Characterizing and Minimizing the Impacts of Diagnostic Computed Tomography Acquisition and Reconstruction Parameter Selection on Quantitative Emphysema Scoring

Hoffman, John

UNIVERSITY OF CALIFORNIA

Los Angeles

Characterizing and Minimizing the Impacts of Diagnostic Computed Tomography Acquisition and Reconstruction Parameter Selection on Quantitative Emphysema Scoring

A dissertation submitted in partial satisfaction of the

requirements for the degree of Doctor of Philosophy

in Biomedical Physics

by

John Marian Hoffman

2018

ABSTRACT OF THE DISSERTATION

Characterizing and Minimizing the Impacts of Diagnostic Computed Tomography Acquisition and Reconstruction Parameter Selection on Quantitative Emphysema Scoring

by

John Marian Hoffman
Doctor of Philosophy in Biomedical Physics
University of California, Los Angeles, 2018
Professor Michael F. McNitt-Gray, Chair

Computed tomography (CT) has proven to be a critical component of clinical care, and at present there is a strong interest in quantitative imaging: augmenting or assisting human readers through the use of quantitative and computational techniques applied to the image data. While quantitative imaging is extremely promising and has been the focus of many research projects, widespread clinical adoption has not yet occurred, in part due to the susceptibility of many tests to CT acquisition and reconstruction parameters (such as radiation dose, reconstruction kernel, reconstruction algorithm, etc.). Previous efforts to illustrate and quantify the effects of parameter selection on various quantitative imaging tests have been limited in their ability to inform broader use of quantitative imaging; this is partly due to the number of parameters investigated (typically one) and/or the cohort size. This work builds on previous efforts by studying one well-established quantitative imaging test, namely emphysema scoring, using a newly-developed, high-throughput, quantitative imaging pipeline. Because of the number of conditions investigated and the cohort side, a key goal of this work is to provide recommendations for the clinical use of quantitative emphysema scoring.

The high-throughput pipeline was utilized to reconstruct a cohort of 142 subjects, scanned using the lung-screening protocol at our institution. Each scan was reconstructed under a variety of conditions: 100%, 50%, 25%, and 10% dose levels; 0.6mm, 1.0mm, and 2.0mm slice thickness; and smooth, medium, and sharp reconstruction kernels. Additionally, two reconstruction approaches were investigated: weighted filtered backprojection (wFBP), and an implementation of iterative reconstruction (Siemens SAFIRE). Thus, each scan was reconstructed using 72 unique parameter configurations.

First, the susceptibility of quantitative emphysema scoring was investigated and characterized by determining “safe” parameter configurations (i.e. resulting in small emphysema score change from a reference value computed on the 1.0mm, smooth kernel, 100% dose, wFBP reconstruction). Second, an adaptive denoising method (bilateral filtering, adjusted based on slice thickness and dose) was applied and safe parameter configurations were reassessed.

It was found that there exist small groupings of parameter combinations near the reference value that produce quantitative emphysema scores similar to the reference. This suggests that careful protocol adherence is not strictly necessary to obtain a reasonably accurate quantitative emphysema score, however there were still many parameter configurations that resulted in large deviations from the reference score. In terms of clinical translation, this suggests that in addition to a standardized, recommended protocol, one or two small changes would not typically compromise the results. However, if a scan or reconstruction was acquired using a parameter configuration deemed “unsafe,” this approach provides no means to obtain a valid emphysema score, other than to reacquire or re-reconstruct the data, which is not typically available.

With adaptive denoising applied, substantially more parameter configurations were found to result in acceptable levels of change. Only the parameter configurations using 10% dose resulted in problematic emphysema score changes. Thus, adaptive denoising provides a means to greatly improve the reliability of quantitative emphysema scoring, most importantly in cases where the scan or reconstruction fall outside of typically accepted standards.

While there is still more investigation needed, this dissertation illustrates that widespread quantitative emphysema scoring could be made more viable via the use of adaptive denoising, and in the absence of denoising only some parameter configurations yield acceptable quantitative results. Additionally, the high-throughput pipeline discussed can be applied to future, similar investigations regarding emphysema scoring, as well as investigations into other quantitative or computational imaging techniques.

