gaussian filter with a FWHM of 0.5mm. The bottom row shows simulated images from 1, 5, 10, and 15-year-olds reconstructed using the same methods.

3.2.7 Defect model

Assessing image quality should be done with respect to the task that will be performed with the images. In the case of DMSA SPECT, the task is to detect functional defects in the renal cortex.

Thus, it is necessary to create defects in the simulated images. Since the ultimate goal of the project is to provide guidance data for selecting minimum AAs commensurate with detecting clinically relevant defects, we developed defects for each age that were challenging but clinically relevant [84]. Challenging defects will tend to be ones that are small, where partial volume effects produce low contrast defects in the reconstructed images. Since the distance from the collimator face to the kidney depends on patient size, resolution will tend to be worse for large patients compared to small ones. In addition, since the thickness of the cortex tends to be greater for larger patients, and partial volume effects depend on the uptake in tissues surrounding the defect that are within approximately two times the full width at half maximum (FWHM), the same physical defect size would be harder to visualize in a large patient than in a smaller patient. Also, when the kidney is larger, the same size defect would tend to be of less consequence in terms of total renal function. Thus, we chose to create defects with sizes that varied depending on the patient size.

Based on input from an experienced pediatric nuclear medicine specialist (S.T. Treves), we selected a defect volume of 0.3 cm^3 as the defect size that is clinically relevant and challenging for a newborn. To create a realistic defect, we used an ellipsoid with one major axis length equal to approximately the thickness of the cortex. The center of the ellipsoid was positioned along the line extending through the center-of-mass of the cortex at the point where it intersects the outer cortical surface. The half-length of the axis of the ellipsoid in this direction was equal to the cortical thickness along this line, meaning that the apex of the ellipse was at the inner surface of the cortex. The half-lengths of the other two axes were the same; for the newborn, the length of these remaining axes was set so that the intersection with the kidney cortex had a volume of 0.3 cm^3. This was verified numerically by creating a voxelized version of the defect where the voxel values were set to unity using sub-voxel sampling by a factor of 2, taking the product of this ellipsoid

with the cortical VOI, and summing the values in the resulting product image. A trans-cortical defect with a realistic shape can be created by scaling the defect image by the desired contrast and subtracting this from the cortical VOI. Note that, due to the linearity of the projection operation, the subtraction can be performed in either the image or projection domain. The position along the cortex was defined by an angular coordinate in the coronal slice containing the defect center of mass. Gradual transitions of function can be modeled by blurring the ellipsoid with a Gaussian kernel prior to the multiplication described above. Sample images containing defects for phantoms representing various ages in Fig. 3.6.

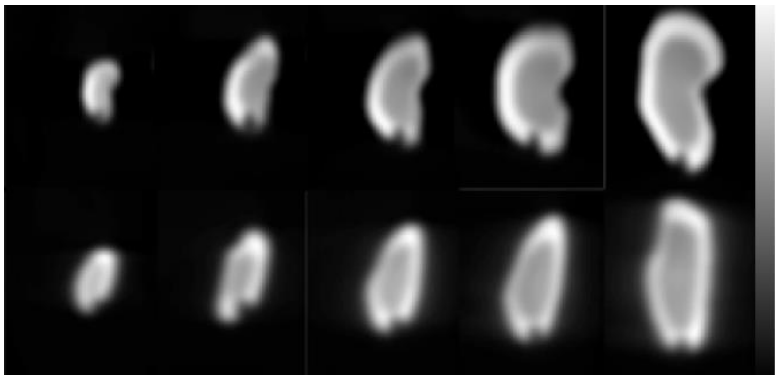

Figure 3.6. Sample lower pole defects in noise-free reconstructed images for newborn, 1-, 5-, 10-, and 15-year-old male phantoms with reference heights in coronal and sagittal views. The defect volumes for ages 1, 5, 10 and 15 were determined by matching their contrasts to the newborn.

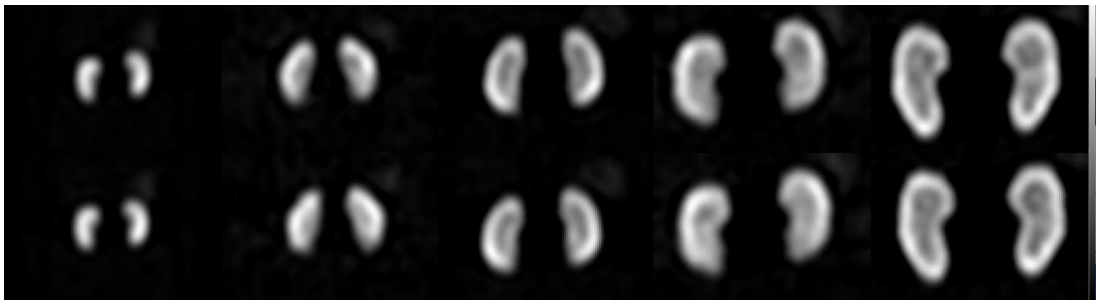

Figure 3.7. Sample reconstructed images from noisy projection data using FBP reconstruction followed by a post-reconstruction 3D Butterworth filter with an order of eight and cutoff frequency of 0.12 cycle/pixel. Negative values were mapped to zero in the display. From left to right, the bottom and top rows shows coronal images with and without, respectively, a (lower pole) defect for the newborn, 1-, 5-, 10-, and 15-year-old male phantoms at the 50th height percentile. The volumes of these defects were chosen to be near the limits of clinical relevance and to have the same defect contrast.

3.2.8 Reconstruction and post-reconstruction processing

SPECT images were reconstructed from the simulated projections using filtered backprojection (FBP) reconstruction and a ramp filter with no apodization. The reconstructed images had cubic voxels with a side length of 0.2 cm. The reconstructed images were post-filtered with 3-D Butterworth filters of order 8 and cutoff frequency 0.12 cycles/pixel. Fig. 3.7 shows a sample set of reconstructed images for the five ages.

3.2.9 Quantitative measures of image quality

Defect detectability depends, in principle and among other factors, on the contrast of the defect and the amount of noise in the reconstructed image. The contrast of the defect depends on the intrinsic contrast of the defect relative to surrounding tissues and the size of the defect with respect to the resolution of the imaging system. In addition, contrast is degraded by the effects of scattered photons from surrounding tissues. In the following we present physical measures of image quality that quantify the noise, image resolution, and scatter, all of which together affect image quality.

We measured the contribution of scattered photons in the kidney region by the scatter-to-primary ratio obtained from projection images that only contain the kidneys. The images of the kidneys resulting from detected scattered photons were obtained by subtracting the kidney projection generated with attenuation, collimator-to-detector response, and scatter from the same projection generated with attenuation and collimator-to-detector response. The numbers of scattered and primary photons were obtained by summing the counts in the resulting images, respectively.

Lastly, we quantified the noise in the kidney region by the kidney count density, i.e., the number of detected primary photons emitted in the kidneys divided by the kidney volume. We used the primary-photon-only projection images of the kidneys, generated as described above, for this calculation. We performed this calculation for all 90 phantoms using the same mean kidney uptake fraction (0.361) and cortex-to-medulla plus pelvis activity concentration ratio (1.68) in all cases. The count density was averaged over the 3 kidney sizes for each phantom.

In SPECT, the image resolution at the center of rotation is proportional to the radius of rotation. Thus, we quantified the system resolution by the distance from the collimator face to the center of rotation averaged over the (body contouring) camera orbit. Note that the phantoms were placed on a camera bed measured using a CT scan, so the camera orbit, especially for the small phantoms, was constrained in some views by the size of the bed.

3.2.10 Model observer study

The model observer study was performed using methods similar to those previously described in [5]. In summary, seven 2-dimensional frequency-domain bandpass difference-of-mesa channels were used to approximately model the human visual system. The starting frequency and width of the first channel was 0.5 cycles per pixel, and subsequent channels had widths that doubled and abutted the previous channel. These channels were applied to the 3 orthogonal slices (transaxial, sagittal, and coronal) that contained the defect centroid. The output of this procedure was a 21-element feature vector.

In previous work, feature vectors were analyzed using a Hotelling Observer (HO) methodology, a combination often referred to as the channelized Hoteling observer (CHO). However, the projections in the database and resulting reconstructed images reflect variations in

anatomy and uptake. This resulted in features vectors that were often multi-modal thus not multivariate normally (MVN) distributed. We have previously demonstrated that this can result in difficulty for the HO [85]. Thus, instead of traditional CHO, we used a multi-template strategy proposed by Li and Jha [86] to handle the non-MVN data. In this strategy, the entire ensemble of input data vectors is decomposed into sub-ensembles that are approximately MVN. A linear discriminant trained using the data for that sub-ensemble is then used to analyze the set of feature vectors for that sub-ensemble. The set of test statistics from each sub-ensemble is then pooled and analyzed using ROC analysis. In this work, the channel output vectors were sorted into sub-ensembles based on defect location, age, and height percentile. It was verified qualitatively that the resulting channel output data were not multi-modal and were nearly MVN distributed. We used a leave-one-out sampling methodology to generate the subsets for each sub-ensemble [87]. We pooled the test statistics for each defect location and height percentile and applied ROC analysis. The AUC was used as a figure-of-merit for task performance.

3.3 Results

3.3.1 Quantification of noise by renal count density

Figure 3.8 shows plots of the average kidney count density as a function of patient age for the different height percentiles for male and female phantoms, respectively. Overall, the plots demonstrate that the weight-based AA produced nearly equal kidney count densities for all ages except for the newborn. The data also shows that gender did not affect count density in patients less than 10 years old. This indicates that there is not much difference in the overall attenuation in the kidney region between the male and female phantoms for these ages. Note that, in theory, we

expect to see rankings of the values of the count density in the order $90^{th} > 50^{th} > 10^{th}$ height percentile: we would expect the short-stout phantoms to allow fewer photons to escape the body than for the reference or the tall-thin phantoms. However, this was not observed in all cases. For example, we see that, for the 10-year-olds, the largest difference was between the 10^{th} and 50^{th} height percentile. This difference in count density was approximately 30% in the male phantoms and 15% in the female phantoms. For the 15-year-olds, the differences were 25% for the male phantom and 15% for the female phantoms, respectively. There was essentially no difference in count density between 10^{th} and 90^{th} height percentile phantoms for both the 10 and 15-year-olds for both genders.

Fig. 3.9 shows transaxial images of phantom slices at the mid-kidney level for the 10^{th}, 50^{th}, and 90^{th} height percentile from male phantoms with ages 0, 1, 5, 10, and 15. From these images, we can see that there was not a significant difference in body circumference (girth) among phantoms for different heights nor a strong correlation between girth and phantom height rankings. That is, the 10^{th} height percentile (short-stout) phantom did not necessarily have a larger girth than the 90^{th} height percentile (tall-thin) phantom or the reference phantom, at the mid-kidney level. These images provide a pictorial illustration of the reason that the observed count density did not vary as expected with patient height.

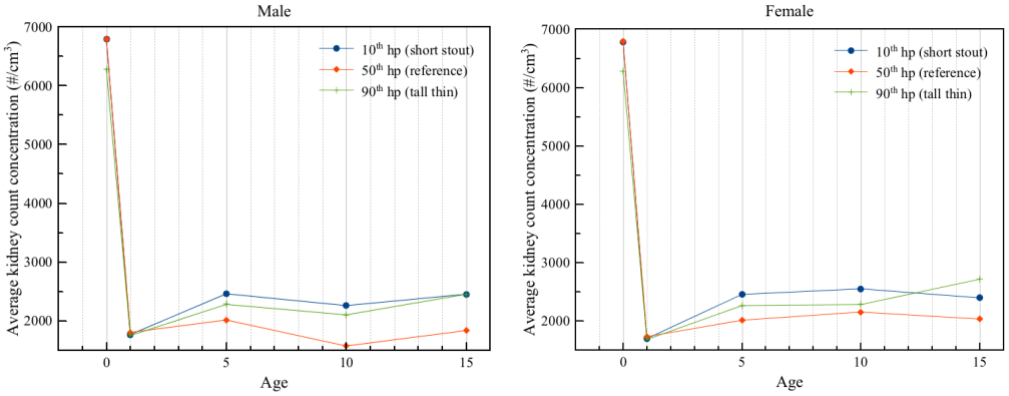

Figure 3.8. Average kidney count density obtained for three different height percentiles as a function of phantom age for male and female phantoms.

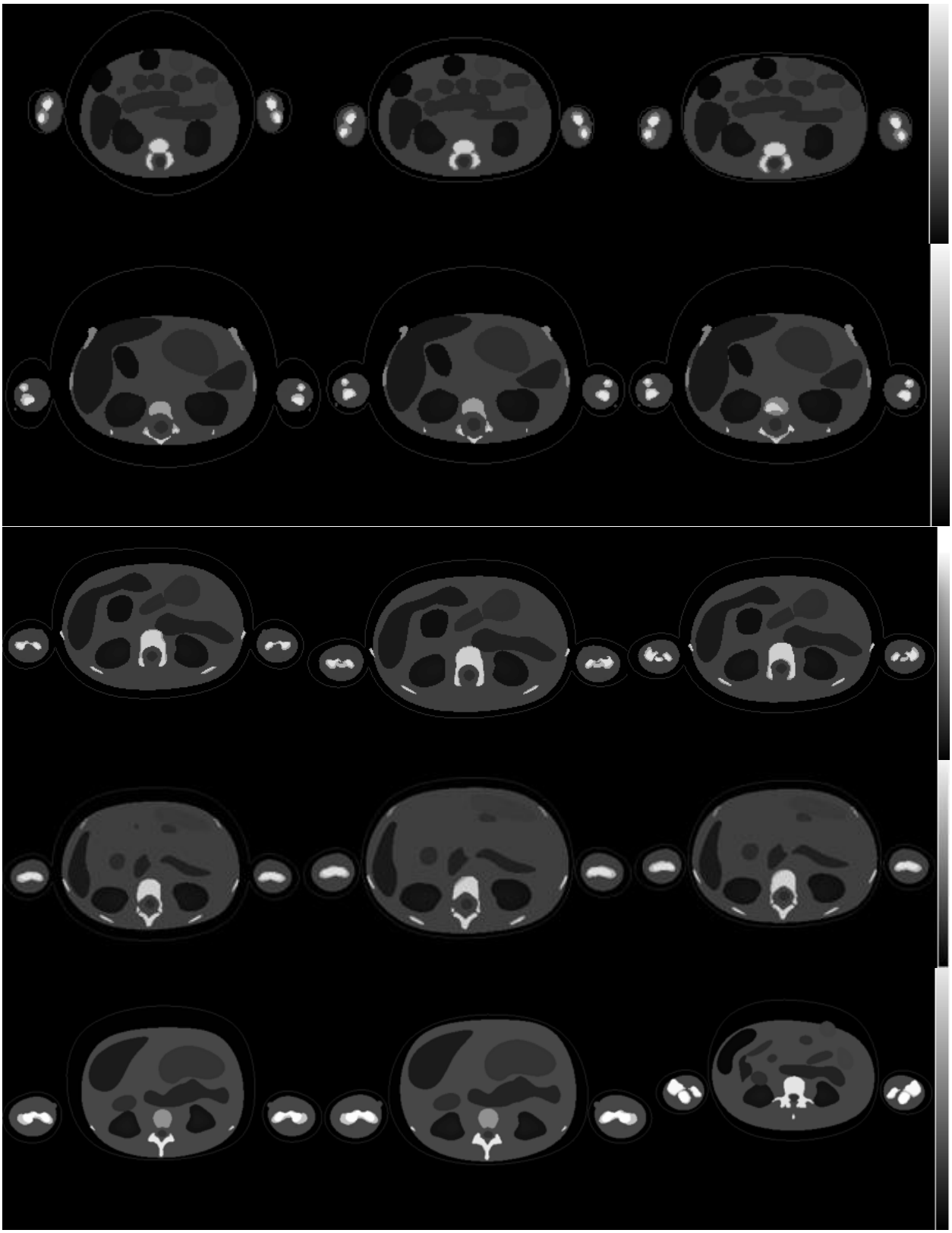

Figure 3.9. Sample transaxial phantom images at mid-kidney level in 10th, 50th, and 90th height percentile (from left to right) from the male phantom of age 0, 1, 5, 10, and 15 (from top to bottom) showing variations in body habitus.

3.3.2 Quantification of scatter by scatter-to-primary ratio

Fig. 3.10 shows plots of the average scatter-to-primary ratio for the three different height percentiles as a function of patient age for male and female phantoms, respectively. It is clear that dosing by weight did not equalize the effects of scatter. These data also demonstrate that the scatter-to-primary ratios depended to a varying degree on height. Just as for count density, the expected rankings were not consistently observed. For example, one would expect the tall-thin patient to have a smaller scatter-to-primary ratio than the short-stout phantom. However, this was not observed for the 15-year-old male, though the differences in the ratios for the different heights were generally small. In any event, these data suggest that weight and height alone may not sufficient for predicting the effects of scatter, and thus their degrading effects on image quality.

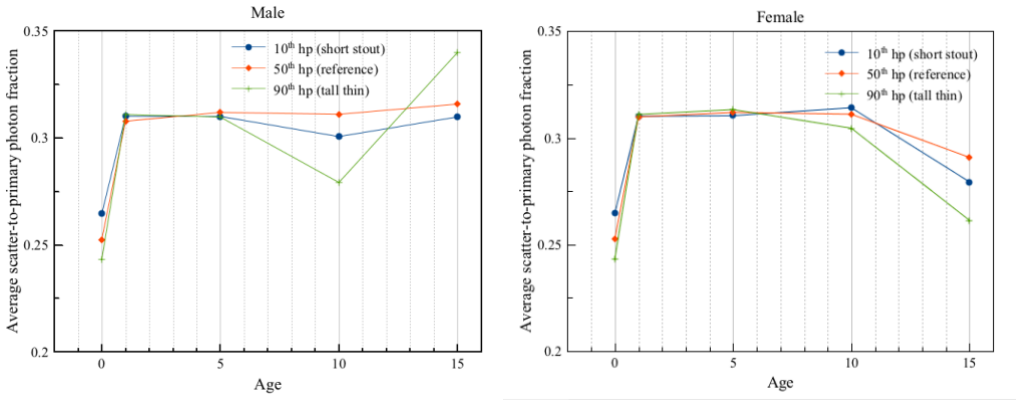

Figure 3.10. Average scatter-to-primary ratio obtained from three different height percentiles as a function of phantom age for male and female.

3.3.3 Quantification of resolution by camera radius of rotation

Fig. 3.11 shows plots of the average camera radius of rotation for the three different height percentiles as a function of patient age for male and female phantoms, respectively. Again, the

expected trend, larger height percentiles having smaller average radii of rotations, was not always observed. The cause is likely for the same reason that differences in count density did not vary as expected with height: the maximum girth of the patient, which determines the distance from the camera to the body for a body contouring orbit, did not vary directly with height percentile. This indicates that height and weight are not sufficient to predict resolution effects. This is especially true for the small phantoms as the radius of rotation was limited by the size of the imaging bed in the lateral direction.

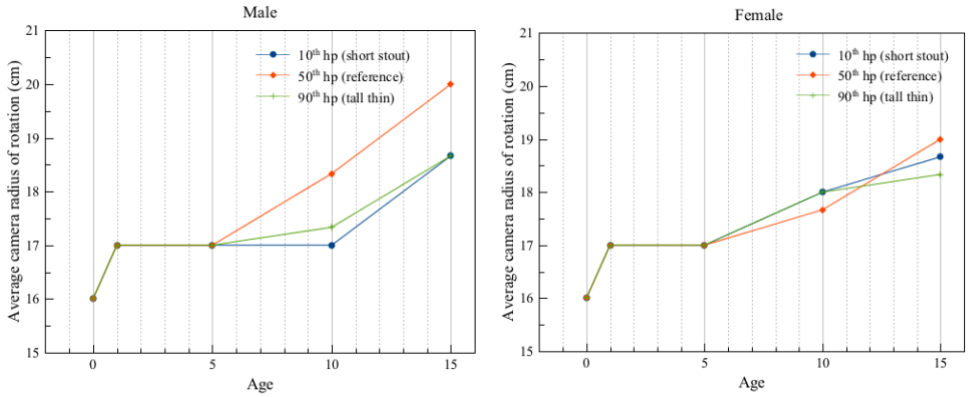

Figure 3.11. Average camera radius of rotation obtained from three different height percentiles as a function of phantom age for male and female.

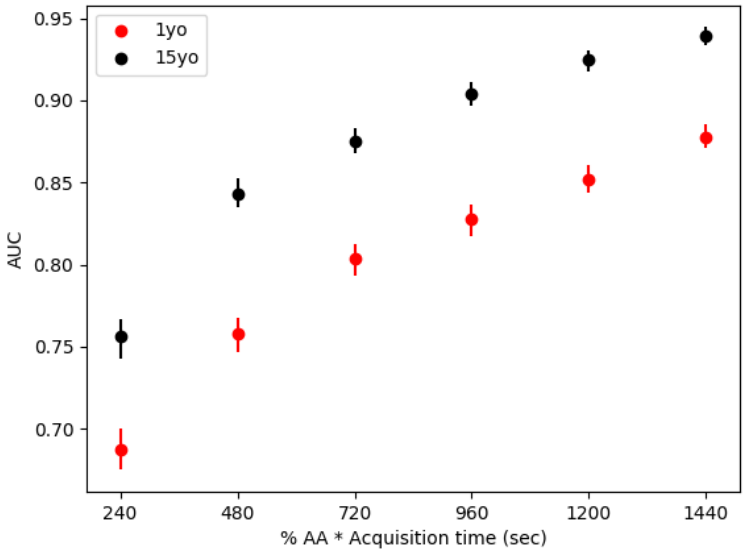

Figure 3.12. Image quality result on a defect detection task for the 1- and 5-year-old phantoms. A 20% defect contrast was modeled for these patients.