The dissertation of John Marian Hoffman is approved.

Matthew Sherman Brown

Jonathan Goldin

Grace Hyun Jung Kim

Anand Prasad Santhanam

Frédéric Noo

Michael McNitt-Gray, Committee Chair

University of California, Los Angeles

2018

For my father, Damian,

whose intellect, ingenuity, care, and craftsmanship are present in all that I do.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I'd like to start by acknowledging the invaluable input and assistance provided by my advisor, Michael McNitt-Gray, without whose guidance this work would not exist in its current form. The road to this point hasn't always been perfectly straight, but Mike's support and patience has been consistent.

Thank you additionally to my entire committee for their participation and input on the final dissertation, Drs. Matthew Brown, Grace Kim, Jonathan Goldin, Frédéric Noo, Anand Santhanam. Their inputs and opinions have helped along the way to improve the work substantially. In particular, I would like to acknowledge the input of Dr. Grace Kim in the development and assistance with the statistical analysis approach employed in chapters three and four.

I would also like to acknowledge the substantial input of Dr. Frédéric Noo throughout this dissertation and my entire time in graduate school. His expertise in CT image reconstruction and the physics of CT was integral to the development of the entirety of FreeCT, and by extension, the infrastructure on which this project is built.

A huge effort has been made on the part of many people to collect the data for this work (and many more projects going forward), and for this I would like to acknowledge the contributions of Wasil Anwar, Nastaran Emaminejad, Anthony Hardy, Grace Lee, Caroline Chou, and Angela Sultan in collecting and organizing the endless stream of image and projection data.

The infrastructure development (i.e. “the pipeline”) enabling this work was the effort of many contributors. I’d like to acknowledge the contributions and insights provided by Wasil Anwar, Nastaran Emaminejad, Stefano Young, Pechin Lo, and Matthew Brown.

Last, but certainly not least, I would like to acknowledge the personal support and encouragement I have received from so many individuals in my life. In particular the love and support of my parents, Shawn and Damian Hoffman who have been there every step of the way. I never would have made it this far were it not for the many conversations about science and life with Jack Neylon over beers at Oldfield’s and Sepi’s. Kirsten Boedeker, for faith and trust in my MATLAB programming skills early on. And finally, my girlfriend Beth Wollman, whose company and patience have made the last few years so much more meaningful and enjoyable… and whose contribution of a coffeemaker is directly responsible for at least half of the text in this dissertation.

Vita

2007-2011	B.S. Mathematics Virginia Tech, Blacksburg, Virginia, USA
2007-2011	B.S. Physics Virginia Tech, Blacksburg, Virginia, USA
2014-2016	Intern, Quantitative Image Quality Toshiba Medical Research Institute, USA Vernon Hills, Illinois, USA
2016-2017	Imaging Scientist (part-time) Toshiba Medical Research Institute, USA Vernon Hills, Illinois, USA
2017	Moses A. Greenfield Award, UCLA Biomedical Physics Los Angeles, California, USA
2017	Dissertation Year Fellowship UCLA Graduate Division Los Angeles, California, USA
2013-Present	Graduate Student Researcher Department of Radiology David Geffen School of Medicine at UCLA Los Angeles, California, USA

Publications

(In preparation) *Hoffman, J.*, Kim, G., Young, S., McNitt-Gray, M. (2018). Evaluating the Effects of a Range of Acquisition and Reconstruction Parameters on a Quantitative CT Imaging Task, Part 1: Development of a Fully-Automated, High-Throughput Pipeline to Simulate a Range of Acquisition and Reconstruction Conditions.

(In preparation) *Hoffman, J.*, Kim, G., Young, S., McNitt-Gray, M. (2018). Evaluating the Effects of a Range of Acquisition and Reconstruction Parameters on a Quantitative CT Imaging Task, Part 2: Application of the Pipeline to Evaluate Quantitative Measures of Emphysema in a Large Lung Screening Cohort.

(In preparation) Zhao, T., *Hoffman, J.*, McNitt-Gray M., Ruan, D. (2018). A Modified Wiener Filter in the BM3D Algorithm for Ultra-Low-Dose CT Image Denoising.

(Accepted, pending revisions) *Hoffman, J.*, Noo, F., Young, S., Hsieh, S., McNitt-Gray, M. (2017). Technical Note: FreeCT_ICD: An Open Source Implementation of a Model-Based Iterative Reconstruction Method Using Coordinate-Descent Optimization for CT Imaging Investigations. Medical Physics, 2018.

Martin, T., *Hoffman, J.*, Alger, J., McNitt-Gray, M., Wang, D. (2017). Low Dose CT Perfusion with Projection View Sharing. Med. Phys., 45: 101–113.

Young, S., Lo, P., Kim, G., Brown, *M., Hoffman*, J., Hsu, W., … McNitt-Gray, M. (2017). The Effect of Radiation Dose Reduction on Computer-Aided Detection (CAD) Performance in a Low-Dose Lung Cancer Screening Population. Medical Physics, 44(4), 1337–1346.

Hoffman, J., Young, S., Noo, F., & McNitt-Gray, M. (2016). Technical Note: FreeCT_wFBP: A Robust, Efficient, Open-Source Implementation of Weighted Filtered Backprojection for Helical, Fan-Beam CT. Medical Physics, 43, 10 pp.

Invited Presentations

"Imaging Is a Numbers Game: Challenges and Breakthroughs in CT Quantitative Imaging." Imaging Elevated: Utah Symposium for Emerging Investigators. September 29-30, 2017. University of Utah, Salt Lake City UT.

"High Throughput Computing in Quantitative CT Imaging: Development and Applications." PKU-UCLA Joint Research Institute in Science and Engineering. 8th Annual Symposium. June 29-30, 2017. Peking University, Beijing, China.

Chapter 1 - Background and Motivation

1.1 Introduction

Clinical x-ray computed tomography (CT) was first introduced in 1971. The original scanner was limited to a pure "pencil-beam" geometry, was only for the imaging of a patient's head, and required approximately 4.5 minutes to render a relatively low-resolution (80x80 pixels) image. Since that time, modern CT scanners have improved to the point where scanners acquire thousands of multi-slice projections in under half a second, most scanners can image an entire adult thorax in one breath hold or less, and radiation doses are low enough that CT scanning is being used for lung cancer screening [1]. Modern computing technology has enabled new iterative reconstruction approaches that could enable substantial further dose reduction and information extraction, and machine learning has substantially increased interest in image processing of CT images. Because of these developments and CT's relative speed, accuracy, and low cost, it has become one of the most widespread imaging technologies in use today.

Since its introduction in in 1971, the gold standard for interpreting clinical CT images has been radiologists. In a research setting, often radiologists are employed alongside other readers, and while radiologists and readers are typically highly-skilled and specially trained for a given task, there is substantial interest in supplementing or enhancing human reader performance with quantitative or automated tools. Recent progress in the fields of machine learning and computing, as well as improved data collection and sharing have led to the rise of topics such as radiomics, disease classification, computer automated detection, and others. With the extraction of detailed quantitative information not readily available to the human eye, it is suspected that diagnosis and treatment of patients can be improved.

While there is no indication that the use of human readers will decline in the near future, there has long been interest in augmenting human readers with quantitative measurements made on the image data from CT scans. Early approaches to quantitative CT were limited to regions of the body that could be kept still for the relatively long periods of time required to acquire the CT scan, and thus were limited to extremities such as the arms, legs, and head. As far back as 1976, quantitative CT measurements were proposed for bone mineral density estimation in the radius and ulna [2]. As CT technology improved through the 1980s and 1990s, clinical imaging of the thorax became possible, and quantitative approaches to bone densitometry in the spine were developed [3]–[5] alongside initial efforts to begin quantifying diseases of the lungs [6]–[8].