3.3.4 Model observer study results

Fig. 3.12 shows plot of AUC as function of the product of percent AA and acquisition duration for the 1- and 15-year-old on a defect detection task. Image quality was studied as a function of count level using the developed database with a fixed acquisition duration of 960 seconds. The saturation of the AUC values indicates that there is a decreasing benefit of increasing the AA, indicating that detection is limited by background variation. These results show that there was a monotonic but modestly saturating increase in AUC with AA. The fact that there was modest saturation indicates that image quality was limited by quantum noise and the effects of object variability were modest [88]. More importantly, the results show that the AUCs for an AA of 100% of the weight-based, consensus dosing guideline were not equal, indicating that the ability to detect a defect with the same contrast, was not the same for the AA recommended by the North American Consensus Guidelines. This, combined with the results of the other image quality surrogates reflecting noise, scatter, and resolution effects, suggest that weight-based scaling is not sufficient to equalize image quality.

3.4 Discussion

This paper describes the design and simulation of a realistic projection database for use in pediatric renal SPECT research. The population included variability in age (and thus weight), gender, kidney size, and height. The specific dataset generated here was focused on matching the defect detectability across the population for a challenging and relevant defect detection task. We thus designed a set of defects, one for each phantom and kidney size, that is clinically relevant but at the limits of what is likely to be detectable. Thus, the curve describing the tradeoff between

image quality and AA is for a difficult case; larger defects are likely to be easier to detect and thus not as affected by reductions in AA.

Using the established projection database, we investigated the AAs based on the current North American consensus weight-based dosing guidelines in terms of impact on surrogates for factors that affect image quality (image noise, resolution, and contrast). As compared to the approach by Jacobs et al. [76], the present study provides a more rigorous evaluation on image quality by adopting a more realistic phantom database, imaging simulations, pharmacokinetic model, and task-based image quality evaluation method. In the Jacob et al. study, 7 phantoms were used, representing newborns and children (male only) of 1, 5, 10, and 15 years, adult females and adult males, corresponding to their reference weights. Though these phantoms included 7 organs, the radionuclide was assumed uniformly distributed in the phantom (no inter-organ uptake variability) for the purposes of estimating count rates. The fraction of energy absorbed by the target organs at the emitted photon energy was computed using Monte Carlo simulations. The count rates that would be obtained with gamma camera imaging were assumed to be proportional to the average number of photons (at energies useful for imaging) that exited the body. These count rates were considered to have potentially contributed to the image. The fraction of exiting photons was computed as a weighted sum of the non-absorbed fractions (one minus the absorbed fraction) at the energy of each emitted photon. The average number of photons emitted per disintegration for each emitted photon energy was used to weight the absorbed fractions, and only emitted photon with abundances greater than or equal to 10% and energies suitable for imaging were included in the calculation. Normalization factors for the count rates were obtained by dividing corresponding count rates by that for an adult male of 70 kg (normalization factor = 1.0).

Using the count rates estimated by this method as the sole surrogate for image quality has several limitations. First, it ignores the effects of other factors, such as scatter and resolution, which can vary with body size. Related to this, energy exiting the body is an incomplete surrogate for primary photon counts. It equates, for example, two or more scattered photons with total energy equal to the primary photon energy to a primary photon. In other words, the use of total energy as a surrogate for primary photon counts is valid if the scatter-to-primary photon fraction is a constant across the entire patient population. However, this is not the case as was demonstrated above. Second, the method assumed uniform radionuclide distribution in the body. This is less than ideal, especially for agents such as DMSA that concentrates in a small number of tissues. Third, there was no variation in organ size or patient height for a given weight. Finally, the suggested minimum AAs were purely based on effective dose and not image quality. The authors did point out in the discussion that the suggested AAs calculated using this method could possibly lead to impractical scanning times or unusable images.

3.5 Conclusion

A realistic projection database has been generated for investigation of relationship between image quality and patient morphometry in ^{99m}Tc-DMSA renal SPECT. A total of 207,360 projection images was generated, encompassing 6 different administrated activities for 90 phantom anatomies. This projection database can be used to study the relationship between the product of AA and acquisition duration and image quality in a way that is impossible with either real patients or via experimental phantoms. The database generated in this work is immediately applicable to other pharmaceuticals labeled with ^{99m}Tc used in pediatric imaging such as ^{99m}Tc-MAG3 or ^{99m}Tc-MDP; only scaling and summing of the organ projections with appropriately scaling factors

reflecting agent biokinetics. Further, the methods used in this study are applicable to studying these tradeoffs for other diagnostic and/or therapeutic radiopharmaceuticals in both pediatric and adult patients.

Using this projection database, we conducted a quantitative analysis of three factors that affect image quality: noise, as measured by kidney count density; scatter, as measured by the scatter-to-primary ratio for photons emitted from the kidneys; and resolution, as measured by the average radius of rotation. The results of this study showed that weight-based dosing was partially able to offset losses in count density due to variations in patient weight. However, it suggested that the kidney count density for newborns was higher than for other ages using weight-based dosing. The results also demonstrated variations in scatter and resolution that depend on body morphometry, but were not well correlated with phantom height. We also performed a task-based image quality study using an anthropomorphic model observer that demonstrated that the weight-based scaling of the AA did not equalize image quality as measured by the AUC. This, combined with the image quality surrogate data on noise, scatter, and resolution, suggests that weight-based scaling is not sufficient, suggesting that a dosing procedure beyond simple weight-based scaling of AA is required to equalize image quality in pediatric renal SPECT. Further, the results also suggest the need for more detailed task-based studies of image quality, and that variables beyond height and weight are needed in order to prescribe AAs that optimize image quality in order to achieve as low as reasonably possible dosing.

Chapter 4

Current pediatric dosing guidelines for ^{99m}Tc-DMSA SPECT based on patient weight do not provide the same task-based image quality

4.1 Introduction

In nuclear medicine imaging, the product of acquisition duration and administered activity (AA) determines the level of quantum noise present in the image. Quantum noise can have a direct impact on diagnostic image quality, and, for the purposes of maximizing image quality, reducing AA, or reducing acquisition duration, it is desirable to study the relationship between these factors.

Over the past decade, there has been an increased interest in reducing patient radiation exposure in diagnostic imaging studies that use ionizing radiation. Therefore, there has been significant interest in the nuclear medicine community in establishing universally accepted and optimized dosing guidelines for pediatric nuclear medicine studies. The European Association of Nuclear Medicine (EANM) and Society of Nuclear Medicine and Molecular Imaging (SNMMI) have, respectively, published the European pediatric dosage card and the North American consensus guidelines for pediatric AA [12, 13]. The goal of these guidelines is to provide a balance between radiation risk and image quality. However, these guidelines were developed either based on a consensus of best practices or a simple estimate of image quality and not on a rigorous evaluation of diagnostic image quality relative to AA.

A second concern in pediatric imaging is the acquisition duration. Sedation is often required, especially for longer acquisitions. Longer acquisition durations increase the chance of patient motion, which can degrade image quality. Shorter acquisition durations are thus desirable.

All else being equal, reducing the product of AA times acquisition duration will increase the Poisson noise in the image. However, the effect of changes in quantum image noise on diagnostic performance are complicated [47]. Similarly, decreasing quantum noise in the images requires increasing AA, acquisition duration, or both. Increasing the AA above that needed to provide acceptable image quality violates the principle of as low exposure as reasonably possible (ALARA). Consequently, appropriate guidelines for pediatric AAs are of significant interest [75]. Similarly, increasing the acquisition duration in pediatric patients to compensate for reduced AA may not be acceptable. Thus, understanding the tradeoff between image quality and the product of AA and acquisition duration is an important problem.

In 2008, the Dosimetry and Pediatrics Committees of the EANM published the first version of the EANM pediatric dosage card to better standardize the AAs in pediatric nuclear medicine procedures. The dosage card was based on data from a publication by Jacobs et al. [76]. In that study, count rates and effective doses were computed as a function of body weight for 10 radionuclides and 95 radiopharmaceuticals, respectively, using 7 hermaphrodite anthropomorphic computational phantoms [77]. Count rate was used as the only surrogate for image quality; a discussion of the details and limitations of that aspect of that work are provided in the discussion section.

A second effort at standardization of pediatric dosages was the 2010 North America Consensus Pediatric Dosing Guidelines [78]. The AAs recommended in that report were slightly

lower for infants and small children as compared to the EANM guidelines, compensating for the higher radiation risk in early childhood. Those guidelines were based on a combination of experience and retrospective analysis of clinical data, taking into account the patient's weight and count rate density per unit area or volume, and using these as the surrogates for radiation risk and count rates as the surrogate for image quality.

In 2011, Sgouros et al. proposed a rigorous method to balance diagnostic image quality with cancer risk using ^{99m}Tc-DSMA as an example [5]. The study showed that weight alone may not be sufficient for optimally scaling AA in children. In that study, nonuniform rational B-spline (NURBS)-based anatomic phantoms, realistic organ uptakes and models of the image formation process, and task-based measures of image quality were used to objectively compare image quality of ^{99m}Tc-DMSA SPECT images. Two 10-year-old females of the same weight but different heights, respectively representing short-stout and tall-thin patients, were used in that study. Several different AAs (25%, 50%, 75%, 100%, 125%, and 150%), defect locations, and lesion severities with different target-to-background activity concentration ratios were simulated to represent clinical imaging. Channelized Hoteling observer methodology was used in a receiver-operating-characteristic (ROC) analysis of lesion detectability to study the relationship between AA and the area under the ROC curve (AUC). The results of the study showed that the same AUC could be obtained for the tall-thin phantom with approximately half the AA as for the short-stout phantom. [5].

In this present study, we have built upon the Sgouros et al. work by developing a realistic pediatric phantom population including variations in age, gender, kidney size, and height. We have also proposed a novel method that produces contrast-matched, clinically-relevant defects in all of the phantoms across different ages, gender, body morphometries, and kidney sizes. The

combination of these methods allows application of task-based image quality methods to rigorously assess current dosing guidelines in terms of their effectiveness for equalizing image quality across patients with different age and body morphometry.

Toward this end, we simulated realistic projections of the pediatric patient population in preparation for future detailed investigations of the tradeoffs between image quality, the product of AA and acquisition duration, patient weight and height, and reconstruction method for ^{99m}Tc-DMSA renal imaging. Using this realistic phantom population and projection database, we investigated the effects of scatter, count density, and radius of rotation as a function of patient morphometry. These studies provide insight into the changes in these surrogate indices for factors affecting image quality and how they change with patient weight and body morphometry and the limitations of weight-based scaling of AA. We also performed a model observer study to investigate further the impact of patient weight on image quality to study the validity of weight-based dose scaling for ^{99m}Tc- DMSA imaging.

4.2 Methods and materials

4.2.1 Series of realistic digital phantoms

The series of pediatric phantoms used was developed at the University of Florida and was based on demographic data from the CDC's National Health and Nutrition Examination Survey (NHANES) data [79]. It consisted of 90 phantoms that included variations in age, gender, height, and kidney mass. For each gender, five groups were modeled: 0, 1, 5, 10, and 15 years of age. All phantoms at a given age had a weight equal to the 50th percentile weight and one of three height

percentiles: 10th (short), 50th (reference) and 90th (tall). The phantoms for each height percentile and age group are shown in Fig. 4.1. The targeted weights for each age are provided in Table 4.1. For each height percentile, we modeled 3 kidney sizes: -15%, average, and +15%. The variation in kidney size was used to model the effects of anatomical variation [79] that would not be externally observable. The phantoms were digitized using 0.1 cm cubic voxels.

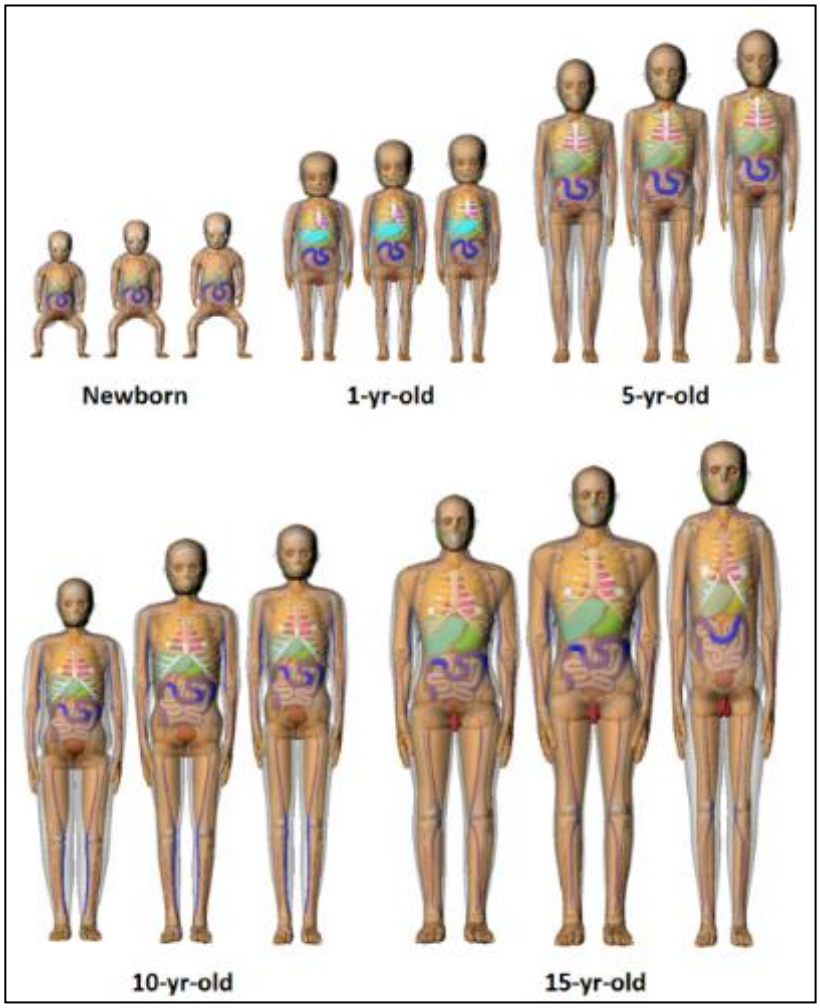

Figure 4.1. Renderings of 10th, 50th, and 90th percentile height at constant 50th percentile weight newborn, 1-yr-old, 5-yr-old, 10-yr-old, and 15-yr-old hybrid phantoms.

Table 4.3. Summary of phantom masses

Age (yr)	Male	Female
Newborn	3.5 kg	3.4 kg
1y	10.4 kg	9.5 kg
5y	20 kg	20 kg
10y	30 kg	35 kg
15y	55 kg	50 kg

4.2.2 Pharmacokinetics model

A new pharmacokinetic (PK) model for ^{99m}Tc-DMSA was used in this study to model kidney uptake [89]. The PK model is based on literature data and was validated using 47 patient datasets acquired at the Boston Children's Hospital (BCH). Tracer uptake in individual organs, i.e., the kidneys, spleen, liver, and body remainder, at 3 hours post injection was computed using the PK model. Variations in tracer uptake based on those seen in the 47 patient datasets were modeled using the coefficient of variation (percent standard deviation) from those data and assuming a truncated normal distribution.

4.2.3 Defect model

We used a defect model described in [84]. In the model, a defect volume of 0.3 cm3 for a newborn patient with the reference kidney size and 50th height percentile was deemed, by an experienced pediatric nuclear medicine specialist, clinically relevant and at the limits of clinical detectability in the newborn phantom. Defect volumes for other ages were determined so that the defect contrast for each age at the 50th height percentile was the same as for that phantom [84]. Using this model, focal renal lesions consisting of areas of reduced uptake were created to simulate focal acute pyelonephritis in three locations (lower pole, upper pole and lateral aspect of the kidney)

along the cortical wall.

4.2.4 Projection data simulation

For each phantom in the population, we simulated noise-free projection data for the renal cortex, medulla, pelvis, liver, spleen, and background (including all other organs), modeling the physics and acquisition parameters appropriate for ^{99m}Tc renal SPECT. The projections were generated using an analytic projection code that modeled attenuation, spatially varying collimator-to-detector response [81], and object-dependent scatter [82]. The code has been previously validated by comparison to Monte Carlo and experimental projection data for imaging of a variety of radionuclides [90-98]. The projections were simulated for a low-energy, ultra-high-resolution collimator at 120 projection views over a 360° body-contouring orbit and a 0.2-cm projection bin size. Prior to simulation, the phantom was placed on a patient bed obtained from a CT scan of the bed on a Siemens Symbia SPECT/CT system. This bed constrained the orbit, especially for small phantoms.

The renal activity and relative activity concentrations for structures inside the kidney (the renal cortex, medulla, and pelvis) were randomly sampled from truncated Gaussian distributions with the means and standard deviations derived from the PK model and 47 sets of patient data acquired at the Boston Children's hospital. These parameters are summarized in Table 3.1. Each individual organ projection was scaled by its relative uptake value and the product of AA, acquisition duration, and scanner sensitivity.

A projection of the entire phantom was generated by summing these individual sets of scaled organ projections. Simulated projections were scaled to represent AA-levels (AA relative to the standard weight-based AA) varying from 25% to 150% in increments of 25%. Poisson noise

was then simulated using a Poisson distributed random number generator. A total of 207,360 sets of projection images were thus generated: 64 uptake realizations × 6 count levels × 5 ages × 3 height percentiles × 2 genders × 3 kidney sizes × 3 defect locations × 2 defect statuses (present or absent).

4.2.5 Image reconstruction and post-reconstruction processing

Images were reconstructed using filtered backprojection (FBP) and post-filtered with 3D Butterworth filters with an order of 8. We determined the optimal cutoff frequency for a 3D post-reconstruction Butterworth filter based on the AA giving the highest AUC at each count level. The optimal cutoff frequency was 0.6 cycles per cm for all the count levels investigated. This cutoff frequency was used for all the AUC values presented below. The reconstructed images had cubic voxels with a side of length of 0.2 cm. Images centered on the defect with a size of 128x128 pixels were extracted from the coronal, transaxial, and sagittal slices containing the defect centroid and used in the image quality evaluation. Samples of these images are shown in Fig. 4.2.

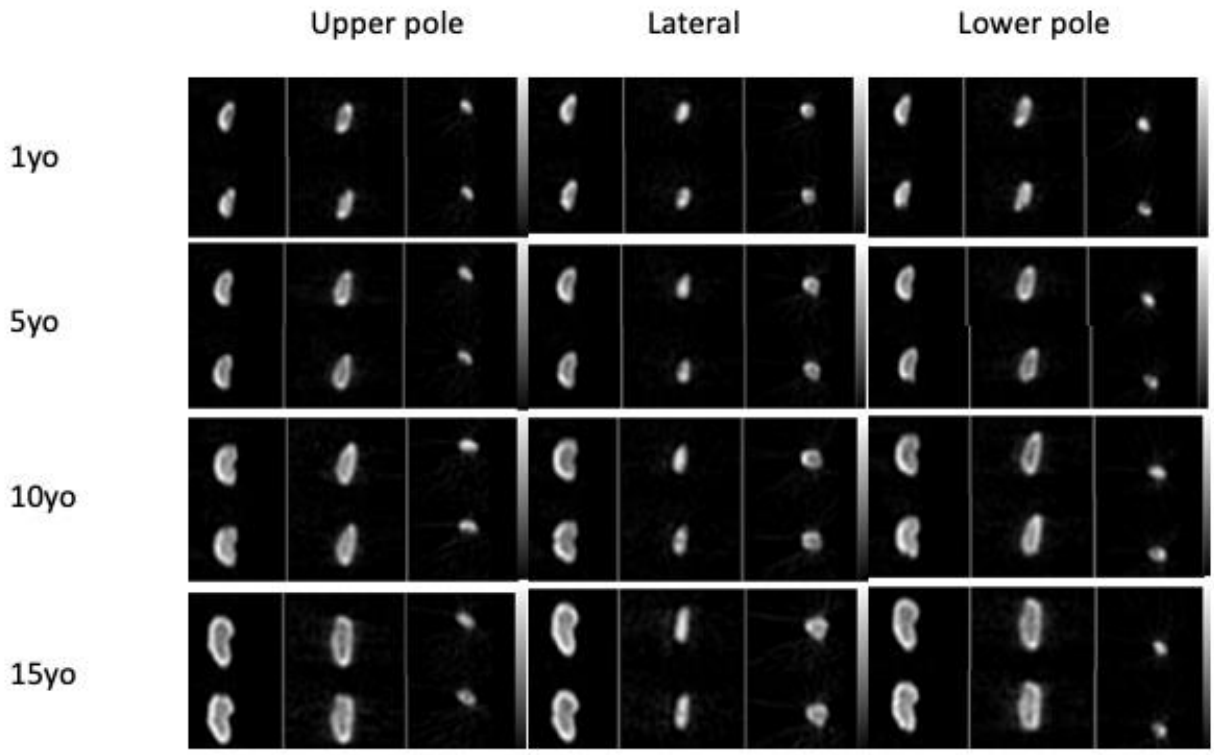

Figure 4.2. From top to bottom the images show upper, lateral, and lower pole (from left to right) defects for the 50th height percentile for the 1- and 5-year-old female and 10- and 15-year-old male phantoms.

4.2.6 Model observer

The channelized Hotelling observer (CHO), first proposed by Myers and Barrett [45] has been shown to provide good predictions of human performance on detection tasks for a variety of nuclear medicine imaging applications[38, 44, 46, 47]. The CHO uses a set of frequency channels applied to input images that model the human visual system combined with the Hotelling Observer, which approximates the Ideal Observer in cases where the input data are multi-variate normally (MVN) distributed with equal covariance matrices.

As noted, the Hotelling Observer is strictly optimal only when the input data (i.e., the vectors of channel outputs) are MVN distributed; conversely, it performs poorly when the input

data are multimodally distributed [86, 99]. The data used in this study, as discussed below, included both background and signal variations and were non-MVN. Thus, instead of the traditional CHO, we used a multi-template strategy proposed by Li et al. to handle the non-MVN data. This strategy involves partitioning the data into sub-ensembles that are approximately MVN and applying the optimal linear discriminant to each sub-ensemble[86]. We used a leave-one-out training-testing strategy. In this strategy, one feature vector was left-out (i.e., not used in the training), and the remaining vectors were used to train the observer. The observer was then applied to the left-out vector to produce a test statistic. This process was repeated with each vector in the ensemble being left-out once. This process was applied to each sub-ensemble and produced a number of test statistics equal to the size of the sub-ensemble. The resulting test statistics produced by this strategy were pooled and analyzed, using ROC analysis to estimate the AUC, which served as a FOM for task performance.

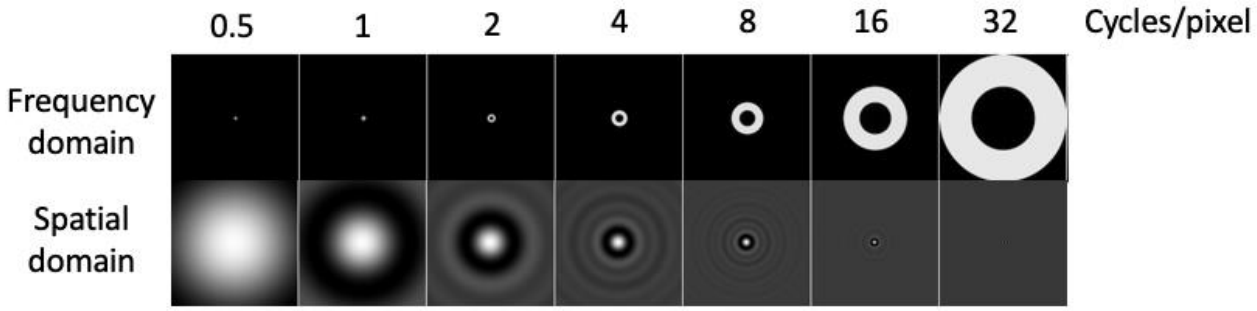

Figure 4.3. Images of the seven anthropomorphic DOM channels used in this work. The top and bottom rows show respectively the frequency channels and the spatial domain templates. From left to right the start frequencies and widths of the channels were 0.5, 1, 2, 4, 8, 16, and 32 cycles/pixel. The spatial templates are analytic inverse Fourier Transform of the frequency channels sampled at the image pixel size.

4.2.7 Evaluation of the multivariate normality assumption of the channel outputs

In the multi-template channelized linear discriminant observer (MTCLDO) strategy, channel output vectors were sorted into sub-ensembles from one defect location, age, and height percentile. We verified visually that the resulting distributions of the channel output vectors in each sub-ensemble were not multi-modal and were nearly MVN distributed, as illustrated in Fig.4.4 below.

4.2.8 ROC and statistical analysis

We applied the MTCLDO to feature vectors in the sub-ensembles described above. Because younger ages have minimal anatomical differences between genders, we combined the sub-ensembles for the two genders. Thus, for newborn and 1-, 3-, and 5-year old phantoms, each sub-ensemble was comprised of 768 channel output vectors (64 realizations × 2 genders × 3 kidney sizes × 2 defect statuses). The sub-ensembles for the 10- and 15-year old phantoms were half as large as separate sub-ensembles used for each gender. The test statistics for all the sub-ensembles for all the height percentiles, genders and defect locations for a given age were pooled, ROC analysis was performed using the LABROC4 code [100], and the AUC calculated. This produced a total 5 AUC values, one for each age and for each of the 6 count levels. Bootstrapping and nonparametric analysis were used to compute 95% confidence intervals for each of these AUC values.

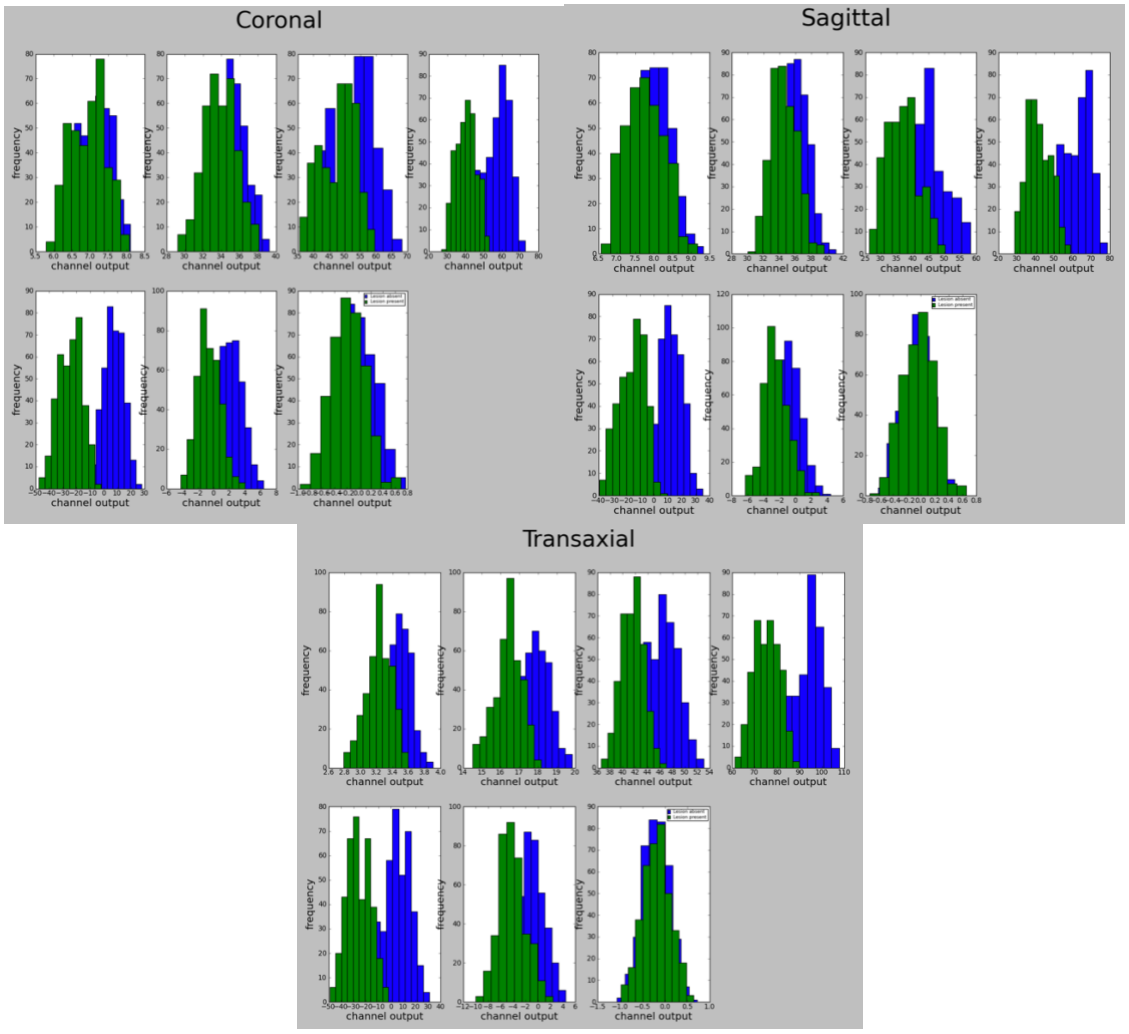

Figure 4.4. Sub-ensemble histograms of the test statistic distributions for the no-defect (green) and with-defect (blue) cases for each of the seven channels. These data are for an upper pole defect in the 50th height percentile 1-year-old phantom (including both male and female). This illustrates the near-MVN distribution of the feature vectors.

4.2.9 Relationship of AUC to AA

The goal of the following is to derive an approximate empirical relationship between the AUC and AA that can be used to fit the data from the model observer studies. When the test statistics are normally distributed under both hypotheses, the AUC under the ROC for the CHOs is related to the Hotelling SNR by [31]

$$AUC = \frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{SNR}{2}\right). \quad (4.1)$$

Rearranging the formula to express SNR as a function of AUC, we have

$$SNR = 2erf^{-1}(2AUC - 1). \quad (4.2)$$

In a binary classification task where the two classes have the same covariance matrices, the SNR of the Hotelling Observer test statistics can be expressed as

$$SNR^2 = \Delta\bar{\boldsymbol{v}} K_{\hat{v}}^{-1} (\Delta\bar{\boldsymbol{v}})^T, \quad (4.3)$$

where $\Delta\bar{\boldsymbol{v}}$ is the difference in the ensemble mean difference of the two classes. Then, we can re-write the above formula using formalism introduced by Barrett [28] to replace the total covariance as a sum of the object covariance matrix and quantum noise covariance matrix:

$$SNR^2 = \Delta\bar{\boldsymbol{v}}(\langle K_a \rangle_f + \langle K_{n|f} \rangle_f)^{-1} (\Delta\bar{\boldsymbol{v}})^T, \quad (4.4)$$

where $\langle \mathrm{K_a} \rangle_\mathrm{f}$ represents the object variability, which includes the effects of anatomical, uptake and count level variability from patient to patient. Here, count level is proportional to the product of AA and acquisition duration for a given patient and imaging system. In (4), $\langle \mathrm{K_{n|f}} \rangle_\mathrm{f}$ denotes the contribution of quantum noise to the ensemble covariance matrix of the reconstructed images. The subscript f denotes averaging over all objects in the sub-ensemble.

Suppose we now change the noise level by scaling the AA by n, such that $\boldsymbol{v} = n\boldsymbol{v}$. Then,

the SNR can be estimated as follows: [47]

$$SNR^2 = n\Delta\overline{\boldsymbol{v}}(n^2\langle K_a\rangle_f + n\langle K_{E|f}\rangle_f)^{-1}n(\Delta\overline{\boldsymbol{v}})^T. \tag{4.5}$$

We now replace n with the AA and assume that the vector $\Delta\overline{\boldsymbol{v}}$ can be replaced with a scalar K_1, representing the mean signal difference, $\Delta\overline{\boldsymbol{v}}$, and the proportionality constant relating n and AA. Similarly, we assume that the two covariance matrices can be replaced by the scalars K_2, representing the object variability noise, $\langle K_a\rangle_f$ and K_3, representing the quantum noise, $\langle K_{E|f}\rangle_f$. This gives

$$SNR^2 = \frac{AA \times K_1}{AA \times K_2 + K_3}. \tag{4.6}$$

Rearranging the formula to express SNR in terms of a function of AUC yields

$$AA = \frac{SNR^2 \times K_3}{K_1 - SNR^2 \times K_2}. \tag{4.7}$$

Combining 6 and 7 gives a relation between the detectability index (SNR^2) and AA:

$$SNR^2 = \frac{AA \times K_1}{AA \times K_2 + K_3}. \tag{4.8}$$

Equations 8 and 1 can be combined to give AUC as a function of AA:

$$AUC = \frac{1}{2} + \frac{1}{2}\,\mathrm{erf}\left(\frac{\sqrt{\frac{AA \times K_1}{AA \times K_2 + K_3}}}{2}\right). \tag{4.9}$$

Inverting the above formula to express AA in terms of AUC yields:

$$AA = \frac{(2erf^{-1}(2AUC - 1))^2 \times K_3}{K_1 - (2erf^{-1}(2AUC - 1))^2 \times K_2}. \tag{4.10}$$

Note that the relative size of the constants K_2 and K_3 indicates the degree that the SNR is limited by quantum noise rather than anatomical variability. It should also be noted that (4.10) is not a rigorous relationship in the sense that it ignores the vector and matrix matures of the defect and covariance matrices. However, as will be shown below, it is suitable for fitting AUC values as a function of the AA, and thus is practically useful.

4.3 Results

The results from the IQ studies are summarized in Fig. 4.5, which shows the AUC for each phantom plotted as a function of the percentage of the AA obtained from the consensus guidelines [13]. Note that the guidelines do not result in the same IQ (as measured by the AUC) for the 100% count level. In this sense, they are sub-optimal. The ultimate goal of this work was to provide a user with the AA needed to give the desired objectively-measured task-based image quality, as specified by the AUC, before imaging. The data in Fig. 4.5 provides a way to do this. The analytic expression relating AUC to AA derived above and given by Equation (4.10) was fit to the data in Fig. 4.5. The results of this fit are shown in Fig. 4.6 for all the patient ages. Note that the fits are visually quite good, and the correlation coefficients are better than 0.99.

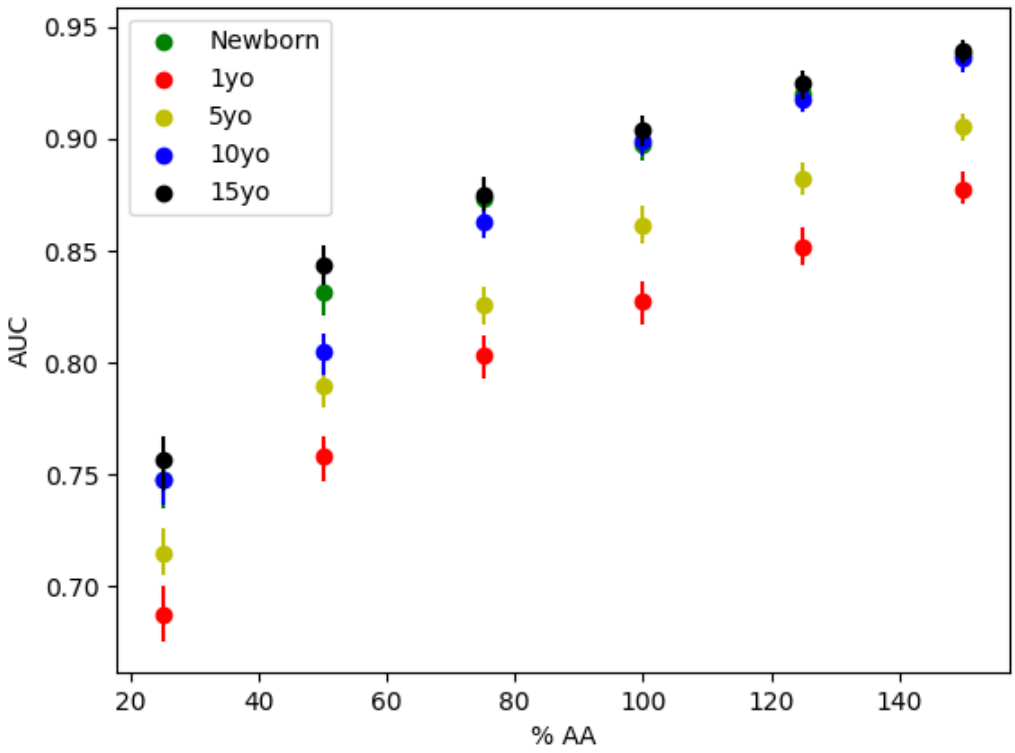

Figure 4.5. The area under the ROC curve (AUC) vs. percent AA plot for all the patient ages. The error bars are the 95% confidence intervals estimated using bootstrapping.

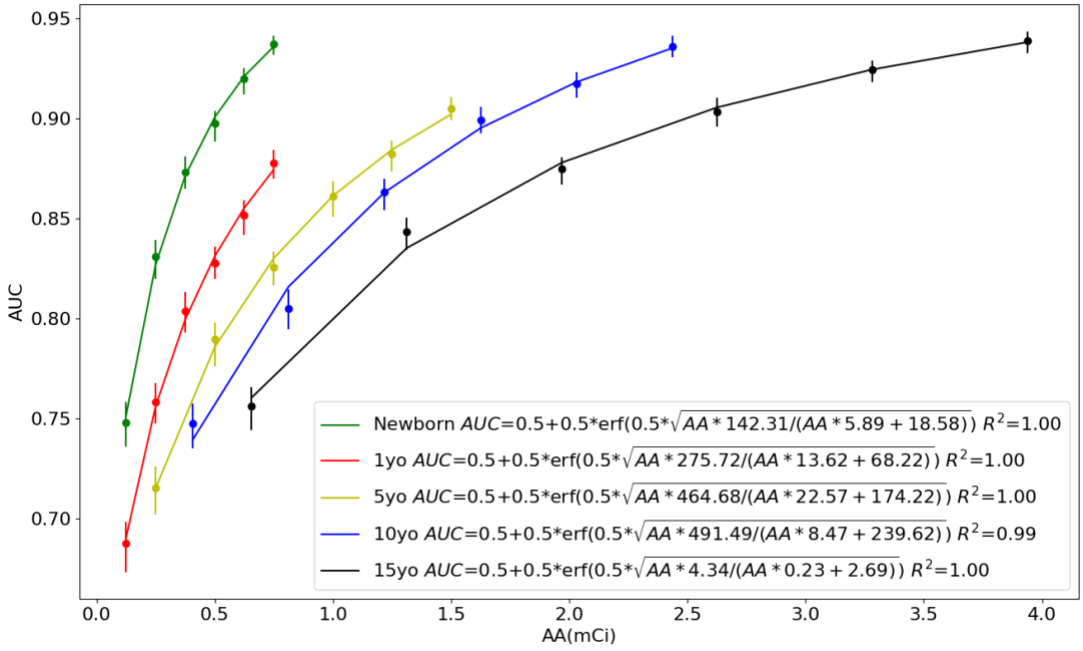

Figure 4.6. AUC vs. AA curves and their fitted functions. The AUC was fitted to the theoretical relationship, as specified in equation 4.9, relating AUC to the mean signal difference (K_1), object variability noise (K_2) and quantum noise (K_3), and AA.

There was a monotonic and modestly saturating increase in AUC with AA, indicating that defect detectability was limited by quantum noise and the effects of object variability were modest over the range of count levels studied. The AA for a given value of the AUC increased with age. The curves in Figure 4.5 indicate that, for the current guidelines, the newborn and 10-year and 15-year phantoms had similar image quality for the same fraction of the AA suggested by the North American expert consensus guidelines, but the 5-year and 1-year phantoms had lower image quality. The fitted functions provide an analytical relationship between AUC and AA, and could potentially be used to determine the AA required to give a desired AUC for a given patient weight.

In previous work [8], we have shown that there were variations in image quality among phantoms with different weights but the same height. In [9], we showed data that suggested that height was not sufficient to explain variations in image quality for phantoms with the same weight over a range of anatomical variations. However, girth (circumference) at the level of the kidneys provides a more consistent correlation. To demonstrate the correlation between girth and the AUC values, we measured the patient girth of each of the phantoms and averaged them over height percentiles within one age. In clinical practice, patient girth could be estimated prior to imaging using a tape measure or from a previous CT image, if available. Fig. 4.7 shows a comparison of AA vs. girth and AA vs. weight at a fixed AUC for all the patient ages. The colored lines connect the nearest phantoms in age. These data indicate that the relationship between girth and AA is more robust than it is between weight and AA. The Pearson product-moment correlation coefficients between AA and weight and girth were 0.941 and 0.985, respectively. This again demonstrates that girth may be more robust for estimating the AA needed to provide a constant image quality.