Improvements in technology began to open up new frontiers of quantitative imaging using CT, and as more companies began to be involved in development of CT technology, new problems began to emerge. In particular, researchers began to notice that there was a need to standardize quantitative results across different scanners in order to produce the same diagnoses at different clinical sites using different technology. The call for standardization in quantitative CT extends as far back as the quantitative methods themselves, such as the proposed standardized phantom for lung nodule densitometry from Zerhouni et al. [9] and a thorough investigation into the impacts of scanner manufacturer, geometry, and acquisition time on CT numbers in lung nodules performed by McCullough and Morin [10]. McCullough and Morin argued that “great care must be exercised in attempting to use CT numbers to characterize tissue type/pathology" and furthermore provided the somewhat prescient recommendation that "if possible, [CT-number data should be accumulated] on your own equipment, for your own patients, and for one 'standard' set of scan parameters."

While both studies [[9], [10]] were conducted in 1983, modern clinical quantitative CT is still challenged and limited by the question of standardization. Further complicating the issue is the massive heterogeneity of modern clinical CT scanners, extending from the lowest hardware levels of CT scanners, such as detector scintillation material or electronics, up through the acquisition parameters, patient size and coaching, and reconstruction parameters, algorithms, and implementations. As the technology and applications of CT have improved, the challenge of standardization has grown more daunting, however in spite of the questions still lingering with regard to standardization of quantitative CT, modern developments in computing and the availability of large CT datasets has only made quantitative imaging more appealing in the last decade. In particular, modern developments in graphics processing units (GPUs) have sparked a potential revolution in quantitative CT. Their suitability for machine learning, image processing, and image analysis has extended the reach of quantitative CT beyond densitometry and into complex and exciting domains such as radiomics (mining image datasets for complex features and correlating values with either genomic data or disease expression), novel automated CAD (computer automated diagnosis/detection) and segmentation algorithms, and quantitative emphysema scoring. Despite the excitement however, only a limited subset of these quantitative CT technologies and techniques have found their way into day-to-day clinical usage (e.g. clinical quantitative calcium scoring, CAD in mammography).

Quantitative imaging has, to date, largely been limited primarily to the domain of research and academic interest, stemming from the fact that results obtained at one site, even for simpler quantitative tests, are not necessarily believed to be comparable in any broad sense to results obtained elsewhere. Buckler and Boellaard captured this challenge in their 2011 opinion article by identifying that “success [for quantitative imaging] will be achieved when results are broadly comparable and are widely disseminated rather than being possible only in highly selective and controlled environments.”[11] However, one area in which quantitative CT has found application is in the world of clinical trials, and in particular with the measurement of chronic obstructive pulmonary disease (COPD). Several large, multicenter trials have employed quantitative CT methods including SPIROMICS (Subpopulations and intermediate outcomes measures in COPD)[12], COPDGene (Genetic Epidemiology of COPD)[13], and others[14], [15]. Efforts to incorporate quantitative imaging of COPD into clinical trials have included a large focus on standardizing CT protocols between sites involved. This is challenging and requires substantial quality assurance and effort to ensure that all sites adhere, making this approach largely infeasible for broader clinical application.

Despite the current infeasibility of large-scale, cross-site standardization, interest in quantitative measurements of COPD remains extremely high, and one of the longest standing approaches to quantitative imaging with CT. With the recent approval of lung cancer screening with low-dose CT for Medicare reimbursement, a large population of subjects likely to benefit from the careful tracking of COPD status has emerged (i.e. former heavy smokers), in addition to other COPD patients. However, in a clinical environment of disparate protocols, non-standard reconstruction algorithms, and aggressive dose reduction, the question remains, are the quantitative measures obtained reliably indicative of a patient's underlying disease status, or are they heavily impacted by the conditions under which they were acquired?

1.2 Quantitative Imaging of COPD

Chronic obstructive pulmonary disease (COPD) is a class of diseases defined as "a preventable and treatable disease state characterized by airflow limitation that is not fully reversible."[16] Furthermore, "the airflow limitation is usually progressive and is associated with an abnormal inflammatory response of the lungs to noxious particles or gases, primarily caused by cigarette smoking."[16] According to the Centers for Disease Control, chronic lower respiratory disease is the third leading cause of death in the United States [17], behind heart disease and cancer, making it a critical target for modern healthcare; while environmental factors often play a role, tobacco smoking is identified as the main risk factor for COPD [16]. Smokers often exhibit both chronic bronchitis and emphysema, both classified under the umbrella of COPD, however emphysema is of particular concern for quantitative imaging.