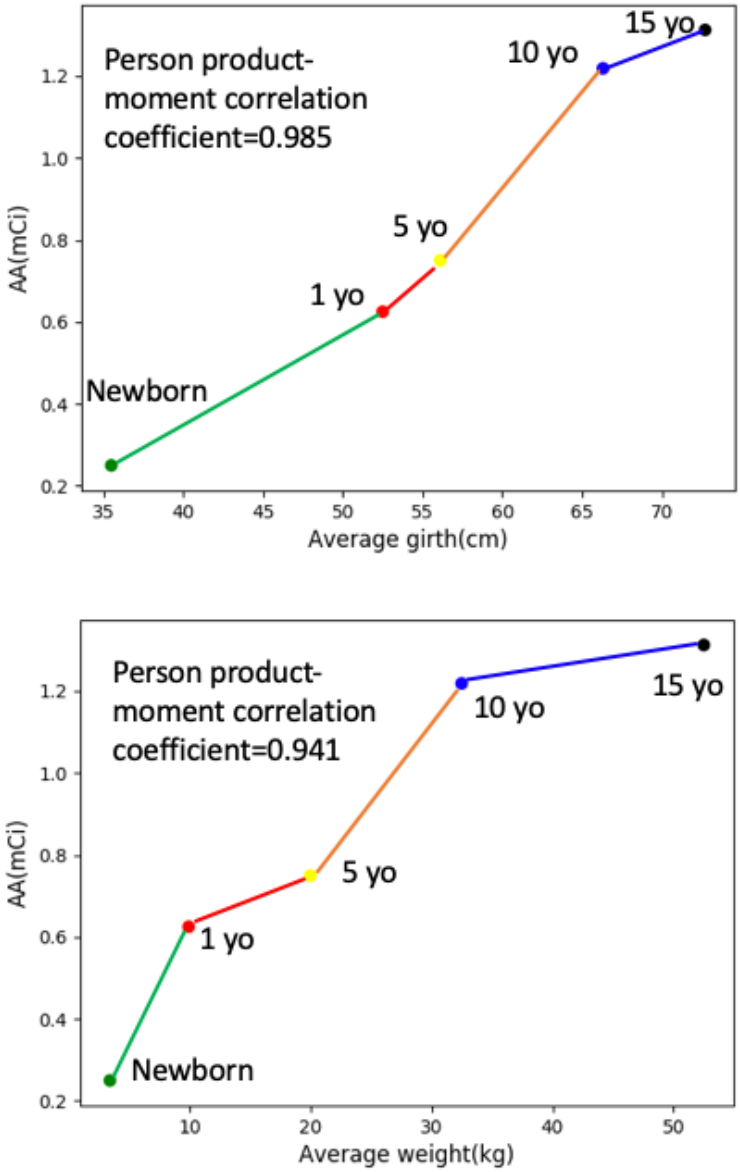

Figure 4.7. AA vs. patient girth (top) and weight (bottom) at a fixed AUC of 0.84.

4.4 Conclusion

This study demonstrated that the current consensus guidelines, which scale activities based on patient weight subject to minimum and maximum activity constraints, do not give the same image quality for patients with different weights. Further, this study provided a relationship

between diagnostic image quality, as measured by AUC, and AA for ^{99m}Tc-DMSA pediatric SPECT for a set of phantoms having different weights. These fitted functions could potentially be used to determine the appropriate AA for desired level of image quality for a given patient weight. However, the data suggest that patient girth at the level of the kidney may ultimately be a better factor to use than weight when selecting AA for this imaging task.

Chapter 5

DeepAMO: a multi-slice, multi-view anthropomorphic model observer for visual detection tasks performed on volume images

5.1 Introduction

Often, the quality of a medical image is measured in terms of the physical properties of the image, such as image contrast, spatial resolution, and noise level [33]. Fidelity-based measures, such as root mean squared error (RMSE), peak signal-to-noise ratio (PSNR), and structural similarity index (SSIM), have also been widely used in the medical imaging community. These measures are appealing because they are relatively easy to compute, have straightforward physical interpretations, and can provide objective quantitative measures of image quality. However, they are not directly related to the diagnostic task that is performed with the images, and thus may not be clinically relevant. Clinically relevant image quality assessment should be with respect to the task that is to be performed [26-32]. Ideally, the observers would be drawn from the population of people performing the task, i.e., for medical images, a radiologist or nuclear medicine physician. However, in practice, especially in large-scale developmental research studies, the use of human observers (and especially physicians) is too time-consuming, inconvenient, and expensive. Thus, a great deal of work has gone into the development of anthropomorphic model observers that predict human observer performance [34-37].

Task-based measures of image quality based on model observers has been advocated by several investigators over the years, starting from Harris [101], and including Hanson and Myers [102], Wager et al. [103], Judy et al. [104], and Myers et al. [34, 105]. However, existing model observers are often not directly applicable to diagnostic tasks [106]. For example, as described below, commonly-used model observers are strictly valid only for signal-location-known (exactly and statistically) tasks. In addition, while these observers predict rankings of human observer performance, they often require the use of concepts such as internal noise to match the absolute performance of human observers.

Of the existing anthropomorphic observer models, the channelized Hotelling observer (CHO) has been the most widely used as a substitute for human observers in signal-location-known tasks in nuclear medicine imaging research [40]. Please refer to section 2.4.3.3 for a detailed introduction to the CHO as well as discussions of its limitations.

Another gap between current anthropomorphic observers and the real clinical task is that current model observers have been primarily designed for analyzing 2D images. By contrast, many clinical tasks require the interpretation of 3D datasets. This often involves reviewing sequences of 2D slices in 3 orthogonal orientations (coronal, sagittal, and transaxial). Existing multi-slice [107, 108] or 3D model observers [109-113] are either for SKE tasks only or single-orientation SKS tasks [107].

In this paper, we propose a novel deep learning-based anthropomorphic model observer (DeepAMO) that evaluates multi-orientation, multi-slice image sets to model the clinical diagnostic process of a radiologist or nuclear medicine physician in a clinically realistic 3D defect detection task. The DeepAMO was evaluated on an SKS/BKS tasks using a realistic anatomical background with variation in organ uptake and defect position (and thus orientation and shape).

We also propose a novel calibration method that 'learns' the underlying distribution of the human observer rating values (including the internal noise) using a Mixture Density Network. Note that in this context a rating value is the raw data from human observer study and is a numeric value expressing the observer's level of confidence that a defect is present or absent in a given image. The entire network is trained using human observer rating values so that the output, when applied to an input image volume, is a rating value designed to reproduce the performance of human observers.

A human observer study was conducted using the volumetric display format routinely used at Boston Children's Hospital (BCH) for clinical interpretation. Quantitative comparisons of the performance between the DeepAMO and human observer are provided in the results section.

5.2 Materials and methods

Image quality in this work was measured in terms of performance on the task of detecting renal functional defects in 99mTc-DMSA SPECT. The images used were simulated based on an anthropomorphic digital phantom of 5-year-old (a typical age in DMSA imaging). The phantom and simulation methods are described in [1]. The simulation modeled administered activities (and thus noise levels) based on the North America Consensus Guidelines[114]. Task performance was evaluated using both human observers and the proposed DeepAMO. Both of these observers produced a set of rating values for images where the true defect status was known. These rating values were analyzed using receiver operating characteristic (ROC) analysis methods [115]. The area under the ROC curve (AUC) served as a figure of merit for task performance.

5.2.1 Materials and methods

The projection data for this study were generated using the Advanced Laboratory for Radiation Dosimetry Studies (ALRADS) UF NHANES-based phantom [116]. The pediatric phantom used was developed at the University of Florida based on demographic data from the CDC's National Health and Nutrition Examination Survey (NHANES) data [79]. For this study, we used a 5-year-old male phantom with average girth and kidney size. The phantom was digitized using 0.1 cm cubic voxels. Activity uptake in the kidneys was modeled using data from a single imaging time point (3 hours post-injection). A dataset of 47 patients acquired at the BCH was used to estimate the means and standard deviations of kidney uptake in units of activity.

The model previously described in [2, 117] was used to simulate defects in the cortical wall of the right kidney consisting of volumes of reduced uptake consistent with focal, acute pyelonephritis. The defects were created at random locations (excluding the area close to the renal pelvis) along the cortical wall. Based on input from an experienced pediatric nuclear medicine specialist, we selected a defect volume of 0.5 cm3 as a defect size that is clinically relevant for the 5-year-old.

Using this model, we created four randomly located focal transmural renal defects at each of the following macro locations on the right kidney cortex: upper pole, lower pole, and lateral. There was a total of 12 random locations for the defects generated in this study, modeling an SKS task. We simulated noise-free projection data for the renal cortex, medulla, pelvis, liver, spleen, and background (including all other organs), modeling the physics and acquisition parameters appropriate for ^{99m}Tc renal SPECT. The renal activity and relative activity concentrations for structures inside the kidney (the renal cortex, medulla, and pelvis) were randomly sampled from truncated Gaussian distributions with the means, standard deviations, minima, and maxima derived

from 47 sets of patient data acquired at BCH. Parameters for the distributions can be found in [2]. Each single-organ projection was scaled by the product of administered activity (AA), acquisition duration, and scanner sensitivity. The projections were generated using an analytic projection code that modeled attenuation, the spatially varying collimator-to-detector response [81], and object-dependent scatter [82]. The code has been previously validated by comparison to Monte Carlo and experimental projection data for imaging of a variety of radionuclides [90-98].

In this study, the projections were simulated to model a Siemens low-energy, ultra-high-resolution (LEUHR) collimator used routinely at BCH for pediatric DMSA studies. Each single-organ projection dataset was generated at 120 projection views over a 360° body-contouring orbit with a 0.1-cm projection bin size and then collapsed to a bin size of 0.2 cm. A model of the patient bed obtained from a CT scan of the bed of a Siemens Symbia SPECT/CT system was added to the attenuation map of each computational phantom. Noise-free projection images of the entire phantom were obtained by summing the individual sets of scaled organ projections. Noisy projections were created by simulating Poisson noise using a Poisson pseudo-random generator.

A total of 384 projection images were thus generated, comprised of 16 uptake realizations × 12 defect locations × 2 defect statuses (present or absent). The mean (noise-free) activity distribution was statistically independent for each of these 384 projection images since the kidney uptake and the activity concentration ratio of the cortex to the medulla plus pelvis activity were randomly sampled.

We followed the clinical reconstruction protocol routinely used at BCH. Projection images were reconstructed using the OS-EM iterative reconstruction algorithm with compensation for the geometric collimator-detector response and post-filtered with a Gaussian filter with an FWHM of 5 mm. The reconstructed images were then interpolated and formatted to match the volumetric

image display used at the BCH. In this display, 10 coronal, 20 sagittal, and 18 transaxial images with sizes of 96 × 96 pixels were generated. These composite images were used for training and testing of the proposed model observer and the human observers. Windowing was used to map the image pixel values to a range of 0 to 255. A sample of BCH's volumetric image display is shown in Fig. 5.1.

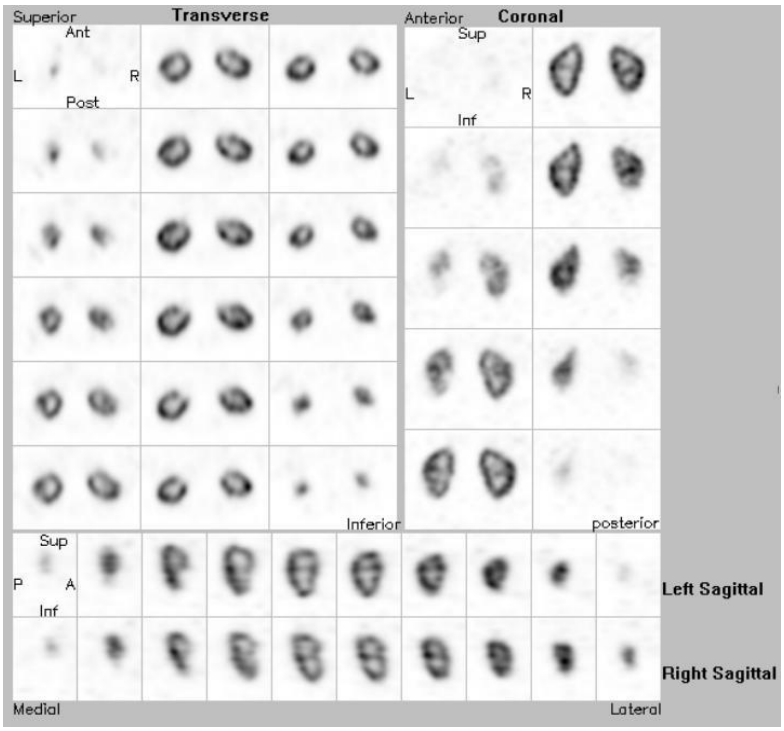

Figure 5.1. A sample 48-slice image shown in the volumetric display format routinely used in clinical practice at the Boston Children's Hospital.

5.2.2 Proposed model observer: overview

The DeepAMO is designed based on a hypothetical model of the image interpretation process of a human observer. One alternative of this approach would be to let the neural network 'learn' how humans interpret 3D image volumes from the data. For example, the most direct approach would be to input the 3D image volume data into a fully connected network, and then to train that network directly with human observer rating values. Such a network would have a large

number of parameters. Since each trial (reading of a set of images by a human observer) provides a single scalar rating value, it provides relatively little information for training the network. A very large number of input rating values would thus be required. Since the rating value data is very expensive to obtain, we have divided the network into stages that are designed to require less human-observer training data. The division of the model is based on how humans interpret the images, as will be described below. The first two stages do not require human observer training data, and the third one maps a low-dimensional feature vector to a scalar rating value.

We hypothesize that a human observer interpreting an image first scans over the orthogonal slices to identify suspicious abnormalities in single slices. If a defect is suspected to be present in one slice (of a particular orientation), the observer confirms that on adjacent slices. The observer would confirm that a defect in one orientation is seen in the other two orthogonal orientations. We suppose that the observer would have more confidence in the presence of a defect if it is found in at least one other orientation.

Thus, we propose to implement this decision-making process in 3 sequential stages. In stage 1, we use a segmentation network to identify defects in three orthogonal slice views. The segmentation is performed using groups of 3 adjacent slices. In stage 2, we use deterministic algorithm that confirms the presence of defects in the 3 orthogonal views and generates a low-dimensional feature vector. In stage 3, we use a Mixture Density Network to learn the mapping of feature vector to rating value, thus calibrating the DeepAMO to reproduce human observer performance.

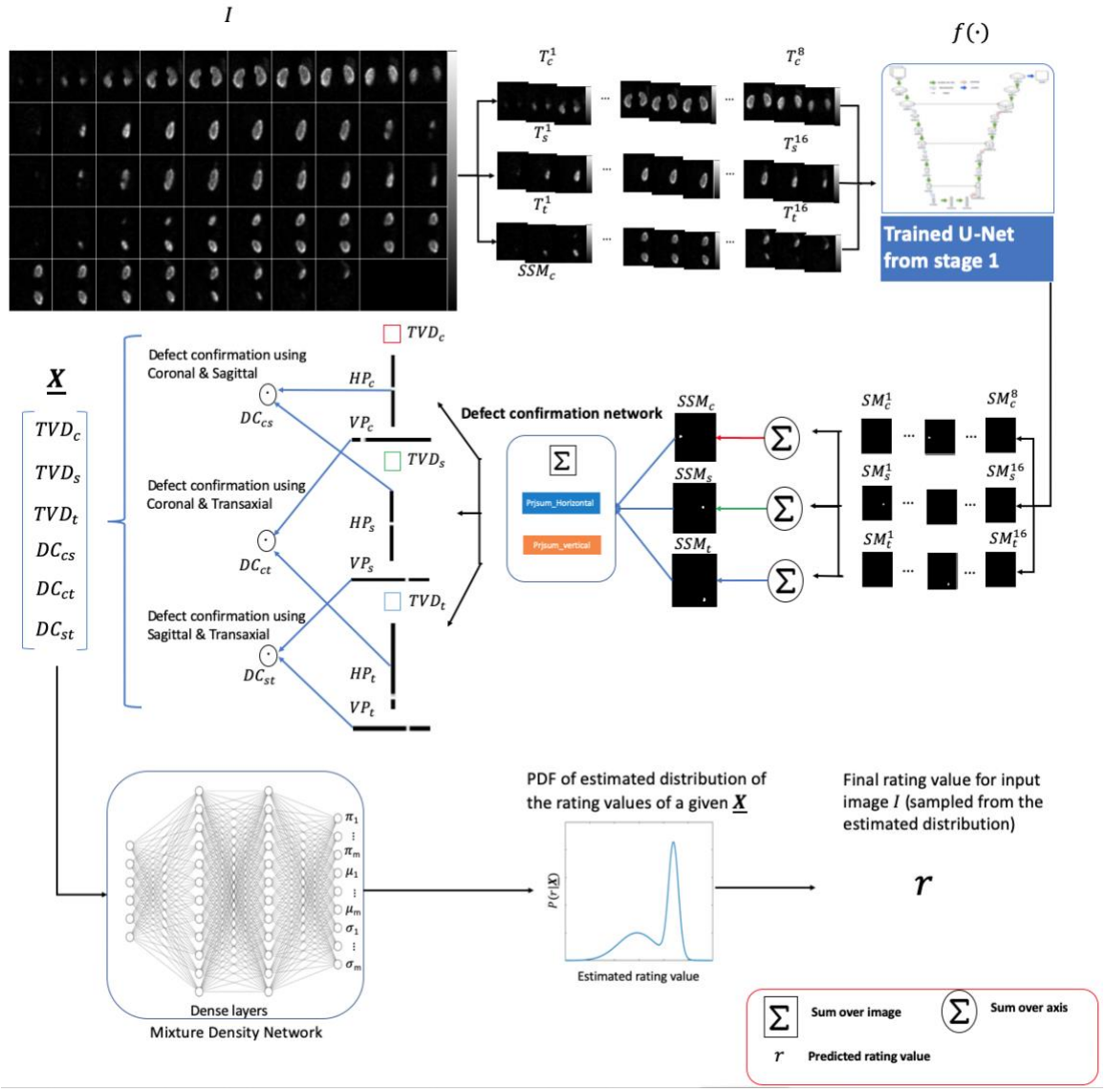

Figure 5.2. A schematic of the proposed model observer: DeepAMO. I is the multi-slice, multi-view input image, T_k^j is the triad where $k \in (c, s, t)$ represents the slicing direction and $j \in [1, N-1]$, where N is the number of slices in each orientation. SM_k^j is the output segmentation mask for each triad T_k^j. TVD_k is the total volume of the defect seen in each slicing direction computed by summing the corresponding SSM_k. SSM_k is the summed segmentation mask along each slicing direction k. HP_k and VP_k are horizontal and vertical projection of the corresponding SSM_k. DC_{cs}, DC_{ct}, and DC_{st} are the three defect confirmation scalars from the defect confirmation network.

5.2.3 Proposed model observer: architecture

A schematic of the proposed DeepAMO is shown in Fig. 5.2. The input to the segmentation network was the same set of slices used in the previously described volume display used in clinical

practice, which consists of multiple slices in each of the three orientations: coronal, sagittal, and transaxial. Mathematically, the slice, $S_k^i(m,n)$, and input composite image, $I(m,n,q)$, are related as follows

$$I\left(m,n,q_k^i\right) = S_k^i(m,n). \tag{5.1}$$

In (1), q_k^i is the index number for the ith slice in the slicing direction $k \in (c,s,t)$, and m, n, and q are pixel indices for the x-, y-, and z-axis, respectively.

For each orientation, $N-2$ (N + 1 slices in each orientation) triads are generated: the first and last slices cannot act as the central slice for a triad. The jth triad in the slicing direction k is:

$$T_k^j(m,n,q) = \left\{S_k^{i-1}(m,n), S_k^i(m,n), S_k^{i+1}(m,n)\right\} \tag{5.2}$$

$$i \in [0,N], j \in [1,N-1].$$

The output segmentation mask (SM) of each triad is a 2D binary mask of pixels thought to be in the defect. The SMs along each orientation are summed to form a summed segmentation mask (SSM) in order to enhance the defect signal(s) that is (are) present in that orientation. That is:

$$SM_k^j(m,n) = f\left(T_k^j(m,n,q)\right), \text{ and} \tag{5.3}$$

$$SSM_k(m,n) = \sum_{j=1}^{n_k} SM_k^j(m,n), \tag{5.4}$$

with j the triad number and k the slicing direction. $T_k^j(m,n,q)$ and n_k represent the jth triad and the number of triads in slicing direction k, respectively. Here, $f(\cdot)$ denotes the segmentation network.

We propose to implement the process of confirming defect presence in other slicing directions, by projecting and comparing defect information from different slicing directions, through a defect confirmation network. Specifically, this is implemented by projecting (i.e.,

summing) each SSM_k vertically and horizontally and calculating the dot products between the resulting horizontal projections (HP) and vertical projections (VP) from different slicing directions. The HPs and VPs are derived as follows:

$$HP_k(n) = \sum_{m=0}^{M-1} SSM_k(m,n), \text{ and} \tag{5.5}$$

$$VP_k(m) = \sum_{n=0}^{N-1} SSM_k(m,n), \tag{5.6}$$

with M and N being the number of pixels in the x- and y-axis directions, respectively.

The projection is constructed so that the projections from the different slicing directions are along the same direction in space. To understand this, consider that any two views always share a common axis, and, by projecting the two views onto this common axis, we can confirm information about defect location that is compatible. For example, consider an L-shape object in a 3D space (Fig. 5.3). By projecting the sagittal and transaxial views vertically, we get two 1D vectors that both contain information about the object's maximum length along the horizontal axis. If the dot product between the two 1D vectors is large, then the object is present at the same location in that direction for both slicing directions. Likewise, we can confirm the object's location along the other two directions via the same projection and dot product operations. This process yields 3 scalar values, representing the defect agreement along the x, y, z-axis, respectively. We named these 3 scalar values the defect confirmation (DC) scalars. They are derived from the HPs and VPs from different slicing directions as follows

$$DC_{cs} = HP_c(n) \cdot VP_s(m), \tag{5.7}$$

$$DC_{ct} = HP_t(n) \cdot VP_c(m), \text{ and} \tag{5.8}$$

$$DC_{st} = VP_t(m) \cdot VP_s(m). \tag{5.9}$$

The DC scalars are concatenated with the total volume of the defect (TVD) seen in each slicing direction to form a single feature vector. The TVD from each slicing direction is computed as follows:

$$TVD_k = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} SSM_k(m,n). \tag{5.10}$$

The resulting 6-element concatenated feature vector is then sent to a Mixture Density Network (MDN) [118] to generate the rating (test statistic) value. The dense layers in the MDN are meant to model the process of a human making the final decision using combined information from the different directions. The output of the MDN is the set of parameters of a statistical distribution, in this case a Gaussian Mixture Model, as described below.