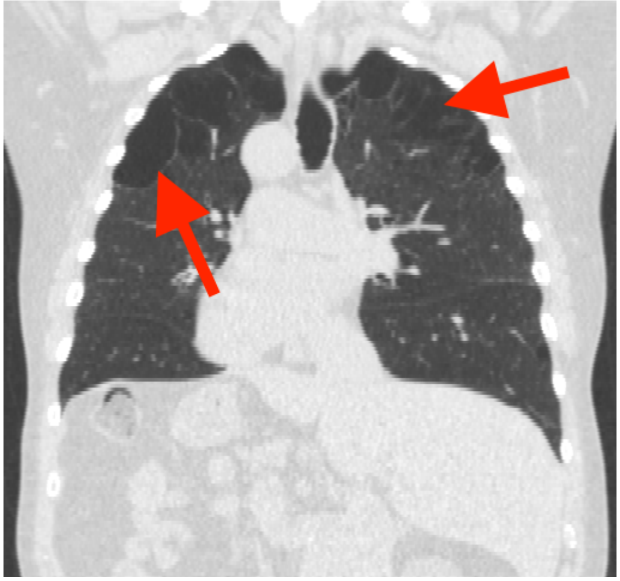


FIGURE 1-1: ILLUSTRATION OF CONTRAST BETWEEN A POCKET OF EMPHYSEMA AND BACKGROUND TISSUE IN A CT SCAN. EMPHYSEMATOUS AREAS HIGHLIGHTED WITH ARROWS.

Emphysema refers to the progressive destruction of alveoli in the lung parenchyma, resulting in trapped air and reduced gas exchange, decreasing blood oxygenation, and the inability to expel air from the lungs normally. As alveoli are destroyed, pockets of empty space are created and air infiltrates and remains. Pockets of emphysema in a CT exam are readily identifiable as darker regions of lower attenuation inside of the lungs, in contrast to the healthy, higher-attenuation regions. An illustration of this contrast is provided in Figure 1-1.

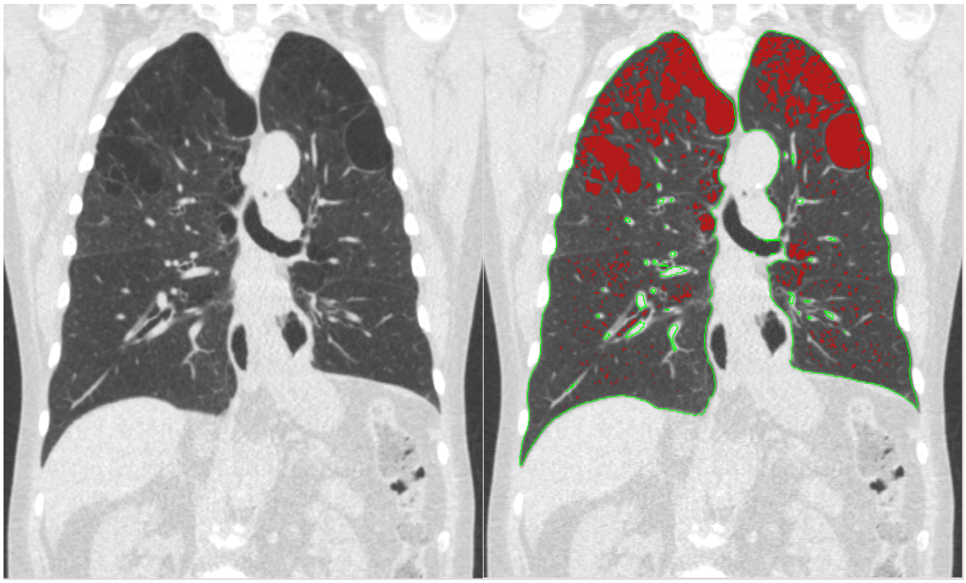


Figure 1-2: Sample density using a threshold of -950HU (i.e. RA-950 scoring). (Left) Original image without density mask, (right) With density mask overlaid. Lung segmentation outline shown in green, voxels of density less than -950HU are shown in red.

While initial efforts to quantitatively measure emphysema concentrated on trying to measure changes in the mean attenuation of the entire lung parenchyma [18], the fundamental approach to quantitative CT for the purposes of evaluating COPD was developed in 1988 by Muller et al. [6] using the concept of a "density mask." The density mask highlights voxels below a given threshold, which are then counted and considered against the entire area of the lung, as illustrated in Figure 1-2. Muller et al. then utilized pathological scoring of resected portions of the lungs and determined that the threshold that yielded the best correlation between the pathological findings and the quantitative CT density mask scores was -910 HU. While this study laid the groundwork for nearly every approach to quantitative emphysema scoring that has followed, it is not without issue when attempting to translate to modern CT. Of note in particular is the use of 10mm slice thicknesses when imaging, which is a slice thickness no longer routinely employed today and often results in substantial partial volume averaging. Additionally, each slice was acquired under a different breath hold (due to the slow acquisition speed of scanners in the 1980s), which has been found to impact emphysema scores derived from density masks [19]. Despite these minor issues however, the work of Muller et al. is still considered the seminal work of quantitative imaging of emphysema using CT [20].

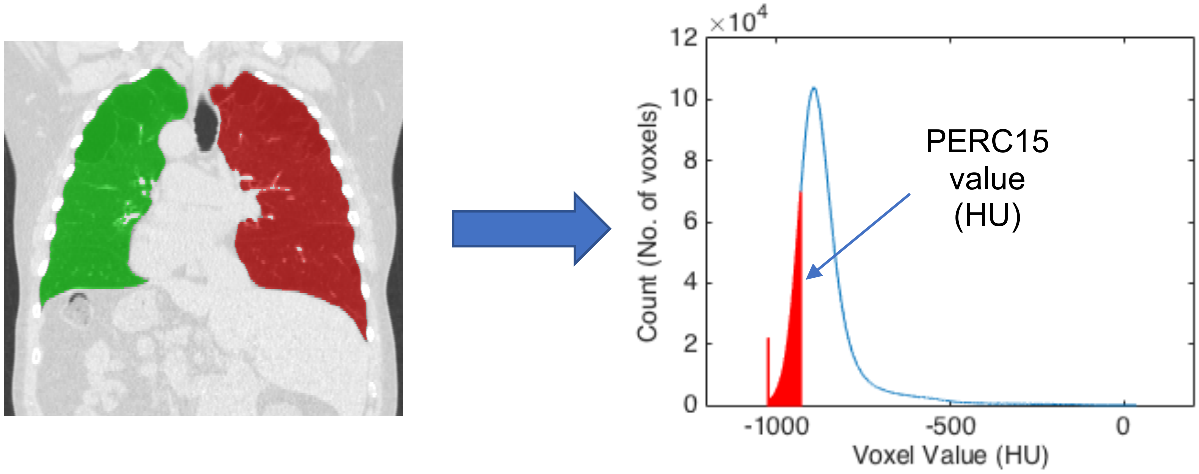


FIGURE 1-3: ILLUSTRATION OF PERC15 CALCULATION IN LUNG HISTOGRAM.