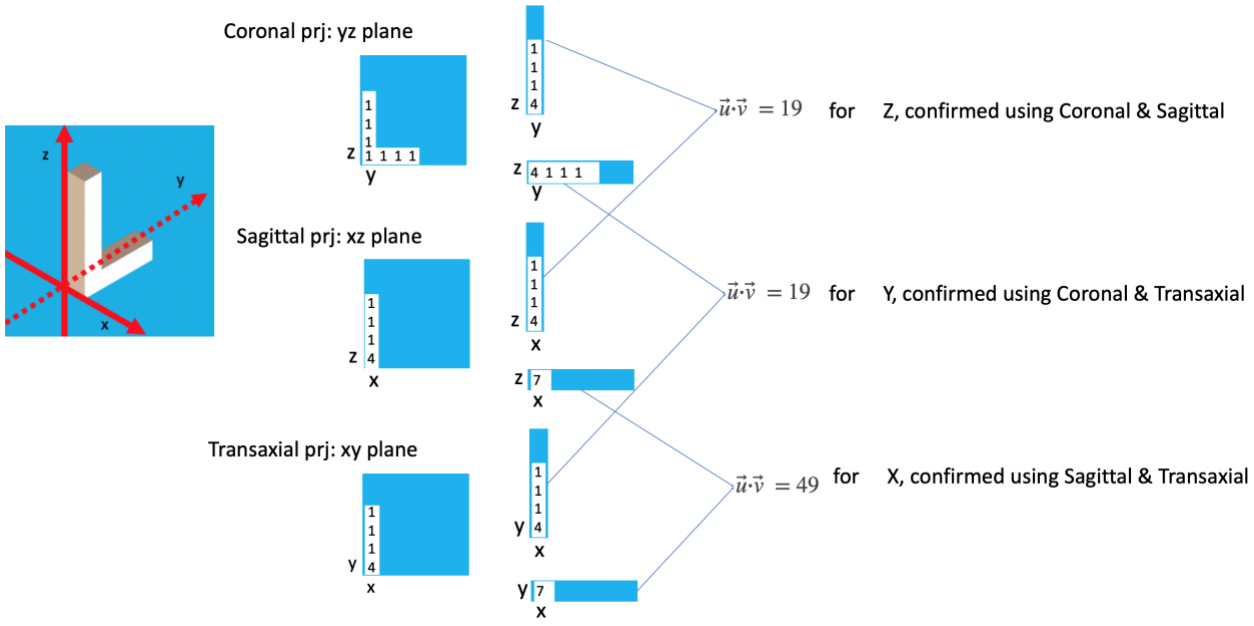

Figure 5.3. An illustration of the process of confirming the defect from different views using projection and dot product in 3D space.

5.2.4 Calibration to human observer data via a mixture density network

For defect detection tasks, the observer performance is usually measured by the AUC, which ultimately depends on the underlying distribution of the rating values given by the observer. Thus, for the purposes of replicating an observer's AUC, we propose to directly learn the mapping of feature vectors to the distribution of the rating values. We hypothesize that more training and testing samples would help better capture the underlying rating value's distribution. However, demonstrating the equivalence of the distributions is a task requiring a large number of rating values. In addition, it is unclear what level of agreement between the true and modeled distribution is required. Thus, we are focusing in this work on verifying that the model observer can replicate the AUC values obtained from the set of rating values resulting from an observer study.

A mixture density network (MDN) was chosen for the task of mapping the input feature vector into a rating value in order to model the fact that a human observer will give a different rating value for the same input image. The MDN provides parameters of a distribution that can then be sampled to provide multiple, continuously valued ratings from a single set of input feature vectors. This can be useful during testing of the DeepAMO to reduce sampling error.

Typically, an MDN learns an entire probability distribution for the output by modeling the conditional probability distribution of the target data conditioned on the input data. In our case, the desired conditional probability distribution is $P(r|\underline{\boldsymbol{X}})$, where is $\underline{\boldsymbol{X}} = [x_{1\ldots}x_6]$ a 6-element feature vector and r is a (continuous) human observer rating value for a given input feature vector. For the purpose of modeling any arbitrary probability distribution, the MDN uses a Gaussian mixture model as the conditional probability density function, which can be represented as a linear combination of kernel functions in the form

$$P(r|\underline{\boldsymbol{X}}) = \sum_{i=1}^{m} \pi_i(\underline{\boldsymbol{X}})\, \phi_i(r|\underline{\boldsymbol{X}}), \tag{5.11}$$

where m is the number of components in the mixture and $\{\pi_i(\underline{\boldsymbol{X}})\}$ is the set of mixture coefficients for the kernel functions, which sum to 1. The set $\{\pi_i(\underline{\boldsymbol{X}})\}$ is derived from the output of the MDN and is converted to a set of probabilities as follows:

$$\pi_i(\underline{\boldsymbol{X}}) = \frac{\pi_i}{\sum_{i=1}^{m} \pi_i}, \tag{5.12}$$

with π_i the output from the last dense layer, as shown in Fig. 5.3. The kernel functions, $\{\phi_i(r|\underline{\boldsymbol{X}})\}$, are in the form of Gaussian distributions

$$\phi_i(r|\underline{\boldsymbol{X}}) = \frac{1}{\sigma_i(\underline{\boldsymbol{X}})\sqrt{2\pi}} \exp\left(-\frac{\left(r - \mu_i(\underline{\boldsymbol{X}})\right)^2}{2\sigma_i(\underline{\boldsymbol{X}})^2}\right), \tag{5.13}$$

where $\sigma_i(\underline{\boldsymbol{X}})$ and $\mu_i(\underline{\boldsymbol{X}})$ are the estimated standard deviation and mean for the input feature vector, $\underline{\boldsymbol{X}}$, and they come from the output of the last dense layer. Note that $\{\pi_i(\underline{\boldsymbol{X}})\}$is a function of $\underline{\boldsymbol{X}}$. So, $\{\pi_i(\underline{\boldsymbol{X}})\}$ can also be regarded as a set of prior probabilities of the target data.

In training, the loss is computed using the human observer rating value, r_{true}, and the predicted mixture distribution $P(r|\underline{\boldsymbol{X}})$ from the MDN as follows

$$L = -logP(r_{true}|\underline{\boldsymbol{X}}). \tag{5.14}$$

In testing, a rating value is predicted by first randomly sampling the mixing coefficients and then sampling from the Gaussian distribution corresponding to that sampled mixing coefficient with its corresponding mean and standard deviation. Multiple sample rating values can be generated to improve the uncertainty in AUC values calculated from the testing data.

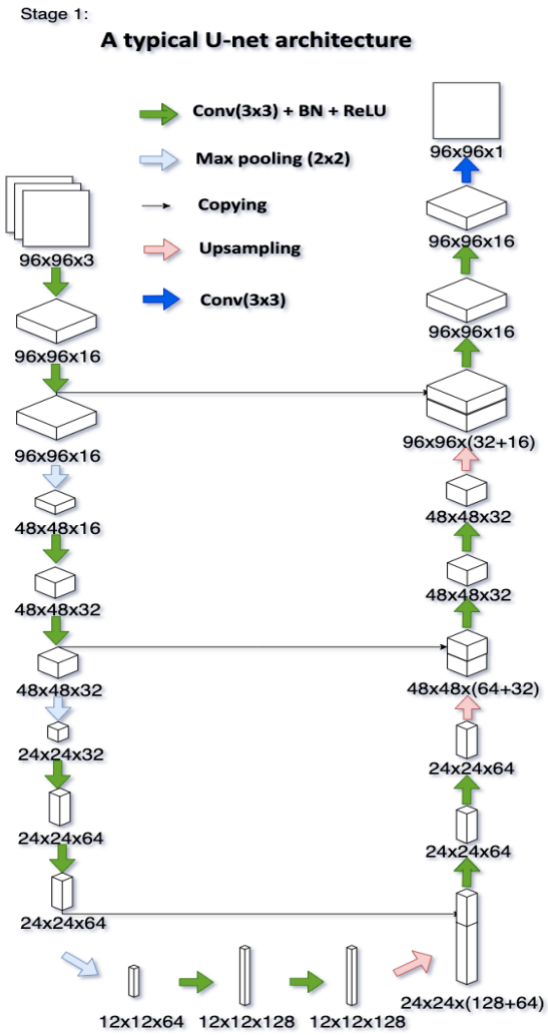

Figure 5.4. Segmentation network architecture used in this study

5.2.5 DeepAMO performance on unseen images

To estimate the number of images needed to train the DeepAMO, we used simulated feature vectors and rating values to train and test the MDN. The criterion for judging the number of images to be sufficient is the statistical confidence level needed in comparing AUC values between the proposed model and human observer. We assumed the elements of the feature vectors and the rating values follow a (unimodal or multi-modal) Gaussian distribution.

The feature vectors were simulated by first generating values for the TVD_k, one for each orientation. Each TVD_k was assumed to be mutually independent and was generated by sampling from independent Gaussian distributions. The sampled TVD_k values were then used to calculate the means and standard deviations of the DC scalars, which were also assumed to follow a Gaussian distribution.

$$\mu_{cs} = TVD_c \times TVD_s, \tag{5.15}$$

$$\sigma_{cs} = \frac{\mu_{cs}}{3}, \tag{5.16}$$

$$\mu_{ct} = TVD_c \times TVD_t, \tag{5.17}$$

$$\sigma_{cs} = \frac{\mu_{cs}}{3}, \tag{5.18}$$

$$\mu_{st} = TVD_s \times TVD_t, \text{ and} \tag{5.19}$$

$$\sigma_{st} = \frac{\mu_{st}}{3}. \tag{5.20}$$

The rating values of each feature vector were sampled from multi- or uni-modal Gaussian distributions. The distribution parameters for these simulated rating values were derived qualitatively from distributions of rating values from human observer studies and are shown in Table 5.1. For each feature vector, we then sampled N rating values from the assumed distribution to simulate the appropriate level of inter- or intra-observer variability in the data. Specifically, in this work, we sampled 2 rating values for each feature vector. So, there were 15,000 (2,500 × 3 feature vector types × 2 repeated samples) feature vector and rating value pairs in total for the case that had 2,500 samples/feature vector type, and 30,000 in total for both the defect-present and defect-absent cases.

In the simulation experiment, we generated 3 types of feature vectors for each class (defect-present and defect-absent): definitely-present, equivocal, and definitely-absent, reflecting different levels of user confidence in making the decision. For example, the feature vectors that belong to

the definitely-present type in the defect-present class were generated by sampling 3 large values for the 3 TVD_ks, modeling a high level of success of the segmentation network in detecting the defect in slices from all 3 orientations. The other two types (equivocal and definitely-absent, respectively) contained 2 and 1 large values (assigned randomly to any of the three orientations) in the TVD_k s to simulate different degrees of success in detecting the defect in the three orientations.

Table 5.4. Summary of distribution parameters for the simulated rating values

Defect-present feature vector type	Definitely-yes		Not-sure		Definitely-no	
Rating value means	7	10	2	4	-3	
Standard deviation	1.2	0.2	1.2	1.2	0.2	
Component weight	0.5	0.5	0.5	0.5	1	
Defect-absent feature vector type	**Definitely-yes**		**Not-sure**		**Definitely-no**	
Rating value means	-10	-8	-2	-4	2	5
Standard deviation	0.2	1.2	0.7	1.2	0.5	0.8
Component weight	0.5	0.5	0.5	0.5	0.5	0.5

5.2.6 Training and testing of DeepAMO

The proposed model observer was trained in two stages. First, the segmentation network was trained given the ground-truth defect segmentation masks. Next, the MDN was trained using the output from the trained segmentation network and the human observer rating values.

The segmentation network was trained with triad images and their corresponding binary defect segmentation labels. Since each defect only contained about 0.5% of the kidney cortex volume, the number of defect-present triads was much smaller than the defect-absent ones, making this a highly imbalanced dataset. Thus, we adopted data augmentation of the defect-present triads to balance the training data. We enriched the data by forming an additional seven sets of raw

images and their labels by rotating each original defect-present triad image by 90, 180, and 270 degrees and flipping them and the original dataset upside down. The exponential logarithmic loss in [119] was adopted to emphasize segmentation of small structures with the best-performing weights ($\omega_{cross} = 0.2$ and $\omega_{Dice} = 0.8$).

For the segmentation network, we adopted a shallow version of the U-Net [120]. We used a shallow (in depth) network due to the relatively small amount of training data available in this study; a deeper network might be needed for a larger number of signal and anatomical variations. The architecture of the segmentation network used in this study is shown in Fig. 5.4. Gaussian noise with a standard deviation of 1.0 was added to the renormalized input image (ranges 0-255) to prevent overfitting. We searched for the optimal network capacity (depth) for the segmentation network. There was a tradeoff between producing the highest Dice score and using the smallest number of parameters. However, it was observed that there was a relatively small increase in Dice score with increased number of parameters in the tested network architectures, and the Dice scores were all reasonably high. So, we adopted the network architecture that had the smallest number of parameters and yet gave a reasonably high Dice score (0.97). The train and test datasets had 12,288 and 3,072 triads, respectively. Data augmentation was done on-the-fly. We used an Adam [121] optimizer with a learning rate of 0.001 and a batch size of 200. The training took about 2 hours (~100 epochs) to converge on a single Tesla K40 GPU.

For the MDN, the number of mixtures was chosen by visually inspecting the distribution of the target human observer's rating values. The number of mixtures was selected to be equal or greater than the number of modes observed in the distribution of the observer's rating values. For this study, we used a MDN with three fully connected dense blocks each with 128 dense units and a dropout rate of 0.5. Each dense block contained a dense layer with the above mentioned dense

units and a batch normalization layer, followed by a ReLU activation and dropout layer. The outputs from the last dense block were then connected to three dense layers which, repsectively, output the mixing coefficients $\pi_i(\underline{\boldsymbol{X}})$, means $\mu_i(\underline{\boldsymbol{X}})$, and sigmas $\sigma_i(\underline{\boldsymbol{X}})$ for the estimated distribution. The number of mixing coefficient was set to 5 since we obseverd about 5 modes in the distribution of huamn observers' rating vlaues.

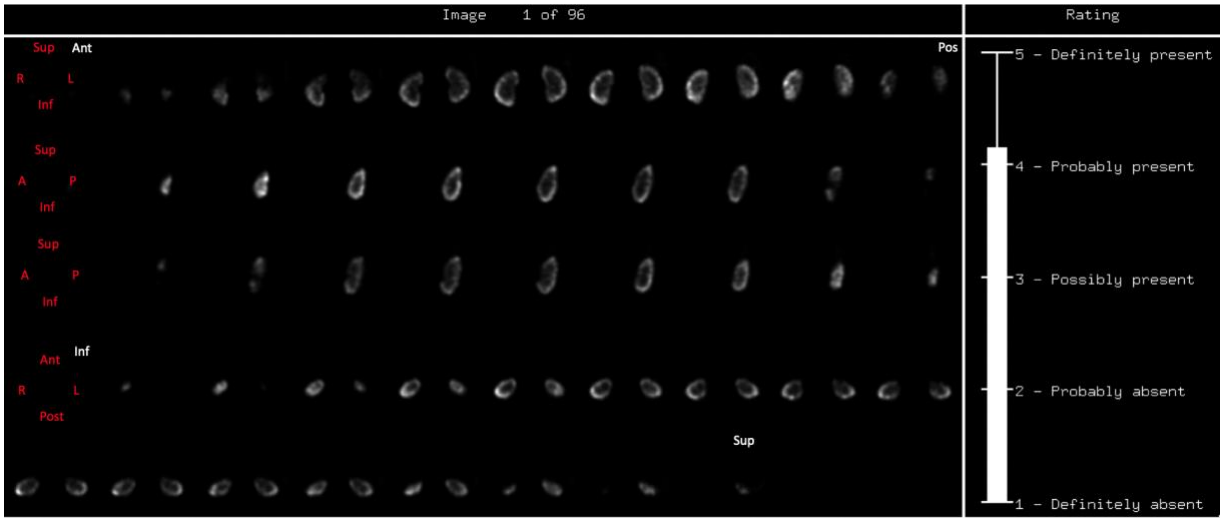

Figure 5.5. A sample image of the GUI used in the human observer study for DeepAMO

5.2.7 Human observer study

The same image display format shown in Fig. 5.1 was used in the human and model observer studies. A sample display of the human observer GUI is shown in Fig. 5.5. In the study, the observer was asked to rate their confidence that a defect was present on a continuous scale ranging between 1 to 5 (later mapped to -10 to 10), with the highest number representing the greatest confidence that a defect was present. To familiarize themselves with the display program and the nature of the clinical defect detection task, all observers participated in an initial training session comprised of 24 images. In the training session, phantom images of the kidney cortex were

provided as ground truth to the observers once their rating value was recorded. Additional training was done as described below. Rating values from the training study were not used in training the network.

Two senior medical imaging physics Ph.D. students participated in the human observer study. A total of 384 of the composite images described in section 5.2.1 were used. To simulate an SKS detection task, the train and test datasets were created without requiring a balance of defect locations. Thus, the test dataset could contain defect locations that were not present in the initial training dataset. The images were divided into an initial training set and three test blocks. The block layout for each observer is shown in Table 5.5. In each test block, a refresher set of 24 images was provided to refresh the observer's memory about the task. A total of 288 rating values was collected from each observer.

Table 5.5. Summary of human observer study block partition

Session	Initial training images	Blocks	Image/block	Total images
	24	1	24 training	24
1	0	1	24 training/96 test	120
2	0	1	24 training/96 test	120
3	0	1	24 training/96 test	384

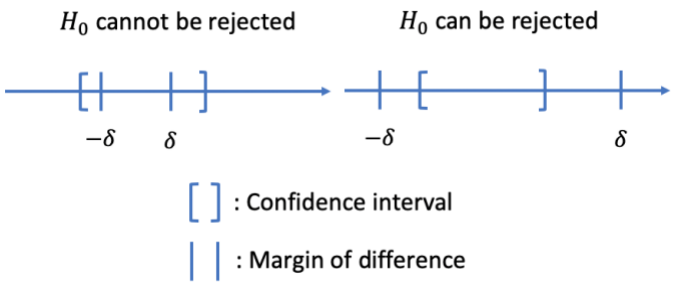

Figure 5.6. A pictorial illustration of the rejectable and unrejectable case in equivalence hypothesis testing.

5.2.8 Equivalence hypothesis testing

An equivalence statistical hypothesis test [122] was conducted to test whether the performance (as measured by the AUC) of the human observer and the proposed model observer was statistically equivalent on a defect detection task. The null hypothesis and alternative hypothesis are expressed as follows:

$$H_0: |AUC_{HO} - AUC_{MO}| = \delta \text{ and} \tag{5.21}$$

$$H_1: |AUC_{HO} - AUC_{MO}| < \delta,$$

where AUC_{HO} and AUC_{MO}, respectively, are the AUC values for the human and proposed model observer; δ is a threshold for an important difference (margin of difference) between AUC_{HO} and AUC_{MO}. The difference parameter was used as it is very difficult, if not impossible, to show statistically that two quantities are exactly equal. In addition, small differences are not practically important. The difference parameter was prespecified and is a determinant of sample size: in order to prove better equivalence (smaller δ), a larger sample size is required. In order to reject the null hypothesis, the confidence intervals of the difference of the AUCs must lie within the interval

defined by the margin of difference parameter, as described in [122] and illustrated in Fig. 5.6. For this study, we set δ to 0.043. That is, as long as the confidence intervals of ΔAUC were found to be smaller than 0.043, the null hypothesis can be rejected and equivalence of the human and model observer can be claimed.

In order to calculate the confidence intervals for the differences in the AUCs (ΔAUCs), we conducted a 5 × 2-fold cross-validation experiment using data generated by the two human observers. A total of 576 rating values (288 images × 2 observers) was used in training and testing of the proposed model observer. The data were partitioned randomly for each of the five trials, and a 50-50 train-to-test fraction was adopted. Within each trial, the train and test data were switched between the 1st and 2nd fold. We used a 50-50 split strategy to divide the data, as we assumed that the number of images in the test dataset should not be too small otherwise the distribution of rating values produced would be too coarse to represent the observer's true performance, thus resulting in unfair AUC comparisons. However, we have not investigated whether the 50-50 splitting is optimal.

5.2.9 Comparison of DeepAMO to a scanning-linear observer

A scanning linear discriminant observer (SLDO) study was conducted using the same reconstructed images as described in section 5.2.1. However, since the scanning observers cannot operate at the location on which they were trained, we had to limit the SLDO input image to only slices that could actually contain a defect. Here, it is worth noting that this input format has significantly reduced the difficulty of the clinical defect detection task by filtering out the defect-absent slices. This eliminates the chance of making a mistake, e.g., due to the presence of a noise artifact in these slices, as described in section 5.2.2.