Recognizing that the -910 HU threshold may need to be revised as technology improved and CT slice thicknesses became thinner, several groups have further investigated the correlation between pathological scoring approaches and threshold selection for density mask scoring. This has yielded slightly different conclusions. Genevois et al. found that optimal correlation is achieved when utilizing a threshold of -950 HU [7], however their results were acquired using 1 mm slice thickness, acquired at intervals of 10 mm, while modern scans typically employ contiguous spacing for a full volumetric acquisition. Furthermore, their study utilized a beam energy of 137 kVp, while most adult thoracic CT studies utilize 120 kVp and shifts in the beam energy are known to cause changes in density measurements. In a similar study, Madani et al. 2006, it was found that either -960 HU or -970 HU should be utilized for optimal correlation between density mask scores and pathologic findings, however non-contiguous acquisition was also present in their study (1.25mm slice thickness spaced at 10mm intervals), and a beam energy of 140kV was utilized. Furthermore, Madani et al. [21] employed substantially lower tube current, 80mAs, than Genevois et al. [7], which, as will be shown later in this work, could have impacted the findings. Although these revisions to the optimal density mask threshold were proposed, the threshold most commonly employed today is -950 HU “in the interests of balancing sensitivity and specificity” [13], [20].

Density mask scoring with a threshold of -950 HU, also known as RA-950 (short for “relative area”), is the most commonly employed quantitative approach to scoring emphysema, however behind RA-950, measuring the location of the 15 percentile point of the histogram of lung parenchyma voxel values (PERC15, illustrated in Figure 1-3) has been found to correlate well with COPD amount. In patients with emphysema, PERC15 has been found to be significantly lower than in patients without, indicating that it also may be a good quantitative metric for evaluating emphysema [22]. Additionally, PERC15 has in some instances proven to be more reproducible and less susceptible to volume changes in the lung (i.e. changes due to breath hold) [23], [24].

Since RA-950 and PERC15 are the two most well-investigated approaches to measuring COPD in the lungs, they will be the focus of the work presented in this dissertation. It is worth noting however that because there are some known limitations in these approaches (discussed below) efforts to develop other methods are ongoing. Additionally, other methods have historically been tried with limited success. One method that has been explored but failed to reach mainstream adoption is the approach of measuring the power-law exponent of the cumulative frequency size distribution of the RA-950 mask. The cumulative frequency size distribution typically displays excellent agreement with the power law model given in [25], however it has been shown that D value does not correlate well with microscopic and macroscopic measures of emphysema [26]. More recently, there has been effort to build a measure of COPD using the mapping of voxels between inspiration and expiration scans [27]. By developing a deformation map between the two reconstructed scans, a voxel can be labeled as healthy, emphysematous, or “functional small airways disease” by assessing its change in value between the two scans. Initial results of the “parametric response map” measure to detect and score emphysema in patients and over time are promising [28] and it has been found to be fairly robust to sources of variability common in clinical CT [29]. Although it is extremely promising, parametric response mapping (PRM) for COPD is a fairly new technique (first published in 2012 by Galbán et al. [27]), and is currently only commercially offered by one vendor (Imbio, LLC, Minneapolis, MN). Implementations of PRM are challenging to develop due to the difficult problem of reliable deformable image registration between the inspiration and expiration scans.

Despite the number of different approaches that have been explored historically and the new approaches that continue to be developed and explored, RA-950 and PERC15 have gained the widest use and are, under proper conditions, reasonably well-trusted. Proper conditions however, are difficult to define, and as will be presented next, without a clear definition and understanding of those conditions, both RA-950 and PERC15 can display large amounts of variation. Because of this variation, quantitative scoring of COPD has failed to gain widespread clinical adoption.