In the SLDO study, we used a 3-slice composite image as the input. The composite image was formed by extracting the coronal, transaxial, and sagittal slices containing the defect centroid from the 3D reconstructed image. All slices had a size of 128 × 128 pixels and their defect centroid shifted to the center of the image. Samples of the defect-present and defect-absent composite image are shown in Fig 5.7. We used seven non-overlapping rotationally symmetric difference-of-mesa channels. The starting frequency and the width of the first channel was 0.5 cycles per pixel, and subsequent channels had widths that doubled and abutted the previous channel. The frequency domain channels and corresponding spatial templates are shown in Fig. 5.8.

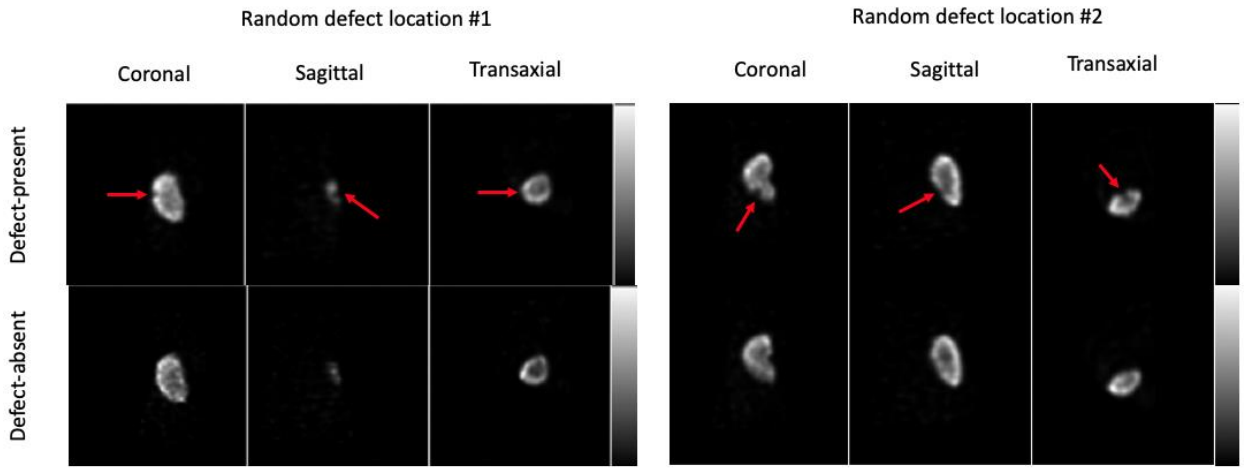

Figure 5.7. Top and bottom row shows the defect-present and defect-absent composite image at two different randomly sampled defect locations, respectively. The red arrows mark the exact location of the defect inside each slice.

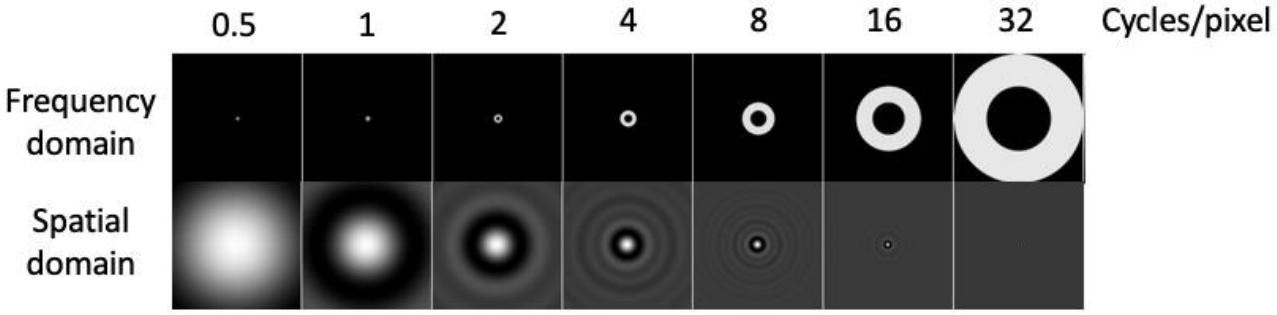

Figure 5.8. Images of the seven anthropomorphic DOM channels used in this work. The top and bottom rows show, respectively, the frequency channels and the spatial domain templates. From left to right, the start frequencies and widths of the channels were 0.5, 1, 2, 4, 8, 16, and 32 cycles/pixel. The spatial templates are the analytic inverse Fourier Transforms of the frequency channels sampled at the image pixel size.

Each of the seven spatial domain templates was applied to each of the 3 images (transaxial, sagittal, and coronal) to give a 21-element feature vector. Each element in the resulting feature vector was obtained by taking the dot product of a spatial domain template with an input composite image. These feature vectors served as inputs to train and test the SLDO as described below.

To apply the SLDO on a test image, we first generated N (N = number of signal variations) feature vectors of each test image, corresponding to features taken at the N different defect locations. Then, we trained a different SLDO on the feature vectors at each of the 12 potential defect locations. Then, for each test image, we applied each of the 12 SLDOs to the feature vectors from each of the potential defect locations, producing a set of 12 test statistics. We then applied the argmax operator to select the largest such test statistic, and this served as the test statistic for this test image. We used a leave-one-out training-testing strategy. In this strategy, one feature vector was left-out (i.e., not used in the training), and the remaining vectors were used to train the observer. In our case, the feature vector corresponding to the ground-truth defect location of the test image was left out in training the SLDO for that defect location. The trained SLDO was then applied to the left-out vector to produce a test statistic for that defect location. ROC analysis was

performed on the test statistics using the LABROC4 code [100], and the AUC calculated. Bootstrapping and nonparametric analysis were used to compute 95% confidence intervals for the AUC value.

A separate human observer study was conducted using the same input format (3-slice composite image) as was used in the SLDO study. Again, two senior medical imaging physics Ph.D. students participated in the human observer study. A total of 384 of the composite images as described above were used. The same block layout as in the human observer study for DeepAMO was used in the human observer study. A sample display of the human observer GUI is shown in Fig. 5.9. A total of 288 rating values was collected from each observer.

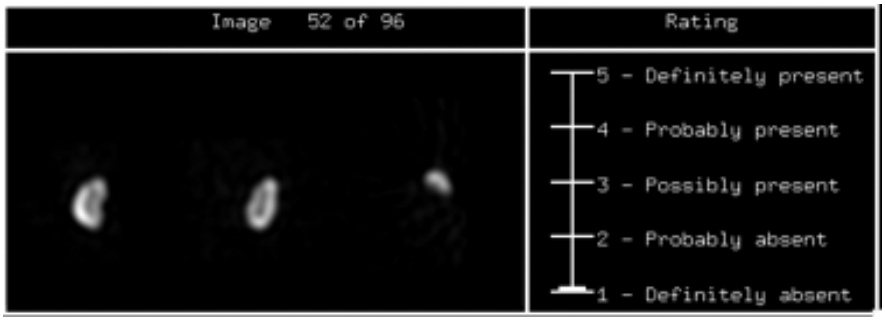

Figure 5.9. . A sample image of the GUI used in the human observer study for SLDO

5.3 Results

5.3.1 DeepAMO on simulated data

The results (Fig. 5.10) show the degree of similarity between the histograms (distributions) of the simulated test data (simulated unseen data); the degree of similarity increased as the total number of samples increased, indicating that the MDN was capable of handling complex

distributions of observer's rating values. This result agrees with the hypothesis that the MDN requires a modest amount of training data in order to learn the underlying behavior of the observer on unseen data. Here, we assumed that the underlying behavior of the observer was encoded in the distribution of that observer's rating values (training data).

The results also demonstrated that there is a tradeoff between ΔAUC and the total number of samples in the dataset. Bootstrapping was used to calculate the non-parametric confidence intervals on the ΔAUC. The ΔAUCs and 95% confidence intervals on the ΔAUCs are summarized in Table 5.6. The results show that the 100, 500, and 2,500 samples/feature vector type cases had decreasing widths of the confidence intervals of ΔAUC, indicating that, as expected, more samples are needed to demonstrate greater equivalence (smaller δ) between the human and proposed model observer. The data also suggest that training set size is an important parameter in determining the bounds of the 95% confidence interval on the ΔAUCs.

Table 5.6. Summary of simulation results

Number of samples per feature vector type	AUC of DeepAMO on simulated test data	AUC of simulated test data (ground truth)	ΔAUC	95% C.I. on ΔAUC	C.I. width
100	0.773	0.769	0.004	[-0.0502, 0.0477]	0.0979
500	0.760	0.776	-0.015	[-0.0352, 0.0261]	0.0613
2500	0.768	0.767	0.001	[-0.0074, 0.0089]	0.0163

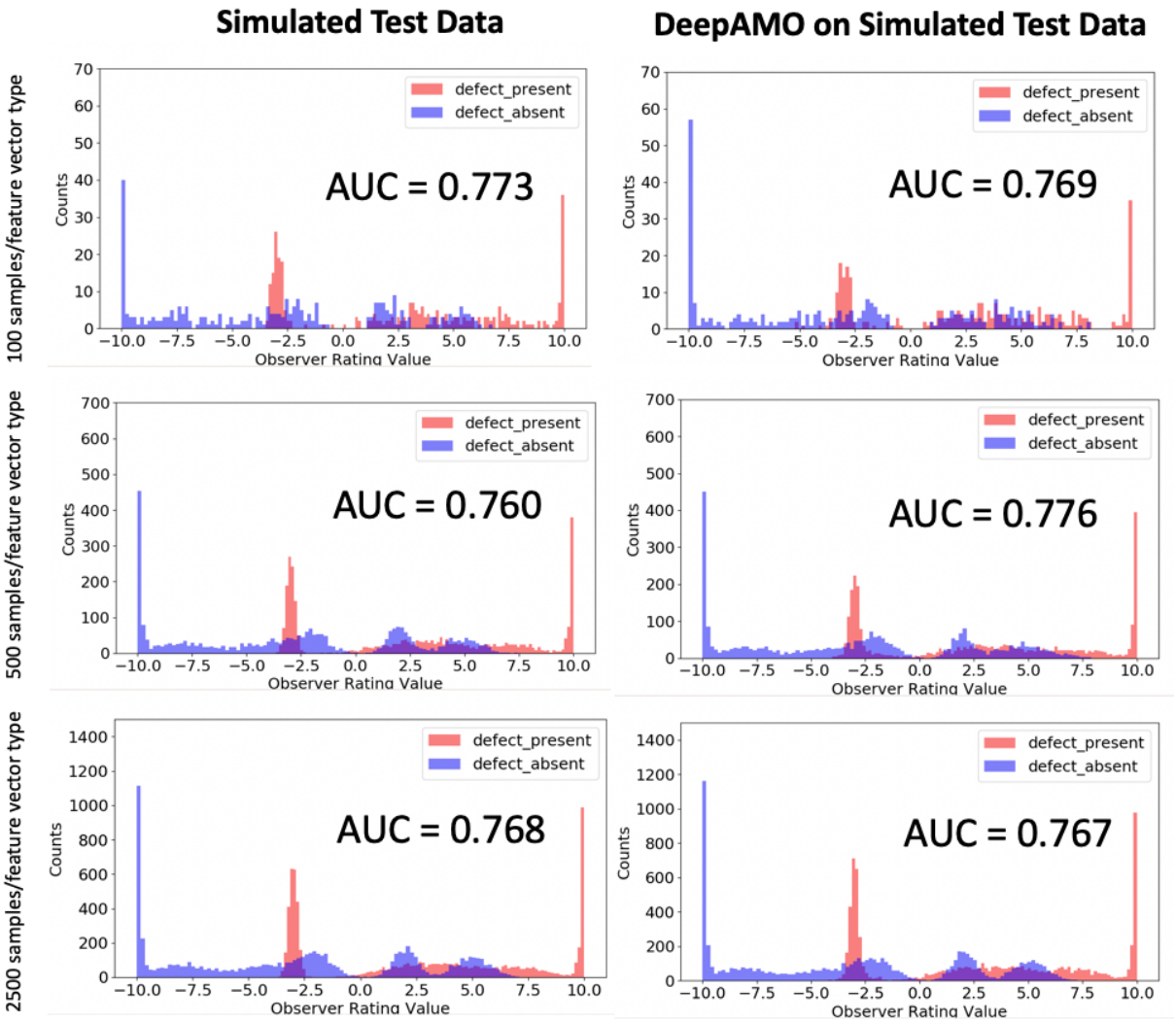

Figure 5.10. A Plots of histograms of the rating values of the simulated feature vectors (test data only) and predicted rating values on these data given by the DeepAMO. The plots show the class 0 and 1(defect present and absent, respectively) as well as the calculated AUC value.

5.3.2 DeepAMO test results

For stage I, the highest dice score achieved on the validation data for the best segmentation network was 0.975. The validation was done on a balanced dataset with 50% of the triads containing a defect.

The AUC values for the human observers and the corresponding DeepAMOs for the 5 × 2-fold cross-validation experiment are summarized in Table 5.7. The mean and standard deviation

of the ΔAUC were 0.03 and 0.0204, respectively. The 95% confidence interval for the ΔAUC was [-0.0174, 0.0426], under the assumption that ΔAUC was normally distributed. The results of the study show that the null hypothesis with a margin of difference (δ) greater than 0.0426 can be rejected at a confidence level of 95%, with this training set comprised of 288 samples. The histograms of the rating values from the human observers and the DeepAMOs for the 5 × 2-fold cross-validation experiment are shown in Fig. 5.11. The AUC value is given at the top of each plot in that figure. The distributions of the rating values for the human and model observer are qualitatively similar.

Table 5.7. Summary of stage II training results

	1st fold		2nd fold				
Trial#	AUC HO	AUC DeepAMO	AUC HO	AUC DeepAMO	ΔAUC 1st fold	ΔAUC 2nd fold	Mean ΔAUC per trial
1	0.829	0.79	0.797	0.75	0.039	0.05	0.045
2	0.814	0.77	0.816	0.78	0.044	0.036	0.04
3	0.814	0.82	0.815	0.77	-0.01	0.045	0.018
4	0.82	0.77	0.809	0.8	0.046	0.007	0.027
5	0.826	0.82	0.806	0.77	0.008	0.035	0.022

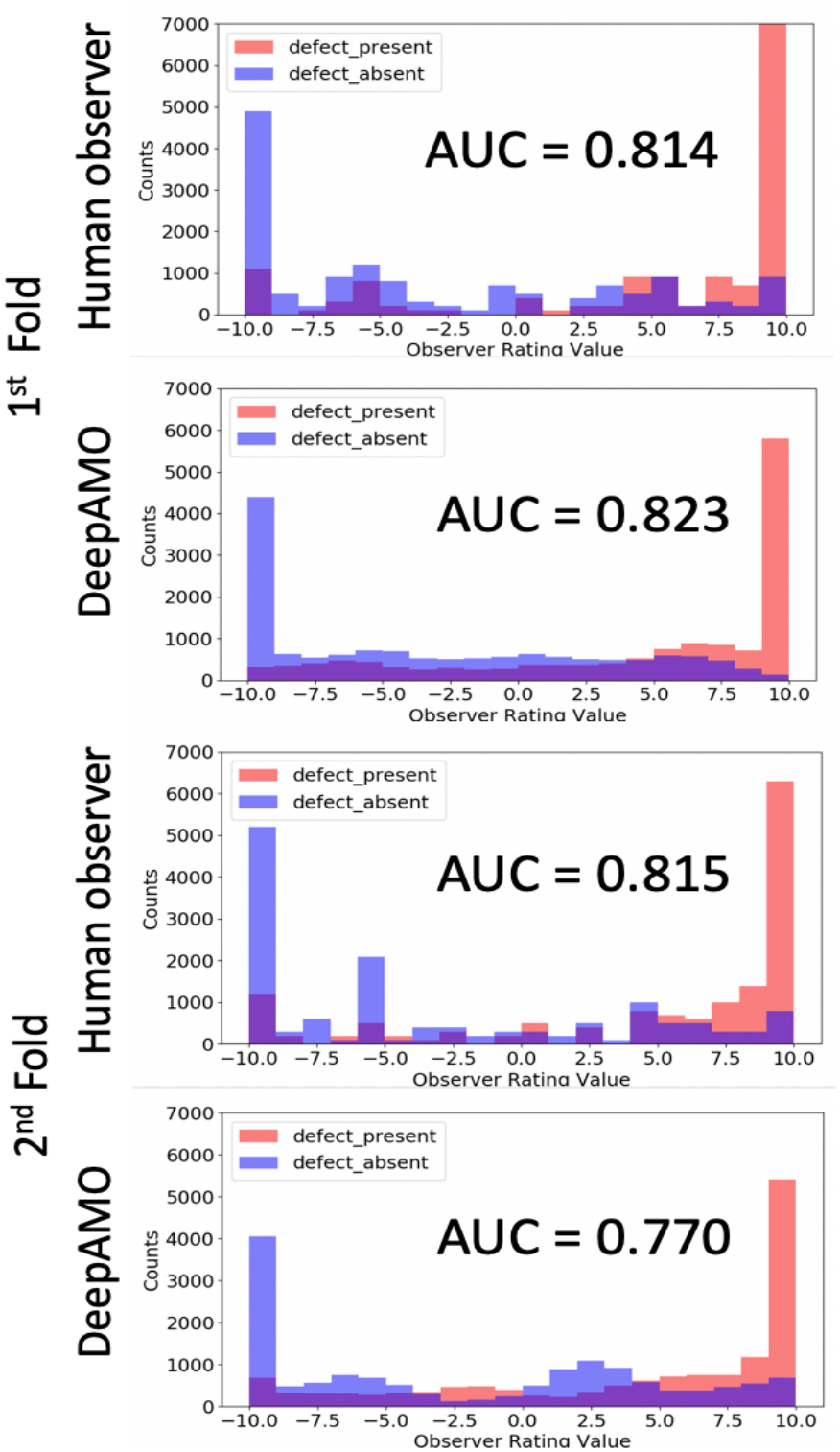

Figure 5.11. Histograms of predicted rating values given by DeepAMO on unseen human observer data from the 3rd trial of the 5 x 2-fold cross validation experiment (other trials have similar patterns). Note that multiple predicted rating values were generated for each test image during testing of the DeepAMO to reduce sampling error. The histograms of the other half of human observer data used for training the DeepAMO are not shown in the plot.

5.3.3 Scanning-linear Observer Test Results and its Human Observer Results

The mean AUC for the scanning-linear discriminant observer and its 95% confidence interval were 0.992 and [1.00, 0.986], respectively. The mean AUC for the human observer study (3-slice composite image as input) was 0.912 with a 95% confidence interval of [0.868, 0.954], which is statistically significantly different from the mean AUC (0.815, 95% C.I. = [0.851, 0.780]) from the human observer study (48-slice composite image as input). The results indicate that the SLDO overestimated human observer performance.

5.4 Discussion

One limitation of this paper is that the simulated dataset has limited background (anatomical) and signal (shape and size) variation. However, we believe that this limitation does not detract from the paper's demonstration that the proposed network architecture can model human observer performance. A dataset with greater anatomical and signal variations might require a different architecture for the segmentation network. However, as long as the segmentation network produced results that distinguish between the defect-present and absent cases at least as well as a human observer, the subsequent stages could still match that performance to human observer performance.

Another limitation of this paper is the use of non-physician observers. Non-physicians were used because of the difficulty of recruiting physician observers to perform a study of this nature. While the lack of physician observers would clearly affect the clinical diagnostic task, the task that the observers performed in this study was limited to identifying defects in images. We believe that well-trained non-physicians, with sufficient training, can perform well on this more limited task.

In addition and more importantly, the purpose of this paper was to validate the ability of the proposed model observer to reproduce human observer defect detection performance, and not to generate data on performance that impacts a clinical task. So, even if the human observers used performed poorly compared to physicians, the data demonstrate that the model can reproduce their performance. The limitations of the human visual system that degrade performance on defect detection are present even for the non-physician observers, and this work demonstrates the ability of the proposed observer to model these limitations. Therefore, we believe that the data from the observers used in this study demonstrate the utility of the proposed method.

A potential concern for the DeepAMO could be the relatively long training time (~ 2 hours) required by the segmentation network. On the contrary, the CHO or scanning forms of the CHO can provide an estimate of relative image quality, e.g., relative rankings of the methods being evaluated. However, the image quality results may not be valid for use in cases where the absolute task performance of the human observer is needed, i.e., selecting administered activity or acquisition duration in clinical practice, as they are assessed using simplified clinical tasks.

5.5 Conclusions

We have proposed a general framework for using deep convolution neural networks as an anthropomorphic model observer for the task of interpreting 3D image volumes and reproducing human observer performance. We applied this framework in the context of a renal functional defect detection task in nuclear medicine imaging using realistic simulated images. The results show that the performances of the proposed model and human observers on unseen images were equivalent with respect to a margin of difference in the AUC (ΔAUC) of 0.0426 at $p < 0.05$, for a training

set of 288 samples. The proposed framework could be readily adapted to model human observer performance on detection tasks for other imaging modalities such as PET, CT or MRI.

Chapter 6

Conclusions

6.1 Summary

Balancing dose reduction and image quality is an unmet need and important goal that has immediate clinical and societal benefits for pediatric patients. Lower radiation exposure to the patient can reduce risk and adverse effects, but can also result in reduced diagnostic image quality. Ultimately, it is desirable to use the lowest dose that gives sufficient image quality for accurate clinical diagnosis.

This dissertation proposed and developed tools for a general framework for optimizing radiation dose with task-based assessment of image quality. In this dissertation, we investigated the tradeoff between image quality and renal defect detectability as a function of administered activity, acquisition duration, and measures of body habitus for pediatric patients undergoing renal molecular imaging procedures.

In Chapter 3, we developed a projection image database modeling imaging of ^{99m}Tc-DMSA, a renal function agent. The database uses a highly-realistic population of pediatric phantoms with anatomical and body morphometry variations in height and weight. Using the developed projection image database, we have explored patient factors that affect image quality. Image quality was measured by three surrogate indices of image quality that quantify the noise (renal count density), image resolution (average radius of rotation), and scatter (scatter-to-primary ratio). The results showed that the current weight-based guidelines, based on scaling the

administered activity by patient weight, are not optimal in the sense that they do not give the same image quality for patients with the same weight. After demonstrating that height and weight did not robustly predict image quality, we explored other externally-measurable factors that could better predict image quality.

In Chapter 4, we have found that factors that are more local to the target organ may be more robust than weight for estimating the administered activity needed to provide a constant image quality across a population of patients. In the case of renal imaging, we discovered that girth at the level of the kidneys is more robust than weight in predicting administered activity needed to provide consistent image quality. In this work, analytical relationships between image quality and administered activity were derived, which could be used to determine the AA required to give a desired image quality for a given patient weight. However, one limitation of this work is that the image quality, as measured by the defect detection performance (quantified using the AUC) of an anthropomorphic model observer, was not verified by humans. To translate the image quality measures to clinical use, it is more meaningful to provide an AUC value that would be obtained for a human observer or ensemble of human observers. Due to the limitations (details are discussed in section 2.5.2) of the current model observers in modeling the clinical task involved in this work, the third part of this dissertation focused on developing a new model observer that can fully model a clinical 3D detection task.

In Chapter 5, we proposed a deep learning-based anthropomorphic model observer to fully and efficiently (in terms of both training data and computational cost) model the clinical 3D detection task using multi-slice, multi-orientation images sets. The proposed model observer is comprised of a segmentation network followed by a regression network. A human observer study using a total of 288 images was conducted, with medical imaging physics graduate students serving

as observers. A 5 × 2-fold cross validation experiment was conducted to test the statistical equivalence in defect detection performance between the proposed model observer and the human observer. The results show that the proposed model observer has the potential to mimic human observer defect detection task performance in a clinically realistic diagnostic task.

The results and tool developed in this dissertation will help provide the data needed by standards bodies to develop improved dosing guidelines for pediatric molecular imaging that result in more consistent image quality and absorbed dose.

6.1.1 A projection database of pediatric renal SPECT

The first aim of this dissertation was to build upon the Sgouros et al. work to investigate more completely the tradeoff between administered activity and image quality as a function of patient height and weight over a wide range of patient heights and weights.

As described in Chapter 3, we generated a realistic projection database modeling pediatric renal ^{99m}Tc-DMSA SPECT imaging from a digital phantom population developed by our collaborator at the University of Florida [79]. The phantom population is comprised of 90 phantoms with realistic variations in height, weight, and organ size. The phantoms model both genders at five ages (newborn, and 1-, 5-, 10-, and 15-years old). The phantoms have median (50th percentile) weight for their age and include variations having 10th, 50th, and 90th height percentiles, simulating patients having the median weight at each age with varying body habitus. The 10^{th}, 50^{th}, and 90^{th} phantoms are referred to as short and stout, average, and tall and thin, respectively. In addition, three kidney masses (-15%, average, and +15%) are modeled for each age and height percentile.

We simulated variations in radiotracer uptake in 6 tissues: cortex, medulla, pelvis, spleen, liver, and body reminder (the remaining soft tissues of the phantom). Projections of each of these tissues were generated separately assuming a uniform activity distribution. The individual projections were then scaled by the relative organ uptakes, which were based on an uptake model obtained from patients. We randomly sampled scale factors to model the variation in organ uptake seen in patient populations. For each phantom, 384 uptake realizations, modeling random variations in the uptakes of organs of interest, were generated, producing 34,560 noise-free projection datasets (384 uptake realizations times 90 phantoms). The resulting images model the projection data for that patient and uptake realization per unit administered activity at a standard acquisition duration. We fixed the acquisition duration and investigated six count levels corresponding to 25%, 50%, 75%, 100%, 125%, and 150% of the original weight-based administered activity as computed using the North American Guidelines [78]. Scaling the projections by the corresponding administered activity gave the mean projections for that count level. Noisy projection images were created by applying a Poisson-distributed pseudorandom number generator.

The results of this work showed that weight-based dosing was partially able to offset losses in count density due to variations in patient weight. However, it suggested that the kidney count density for newborns was higher than for other phantoms, suggesting that current values of minimum administered activity in dosing guidelines may result in over-dosing. The results also demonstrated variations in scatter and resolution that depend on body morphometry, and is not correlate completely with phantom height. The results suggested the need for more detailed task-based studies of image quality, and that variables beyond height and weight are needed in order to

prescribe administered activities that equalize image quality and thus achieve as little as reasonably possible dosing.

In addition to the above results, the work also provided a comprehensive method for efficiently simulating data from a population of realistic phantoms in the context for renal SPECT imaging. The set of digital phantoms, the simulation methods themselves, and the set of simulated DMSA projections provided tools and methods needed to expand the applications of realistic simulation in the optimization and evaluation of nuclear medicine and SPECT imaging.

6.1.2 An investigation of the externally-measurable factors that could better predict image quality

After demonstrating that height and weight did not robustly predict image quality, we considered other externally-measurable factors that could better predict image quality. Our general hypothesis is that patient body factors that closely describe morphometry in the region of the target organ would be most closely related to image quality. The hypothesis is based on the fact that local body morphometry would affect attenuation, system resolution, and scatter, and that morphometry away from the kidney would have little effect to image quality. For example, in the case of renal imaging, patients having large girth in the renal region would have more attenuating medium between the kidneys and the gamma camera than patients with small girth. This should result in 1) fewer photons escaping the body (higher noise), 2) larger camera radius of rotation (poorer resolution), and 3) higher scatter (poorer contrast). On the other hand, large head size would not affect renal image quality. Thus, in the second aim of this dissertation, we investigated whether

patient waist circumference (girth), kidney size and kidney depth would strongly affect image quality in DMSA SPECT.

In Chapter 4, we applied task-based image quality assessment method on the simulated projection database as described in Chapter 3. Using this realistic phantom population and projection database, we conducted two experiments in order to test the hypothesis that weight and height are not as important factors as girth to IQ. First, we used the existing projection database and treated the height variations as part of the population's anatomical variation by pooling the test statistics from different height percentiles together. Then, we calculated the detectability index (SNR^2) from the resulting AUC and fitted the following theoretical relationship relating DI to AA (full derivation of the theoretical relationship is not shown in the summary and is available in [2]).

$$SNR^2 = \frac{AA \times K_1}{AA \times K_2 + K_3}, \tag{6.1}$$

where K_1 is the mean signal difference; K_2 is the object variability noise; and K_3 is the quantum noise. Figure 6.1 shows the area under the ROC curve (AUC) vs. percent AA plot for all the patient ages and the Detectability Index (SNR^2) vs. percent AA curves and their fitted functions, respectively. Note that the Detectability Indices did not cross at the 100% count level, suggesting that the current weight-based guidelines are not optimal. That is, they do not provide the same IQ for all patients. From the plot of DI vs. AA (Fig. 6.1, right), it is evident that the curves have different shapes. Thus, scaling of the AA by a constant factor for each age could not equalize the IQ (by providing the same DI).

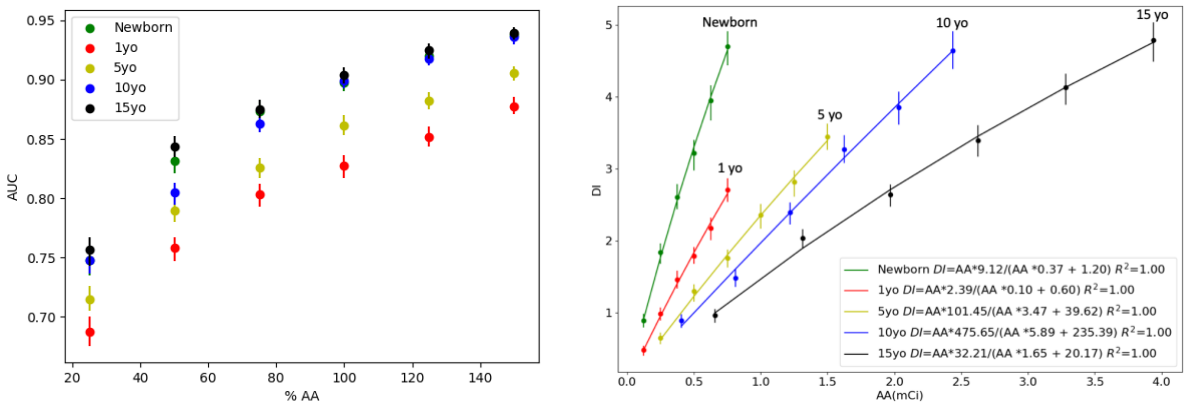

Figure 6.1. The area under the ROC curve (AUC) vs. percent AA plot for all the patient ages and DI (SNR^2) vs. AA curves and their fitted functions. The detectability index (DI) was fitted to the following theoretical relationship relating DI to the mean signal difference (K_1), object variability noise (K_2) and quantum noise (K_3), and AA.

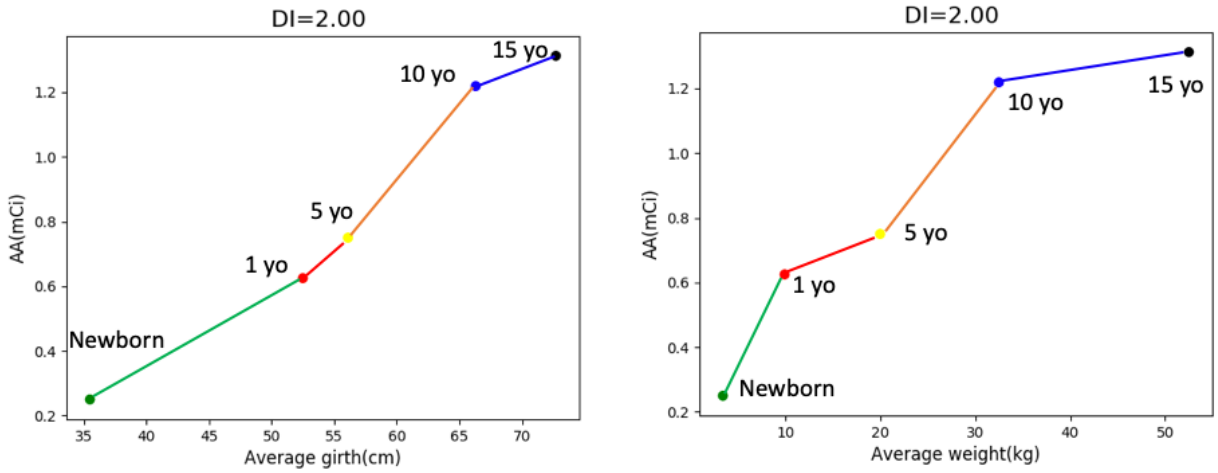

Figure 6.2. . AA vs. patient girth and weight at a fixed DI of 2.0.

Fig. 6.2 shows a comparison plot of AA vs. girth and AA vs. weight at a fixed DI of 2.0 for all the patient ages. The colored lines connect the nearest phantoms in age. These data indicate that the relationship between girth and AA is simpler and more robust than it is between weight and AA. The Pearson product-moment correlation coefficients between AA and weight and girth are 0.941 and 0.985, respectively. This again demonstrates that girth may be more robust for estimating the AA needed to provide a constant image quality.

This study demonstrated that the current consensus guidelines, which scale activities based on patient weight subject to minimum and maximum activity constraints, do not give the same image quality for patients with different weights. Further, this study provided a relationship between diagnostic image quality, as measured by AUC, and administered activity for ^{99m}Tc-DMSA pediatric SPECT for a set of phantoms having different weights. These fitted functions could potentially be used to determine the appropriate administered activity for the desired level of image quality for a given patient weight. However, more importantly, the data suggested that patient girth at the level of the kidney may ultimately be a better factor to use than weight when selecting administered activity for this imaging task.

6.1.3 DeepAMO: A multi-slice, multi-view anthropomorphic model observer for visual detection tasks performed on volume images

Due to the limitations (details are available in section 2.5.2) of the current model observers, the third aim of this dissertation focused on developing a model observer that can efficiently (both training data and training computational cost) simulate a realistic clinical realistic 3D detection task using multi-slice, multi-orientation image sets.

In Chapter 5, we developed a deep learning-based anthropomorphic model observer (DeepAMO) for image quality evaluation of multi-orientation, multi-slice image sets with respect to a clinically realistic 3D defect detection task. The input to the DeepAMO is a composite image, typical of that used to view 3D volumes in clinical practice. The output is a rating value designed to mimic human observer's defect detection performance. The main contributions of this work are threefold. First, we proposed a hypothetical model of the decision process of a reader performing a detection task using a 3D volume. Second, we proposed a projection-based defect confirmation

network architecture to confirm defect present in two different slicing orientations. Third, we proposed a novel calibration method that 'learns' the underlying distribution of observer ratings from the human observer rating data (thus modeling inter- or intra- observer variability) using a Mixture Density Network. We implemented and evaluated the DeepAMO in the context of ^{99m}Tc-DMSA SPECT imaging. A human observer study was conducted, with two medical imaging physics graduate students serving as observers. A 5×2-fold cross-validation experiment was conducted to test the statistical equivalence in defect detection performance between the DeepAMO and the human observer. The results show that the DeepAMO's and human observer's performances on unseen images were statistically equivalent with a margin of difference (ΔAUC) of 0.0426 at $p < 0.05$, using 288 training images. The results show that the DeepAMO has the potential to mimic human observer defect detection task performance in a clinically realistic diagnostic task.

6.2 Contributions

Through the course of this work, we have made several major contributions to the development of an improved dosing guidelines for pediatric molecular imaging that result in more consistent image quality and absorbed dose.

First, we developed a realistic projection database for investigation of relationship between image quality and patient morphometry in ^{99m}Tc-DMSA renal SPECT. The database generated in this work is immediately applicable to other pharmaceuticals labeled with ^{99m}Tc used in pediatric imaging such as ^{99m}Tc- MAG3 or ^{99m}Tc-MDP; only scaling and summing of the organ projections with appropriate scaling factors reflecting agent biokinetics. Further, the methods used in this study

are applicable to studying these tradeoffs for other diagnostic and/or therapeutic radiopharmaceuticals in both pediatric and adult patients.

Second, we demonstrated that the current consensus guidelines, which scale activities based on patient weight subject to minimum and maximum activity constraints, do not give the same IQ for patients with different weights. Furthermore, this study provides a relationship between diagnostic IQ, as measured by AUC, and AA for ^{99m}Tc- DMSA pediatric SPECT for a set of phantoms having different weights. These fitted functions could potentially be used to determine the appropriate AA for desired level of IQ for a given patient weight. However, the data suggest that patient girth at the level of the kidney may be a better factor to use than weight when selecting AA for this imaging task.

Third, we proposed a general framework for using deep convolution neural networks as an anthropomorphic model observer for the task of interpreting 3D image volumes and reproducing human observer performance, and good results were obtained. The results showed that the DeepAMO has the potential to reproduce the performance of human observers on a clinically-realistic defect detection task; absolute performance was not reproduced by a scanning model observer based on the optimal linear discriminant. The proposed framework could be readily adapted to model human observer performance on detection tasks for other imaging modalities such as PET, CT or MRI

While this work provided several important steps towards the development of an improved dosing guidelines for pediatric molecular imaging, there is still work that remains to be done for establishing the data needed by standards bodies to develop improved dosing guidelines for pediatric molecular imaging that result in more consistent image quality and absorbed dose.

6.3 Future works

The findings in this dissertation suggest two areas of future work.

First, the findings in Chapter 4 suggest a new direction to investigate the IQ-RD tradeoff relationships as functions of patient girth.

Second, the major work in this dissertation was done using model observer. Image quality data only showed rankings as functions of AA but not absolute performance representing human performance. To translate the image quality measures to clinical use, it is more meaningful to provide an AUC value that would be obtained for a human observer. Thus, a human observer study is desired to calibrate the model observer to be used to calculate the absolute IQ-RD tradeoff relationships.

6.4 Conclusions

This dissertation has provided useful direction and tool for a general framework for optimizing radiation dose with task-based assessment of image quality. First, we demonstrated that the weight-based dose scaling does not equalize image quality, as measured by defect detectability, for patients with different weights. Second, we have found that patient body factors that are more local to the target organ may be more robust than weight for estimating the administered activity needed to provide a constant image quality across a population of patients. In the case of renal imaging, we have discovered that girth is more robust than weight in predicting administered activity needed to provide a desired image quality. Third, we have proposed a novel

deep learning-based anthropomorphic model observer that can efficiently simulate a realistic clinical realistic 3D detection task using multi-slice, multi-orientation image sets.

The results of this dissertation provide a general framework, a new investigative direction (patient body factors local to the target organ), as well as tools (database, DeepAMO) for optimizing radiation dose with task-based assessment of image quality for nuclear medicine imaging. These results and methods from this dissertation will help provide the data needed by standards bodies to develop improved dosing guidelines for pediatric molecular imaging that result in more consistent image quality and low absorbed dose.

Bibliography

1. Li, Y., et al., *A projection image database to investigate factors affecting image quality in weight-based dosing: application to pediatric renal SPECT.* Phys Med Biol, 2018. **63**(14): p. 145004.
2. Li, Y., et al., *Current pediatric administered activity guidelines for (99m) Tc-DMSA SPECT based on patient weight do not provide the same task-based image quality.* Med Phys, 2019. **46**(11): p. 4847-4856.
3. Treves, S.T., *Pediatric nuclear medicine and molecular imaging*. Fourth edition. ed. 2014, New York: Springer. xxiv, 712 pages.
4. Fahey, F.H., et al., *Dose Estimation in Pediatric Nuclear Medicine.* Semin Nucl Med, 2017. **47**(2): p. 118-125.
5. Sgouros, G., et al., *An approach for balancing diagnostic image quality with cancer risk: application to pediatric diagnostic imaging of 99mTc-dimercaptosuccinic acid.* J Nucl Med, 2011. **52**(12): p. 1923-9.
6. Bolch, W.E., et al., *MIRD pamphlet No. 21: a generalized schema for radiopharmaceutical dosimetry--standardization of nomenclature.* J Nucl Med, 2009. **50**(3): p. 477-84.
7. Fahey, F.H., et al. *Dose Estimation in Pediatric Nuclear Medicine*. in *Seminars in Nuclear Medicine*. 2016. Elsevier.
8. Yanagimachi, R., *The Sperm Cell Production, Maturation, Fertilization, Regeneration Foreword.* Sperm Cell: Production, Maturation, Fertilization, Regeneration, 2nd Edition, 2017: p. X-Xi.
9. Alzen, G. and G. Benz-Bohm, *Radiation protection in pediatric radiology.* Dtsch Arztebl Int, 2011. **108**(24): p. 407-14.
10. Li, L., M. Story, and R.J. Legerski, *Cellular responses to ionizing radiation damage.* International Journal of Radiation Oncology Biology Physics, 2001. **49**(4): p. 1157-1162.
11. Pray, L.A., *DNA Replication and Causes of Mutation.* Nature Education 2008. **1**(1): p. 214.
12. Lassmann, M., et al., *The new EANM paediatric dosage card.* Eur J Nucl Med Mol Imaging, 2007. **34**(5): p. 796-8.
13. Treves, S.T., et al., *2016 Update of the North American Consensus Guidelines for Pediatric Administered Radiopharmaceutical Activities.* J Nucl Med, 2016. **57**(12): p. 15N-18N.
14. Nairne, J., P.B. Iveson, and A. Meijer, *Imaging in drug development.* Prog Med Chem, 2015. **54**: p. 231-80.
15. Lin, T.H., A. Khentigan, and H.S. Winchell, *A 99mTc-chelate substitute for organoradiomercurial renal agents.* J Nucl Med, 1974. **15**(1): p. 34-5.
16. Maisey, M.N., K.E. Britton, and B.D. Collier, *Clinical nuclear medicine*. 1998, Chapman & Hall,: London. p. 1 online resource (xii, 4 , 752 p.).
17. Arnold, R.W., et al., *Comparison of 99mTc complexes for renal imaging.* J Nucl Med, 1975. **16**(5): p. 357-67.
18. Handmaker, H., B.W. Young, and J.M. Lowenstein, *Clinical experience with 99mTc-DMSA (dimercaptosuccinic acid), a new renal-imaging agent.* J Nucl Med, 1975. **16**(1): p. 28-32.

19. Enlander, D., P.M. Weber, and L.V. dos Remedios, *Renal cortical imaging in 35 patients: superior quality with 99mTc-DMSA.* J Nucl Med, 1974. **15**(9): p. 743-9.
20. Treves, S.T., *Pediatric Nuclear Medicine/PET, 3rd edition*. 2007, New York NY: Springer Science+Business Media LLC.
21. Willis, K.W., et al., *Renal localization of 99mTc-stannous glucophetonate and 99mTc-stannous dimercaptosuccinate in the rat by frozen section autoradiography. The efficiency and resolution of technetium-99m.* Radiat Res, 1977. **69**(3): p. 475-88.
22. ICRP, *Publication 53: Radiation Dose To Patients from Radiopharmaceuticals, 53.* Ann ICRP, 1988. **18**(1-4).
23. Dart, R.C., et al., *Pharmacokinetics of Meso-2,3-Dimercaptosuccinic Acid in Patients with Lead-Poisoning and in Healthy-Adults.* Journal of Pediatrics, 1994. **125**(2): p. 309-316.
24. Cherry, S.R., J.A. Sorenson, and M.E. Phelps, *Physics in nuclear medicine*. 2012, Elsevier/Saunders,: Philadelphia. p. 1 online resource (xvii, 523 p.).
25. Sprawls, P., *Physical principles of medical imaging*. 2nd ed. 1993, Gaithersburg, Md.: Aspen Publishers. xv, 656 p.
26. Barrett, H.H., et al., *Objective Assessment of Image Quality .2. Fisher Information, Fourier Crosstalk, and Figures of Merit for Task-Performance.* Journal of the Optical Society of America a-Optics Image Science and Vision, 1995. **12**(5): p. 834-852.
27. Barrett, H.H., C.K. Abbey, and E. Clarkson, *Objective assessment of image quality. III. ROC metrics, ideal observers, and likelihood-generating functions.* Journal of the Optical Society of America a-Optics Image Science and Vision, 1998. **15**(6): p. 1520-1535.
28. Barrett, H.H., *Objective Assessment of Image Quality - Effects of Quantum Noise and Object Variability.* Journal of the Optical Society of America a-Optics Image Science and Vision, 1990. **7**(7): p. 1266-1278.
29. Barrett, H.H., et al., *Objective assessment of image quality. IV. Application to adaptive optics.* Journal of the Optical Society of America a-Optics Image Science and Vision, 2006. **23**(12): p. 3080-3105.
30. Barrett, H.H., et al., *Objective assessment of image quality VI: imaging in radiation therapy.* Phys Med Biol, 2013. **58**(22): p. 8197-213.
31. Barrett, H.H. and K.J. Myers, *Foundations of image science*. Wiley series in pure and applied optics. 2004, Hoboken, NJ: Wiley-Interscience. xli, 1540 p.
32. Barrett, H.H., et al., *Task-based measures of image quality and their relation to radiation dose and patient risk.* Physics in Medicine and Biology, 2015. **60**(2): p. R1-R75.
33. He, X. and S. Park, *Model observers in medical imaging research.* Theranostics, 2013. **3**(10): p. 774-86.
34. Myers, K.J. and H.H. Barrett, *Addition of a Channel Mechanism to the Ideal-Observer Model.* Journal of the Optical Society of America a-Optics Image Science and Vision, 1987. **4**(12): p. 2447-2457.
35. Sachs, M.B., J. Nachmias, and J.G. Robson, *Spatial-frequency channels in human vision.* J Opt Soc Am, 1971. **61**(9): p. 1176-86.
36. Park, S., et al., *Channelized-ideal observer using Laguerre-Gauss channels in detection tasks involving non-Gaussian distributed lumpy backgrounds and a Gaussian signal.* J Opt Soc Am A Opt Image Sci Vis, 2007. **24**(12): p. B136-50.
37. Burgess, A.E., *Visual Perception Studies and Observer Models in Medical Imaging.* Seminars in Nuclear Medicine, 2011. **41**(6): p. 419-436.

38. Sankaran, S., et al., *Optimum compensation method and filter cutoff frequency in myocardial SPECT: a human observer study.* J Nucl Med, 2002. **43**(3): p. 432-8.
39. Zhang, Y., et al., *Correlation between human and model observer performance for discrimination task in CT.* Physics in Medicine and Biology, 2014. **59**(13): p. 3389-3404.
40. Barrett, H.H., et al., *Model observers for assessment of image quality.* Proc Natl Acad Sci U S A, 1993. **90**(21): p. 9758-65.
41. Frey, E.C., K.L. Gilland, and B.M. Tsui, *Application of task-based measures of image quality to optimization and evaluation of three-dimensional reconstruction-based compensation methods in myocardial perfusion SPECT.* IEEE Trans Med Imaging, 2002. **21**(9): p. 1040-50.
42. Burgess, A.E., X. Li, and C.K. Abbey, *Visual signal detectability with two noise components: anomalous masking effects.* J Opt Soc Am A Opt Image Sci Vis, 1997. **14**(9): p. 2420-42.
43. Fukunaga, K., *Introduction to statistical pattern recognition*. 2nd ed. Computer science and scientific computing. 1990, Boston: Academic Press. xiii, 591 p.
44. Gifford, H.C., et al., *Channelized hotelling and human observer correlation for lesion detection in hepatic SPECT imaging.* J Nucl Med, 2000. **41**(3): p. 514-21.
45. Myers, K.J. and H.H. Barrett, *Addition of a channel mechanism to the ideal-observer model.* J Opt Soc Am A, 1987. **4**(12): p. 2447-57.
46. Yao, J. and H.H. Barrett, *Predicting Human-Performance by a Channelized Hotelling Observer Model.* Mathematical Methods in Medical Imaging, 1992. **1768**: p. 161-168.
47. He, X., J.M. Links, and E.C. Frey, *An investigation of the trade-off between the count level and image quality in myocardial perfusion SPECT using simulated images: the effects of statistical noise and object variability on defect detectability.* Physics in Medicine and Biology, 2010. **55**(17): p. 4949-4961.
48. Eckstein, M.P., C.K. Abbey, and J.S. Whiting, *Human vs. model observers in anatomic backgrounds.* Image Perception, 1998. **3340**: p. 16-26.
49. Wollenweber, S.D., et al., *Comparison of hotelling observer models and human observers in defect detection from myocardial SPECT imaging.* Ieee Transactions on Nuclear Science, 1999. **46**(6): p. 2098-2103.
50. Park, S., et al., *Efficiency of the human observer detecting random signals in random backgrounds.* Journal of the Optical Society of America a-Optics Image Science and Vision, 2005. **22**(1): p. 3-16.
51. Sankaran, S., et al., *Optimum compensation method and filter cutoff frequency in myocardial SPECT: A human observer study.* Journal of Nuclear Medicine, 2002. **43**(3): p. 432-438.
52. Sen, A., F. Kalantari, and H.C. Gifford, *Task Equivalence for Model and Human-Observer Comparisons in SPECT Localization Studies.* Ieee Transactions on Nuclear Science, 2016. **63**(3): p. 1426-1434.
53. Gifford, H.C., *Efficient visual-search model observers for PET.* Br J Radiol, 2014. **87**(1039): p. 20140017.
54. Gifford, H.C., A. Lehovich, and M.A. King, *A Multiclass Model Observer for Multislice-Multiview Images.* IEEE Nucl Sci Symp Conf Rec (1997), 2006. **3**: p. 1687-1691.
55. Gifford, H.C., *A visual-search model observer for multislice-multiview SPECT images.* Med Phys, 2013. **40**(9): p. 092505.

56. Zhang, Y., B.T. Pham, and M.P. Eckstein, *Evaluation of internal noise methods for Hotelling observer models.* Medical Physics, 2007. **34**(8): p. 3312-3322.
57. Brankov, J.G., *Evaluation of the channelized Hotelling observer with an internal-noise model in a train-test paradigm for cardiac SPECT defect detection.* Phys Med Biol, 2013. **58**(20): p. 7159-82.
58. Brankov, J.G. *Optimization of the internal noise models for channelized Hotelling observer*. in *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. 2011. Chicago, IL, USA: IEEE.
59. Brankov, J.G. *Comparison of the internal noise models for channelized Hotelling observer*. in *2011 IEEE Nuclear Science Symposium Conference Record*. 2011. Valencia, Spain.
60. Krizhevsky, A., I. Sutskever, and G.E. Hinton, *ImageNet Classification with Deep Convolutional Neural Networks.* Communications of the Acm, 2017. **60**(6): p. 84-90.
61. Lu, L., et al., *Deep learning and convolutional neural networks for medical image computing : precision medicine, high performance and large-scale datasets*, in *Advances in computer vision and pattern recognition,*. 2017, Springer,: Cham. p. 1 online resource.
62. Shin, H.C., et al., *Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning.* Ieee Transactions on Medical Imaging, 2016. **35**(5): p. 1285-1298.
63. Zhou, S.K., H. Greenspan, and D. Shen, *Deep Learning for Medical Image Analysis*. Elsevier and MICCAI Society book series. 2017, London ; San Diego: Elsevier/Academic Press. xxiii, 433 p.
64. Murphy, K.P., *Machine Learning: A Probabilistic Perspective.* Machine Learning: A Probabilistic Perspective, 2012: p. 1-1067.
65. Rosenblatt, F., *The perceptron: a probabilistic model for information storage and organization in the brain.* Psychol Rev, 1958. **65**(6): p. 386-408.
66. White, H., *Artificial neural networks : approximation and learning theory*. 1992, Oxford, UK ; Cambridge, USA: Blackwell. x, 329 p.
67. Y. Lecun, L.B., Y. Bengio, P. Haffner, *Gradient-based learning applied to document recognition.* Proceedings of the IEEE, 1998. **86**(11).
68. Bishop, C.M., *Pattern recognition and machine learning*. Information science and statistics. 2006, New York: Springer. xx, 738 p.
69. Alnowami, M., et al., *A Deep Learning Model Observer for use in Alterative Forced Choice Virtual Clinical Trials.* Medical Imaging 2018: Image Perception, Observer Performance, and Technology Assessment, 2018. **10577**.
70. Massanes, F. and J.G. Brankov, *Evaluation of CNN as anthropomorphic model observer.* Medical Imaging 2017: Image Perception, Observer Performance, and Technology Assessment, 2017. **10136**.
71. Kopp, F.K., et al., *CNN as model observer in a liver lesion detection task for x-ray computed tomography: A phantom study.* Medical Physics, 2018. **45**(10): p. 4439-4447.
72. Zhou, W.M., H. Li, and M.A. Anastasio, *Approximating the Ideal Observer and Hotelling Observer for Binary Signal Detection Tasks by Use of Supervised Learning Methods.* Ieee Transactions on Medical Imaging, 2019. **38**(10): p. 2456-2468.
73. Gong, H., et al., *Deep-learning-based model observer for a lung nodule detection task in computed tomography.* J Med Imaging (Bellingham), 2020. **7**(4): p. 042807.

74. He, K., et al., *Deep Residual Learning for Image Recognition.* CoRR, 2015. **abs/1512.03385 %6**: p. %&.
75. Treves, S.T., R.T. Davis, and F.H. Fahey, *Administered radiopharmaceutical doses in children: a survey of 13 pediatric hospitals in North America.* J Nucl Med, 2008. **49**(6): p. 1024-7.
76. Jacobs, F., et al., *Optimised tracer-dependent dosage cards to obtain weight-independent effective doses.* Eur J Nucl Med Mol Imaging, 2005. **32**(5): p. 581-8.
77. Cristy, M. and K.F. Eckerman, *Specific Absorbed Fractions of Energy at Various Ages for Internal Photon Sources*. 1987, Oak Ridge National Laboratory.
78. Gelfand, M.J., et al., *Pediatric radiopharmaceutical administered doses: 2010 North American consensus guidelines.* J Nucl Med, 2011. **52**(2): p. 318-22.
79. O'Reilly, S.E., et al., *A risk index for pediatric patients undergoing diagnostic imaging with (99m)Tc-dimercaptosuccinic acid that accounts for body habitus.* Phys Med Biol, 2016. **61**(6): p. 2319-32.
80. Evans, K., et al., *Biokinetic behavior of technetium-99m-DMSA in children.* J Nucl Med, 1996. **37**(8): p. 1331-5.
81. Frey, E.C., Z.W. Ju, and B.M.W. Tsui, *A Fast Projector-Backprojector Pair Modeling the Asymmetric, Spatially Varying Scatter Response Function for Scatter Compensation in SPECT Imaging.* IEEE Transactions on Nuclear Science, 1993. **40**(4): p. 1192-1197.
82. Frey, E.C. and B.M.W. Tsui, *A new method for modeling the spatially-variant, object-dependent scatter response function in SPECT.* 1996 Ieee Nuclear Science Symposium - Conference Record, Vols 1-3, 1997: p. 1082-1086.
83. ICRP, *Basic anatomical and physiological data for use in radiological protection: reference values. A report of age- and gender-related differences in the anatomical and physiological characteristics of reference individuals. ICRP Publication 89.* Ann ICRP, 2002. **32**(3-4): p. 5-265.
84. Li, Y., et al. *Development of a defect model for renal pediatric SPECT imaging research*. in *Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), 2015 IEEE*. 2015. IEEE.
85. Elshahaby, F.E., et al., *Factors affecting the normality of channel outputs of channelized model observers: an investigation using realistic myocardial perfusion SPECT images.* J Med Imaging (Bellingham), 2016. **3**(1): p. 015503.
86. Li, X., et al., *Use of Sub-Ensembles and Multi-Template Observers to Evaluate Detection Task Performance for Data That are Not Multivariate Normal.* IEEE Trans Med Imaging, 2017. **36**(4): p. 917-929.
87. Elshahaby, F.E.A., et al., *A comparison of resampling schemes for estimating model observer performance with small ensembles.* Phys Med Biol, 2017. **62**(18): p. 7300-7320.
88. He, X., J.M. Links, and E.C. Frey, *An investigation of the trade-off between the count level and image quality in myocardial perfusion SPECT using simulated images: the effects of statistical noise and object variability on defect detectability.* Phys Med Biol, 2010. **55**(17): p. 4949-61.
89. Plyku, D., et al., *Pharmacokinetic modeling of pediatric imaging agents.* Journal of Nuclear Medicine, 2014. **55**(supplement 1): p. 1134-1134.
90. Du, Y., et al., *Combination of MCNP and SimSET for Monte Carlo simulation of SPECT with medium- and high-energy photons.* IEEE Transactions on Nuclear Science, 2002. **49**(3): p. 668-674.

91. Du, Y., B.M.W. Tsui, and E.C. Frey, *Model-based compensation for quantitative I-123 brain SPECT imaging.* Physics in Medicine and Biology, 2006. **51**(5): p. 1269-1282.
92. Du, Y., B.M.W. Tsui, and E.C. Frey, *Model-based crosstalk compensation for simultaneous Tc-99m/I-123 dual-isotope brain SPECT imaging.* Medical Physics, 2007. **34**(9): p. 3530-3543.
93. He, B., et al., *A Monte Carlo and physical phantom evaluation of quantitative In-111SPECT.* Physics in Medicine and Biology, 2005. **50**(17): p. 4169-4185.
94. Mok, G.S.P., et al., *Development and Validation of a Monte Carlo Simulation Tool for Multi-Pinhole SPECT.* Molecular Imaging and Biology, 2010. **12**(3): p. 295-304.
95. Rong, X., et al., *Development and evaluation of an improved quantitative (90)Y bremsstrahlung SPECT method.* Med Phys, 2012. **39**(5): p. 2346-58.
96. Song, N., et al., *Development and evaluation of a model-based downscatter compensation method for quantitative I-131 SPECT.* Med Phys, 2011. **38**(6): p. 3193-204.
97. Song, N., et al., *EQPlanar: a maximum-likelihood method for accurate organ activity estimation from whole body planar projections.* Phys Med Biol, 2011. **56**(17): p. 5503-24.
98. Wang, W.T., et al., *Parameterization of Pb X-ray contamination in simultaneous Tl-201 and Tc-99m dual-isotope imaging.* IEEE Transactions on Nuclear Science, 2002. **49**(3): p. 680-692.
99. Elshahaby, F.E.A., et al., *Factors affecting the normality of channel outputs of channelized model observers: an investigation using realistic myocardial perfusion SPECT images.* Journal of Medical Imaging, 2016. **3**(1).
100. Metz, C.E., B.A. Herman, and J.H. Shen, *Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data.* Stat Med, 1998. **17**(9): p. 1033-53.
101. Harris, J.L., *Resolving Power + Decision Theory.* Journal of the Optical Society of America, 1964. **54**(5): p. 606-&.
102. Hanson, K.M. and K.J. Myers, *Rayleigh Task-Performance as a Method to Evaluate Image-Reconstruction Algorithms.* Maximum Entropy and Bayesian Methods //, 1991. **43**: p. 303-312.
103. Wagner, R.F., K.J. Myers, and K.M. Hanson, *Task-Performance on Constrained Reconstructions - Human Observer Performance Compared with Suboptimal Bayesian Performance.* Medical Imaging Vi : Image Processing, 1992. **1652**: p. 352-362.
104. Judy, P.F., R.G. Swensson, and M. Szulc, *Lesion Detection and Signal-to-Noise Ratio in Ct Images.* Medical Physics, 1981. **8**(1): p. 13-23.
105. Myers, K.J., et al., *Effect of Noise Correlation on Detectability of Disk Signals in Medical Imaging.* Journal of the Optical Society of America a-Optics Image Science and Vision, 1985. **2**(10): p. 1752-1759.
106. Sen, A., F. Kalantari, and H.C. Gifford, *Task Equivalence for Model and Human-Observer Comparisons in SPECT Localization Studies.* IEEE Trans Nucl Sci, 2016. **63**(3): p. 1426-1434.
107. Zhang, L., et al., *A Multi-Slice Model Observer for Medical Image Quality Assessment.* 2015 Ieee International Conference on Acoustics, Speech, and Signal Processing (Icassp), 2015: p. 1667-1671.

108. Han, M. and J. Baek, *A performance comparison of anthropomorphic model observers for breast cone beam CT images: A single-slice and multislice study.* Medical Physics, 2019. **46**(8): p. 3431-3441.
109. Kim, J.S., et al., *A comparison of planar versus volumetric numerical observers for detection task performance in whole-body PET imaging.* Ieee Transactions on Nuclear Science, 2004. **51**(1): p. 34-40.
110. Liang, H.Y., et al., *Image browsing in slow medical liquid crystal displays.* Academic Radiology, 2008. **15**(3): p. 370-382.
111. Lartizien, C., P.E. Kinahan, and C. Comtat, *Volumetric model and human observer comparisons of tumor detection for whole-body positron emission tomography.* Academic Radiology, 2004. **11**(6): p. 637-648.
112. Chen, M., et al., *Using the Hotelling observer on multislice and multiview simulated SPECT myocardial images.* Ieee Transactions on Nuclear Science, 2002. **49**(3): p. 661-667.
113. Gifford, H.C., et al., *A comparison of human and model observers in multislice LROC studies.* Ieee Transactions on Medical Imaging, 2005. **24**(2): p. 160-169.
114. Treves, S.T., et al., *Standardization of pediatric nuclear medicine administered radiopharmaceutical activities: the SNMMI/EANM Joint Working Group.* Clinical and Translational Imaging, 2016. **4**(3): p. 203-209.
115. Metz, C.E., *Basic principles of ROC analysis.* Semin Nucl Med, 1978. **8**(4): p. 283-98.
116. Brown JL, S.-S.B., Li Y, Frey EC, Treves ST, Fahey FH, Plyku D, Sgouros G, and Bolch WE, *A pediatric library of phantoms for renal imaging incorporating waist circumference, renal volume, and renal depth*, in *Annual Meeting of the European Association of Nuclear Medicine*. 2018: Düsseldorf, Germany.
117. Li, Y., et al., *; Development of a Defect Model for Renal Pediatric SPECT Imaging Research.* 2015 Ieee Nuclear Science Symposium and Medical Imaging Conference (Nss/Mic), 2015.
118. Bishop, C., *Mixture density networks*. 1994, Aston University: Neural Computing Research Group.
119. Wong, K.C.L., et al., *3D Segmentation with Exponential Logarithmic Loss for Highly Unbalanced Object Sizes.* Medical Image Computing and Computer Assisted Intervention, Pt Iii, 2018. **11072**: p. 612-619.
120. Ronneberger, O., P. Fischer, and T. Brox, *U-Net: Convolutional Networks for Biomedical Image Segmentation.* Medical Image Computing and Computer-Assisted Intervention, Pt Iii, 2015. **9351**: p. 234-241.
121. Ba, D.P.K.a.J., *Adam: A Method for Stochastic Optimization.* ArXiv e-prints, 2014. **1412.6980**.
122. Chen, W.J., N.A. Petrick, and B. Sahiner, *Hypothesis Testing in Noninferiority and Equivalence MRMC ROC Studies.* Academic Radiology, 2012. **19**(9): p. 1158-1165.

Vita

Ye Li (b. 1989 in Xi'an, Shaanxi, China) earned his B.S. degree in Radiological Engineering with a minor in Physics from the University of Illinois at Urbana-Champaign in 2013. In 2014, he began his Ph.D. studies in the Department of Electrical and Computer Engineering at the Johns Hopkins University. He is also a graduate research assistant in the Division of Medical Imaging Physics, Department of Radiology and Radiological Science at the Johns Hopkins University School of Medicine. His Ph.D. research focuses on task-based optimization of imaging systems. In particular, he is interested in studying anthropomorphic model observers that can be used as surrogate for radiologist(s) to perform clinical diagnostic tasks. His general research interests include AI for medical imaging, machine learning, and statistics. He also gained industrial experience from his research internship with IBM Research in the summer of 2018.