1.3 Clinical Challenges

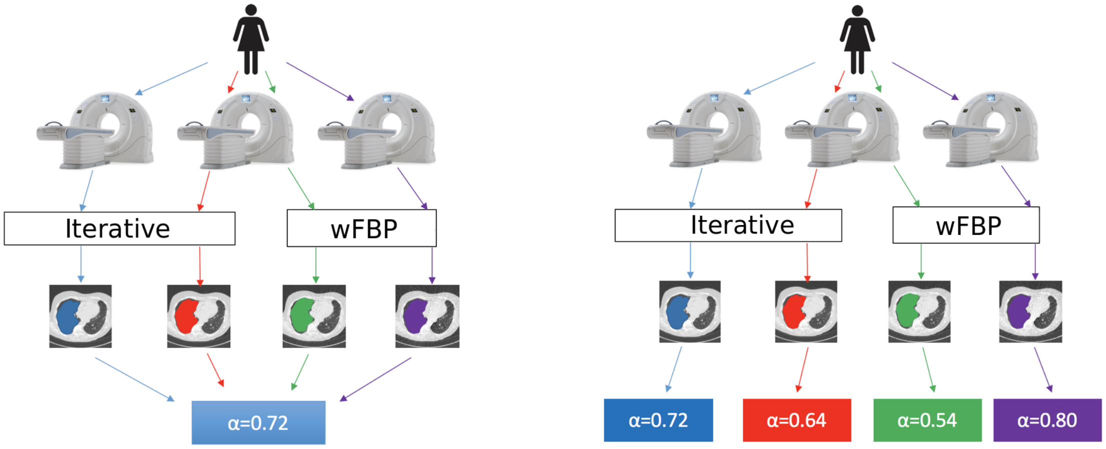


Figure 1-4: Ideal (left) versus realistic (right) scenarios for quantitative imaging. Variations in scanner utilized, reconstruction algorithm and implementation of the quantitative test lead to changes in final scoring values. Ideally, a quantitative test would only reflect the only the patient.

An ideal quantitative imaging test would reflect only the subject's underlying disease state or biology. This means that the quantitative imaging test would not be impacted by the chosen acquisition parameters, model or manufacturer of the scanner, method and settings of the reconstruction, or any other factors such as breath hold, timing of contrast, etc. (illustrated in Figure 1-4 (left)). In practice however, this is not the case, and most of these parameters have minor or major impacts on quantitative scoring approaches, illustrated in Figure 1-4 (right).

For an individual clinical site, common sources of variation for CT are typically tied to the acquisition and reconstruction parameters of a given scanner. Acquisition parameters include but are not limited to beam energy, tube current, detector collimation, focal spot size, bowtie filter, and use of tube current modulation. If a site has multiple scanners, scanner calibration, scanner model, and detector technology can also have large impacts on the final images produced for quantitative evaluation. Reconstruction parameters typically result in even larger changes in the final image. For the more common filtered backprojection reconstruction algorithms, these variations typically result from choice of reconstruction kernel, or selection of slice thickness. By adjusting the reconstruction kernel, a site can vary the tradeoff between image noise and resolution, optimizing based on the type of image and task required. Changing the slice thickness also affects image noise and resolution, however in the "longitudinal" direction; thicker slices result in less-noisy images, however also typically result in more partial volume averaging.

Today however, in addition to the long-standing filtered backprojection reconstruction approaches, iterative reconstruction algorithms are now being employed clinically. Iterative reconstruction algorithms are being employed for their apparent ability to improve image quality at current doses, or maintain image quality at substantially reduced dose. These iterative algorithms are either statistical (such as GE's ASiR algorithm) or model-based (such as Toshiba/Canon's FIRST algorithm) and often are strongly influenced by the selection of cost function and penalty term used for optimization. Furthermore, each penalty term typically includes its own set of parameters that can be adjusted to balance image features such as edge detail, noise, and contrast. While iterative algorithms appear to have the potential to improve or maintain image quality under substantial CT dose reduction, it is not well understood what their impacts are on the underlying quantitative properties of the reconstructed image, further complicating the already heterogeneous clinical environment of CT scanning.

While sites typically implement a standard protocol for most commonly-used scan types (e.g. routine chest, routine head, pediatric abdomen, etc.), variation within an individual site might be caused by a patient being scanned with different protocols or changes in protocols over time; or being scanned on different scanners can introduce drastic protocol changes (e.g. the lung screening protocol on the Siemens Force scanner versus the lung screening protocol on the Siemens Definition AS); or finally changes in patient coaching for reasons such as breath holds, or IV contrast. When multiple clinical sites are involved in imaging, such as for a multicenter longitudinal clinical trial, handling variation becomes even more challenging. Variability across sites' "standard" protocols is often substantial in regard to even the more physical acquisition parameters, namely beam energy and tube current. Often, multiple CT scanners from different manufacturers are utilized, which can have different beam spectra and manufacturers often implement new features such as tube current modulation in unique and non-standard ways. Reconstruction introduces further variability in algorithm use and implementation, including non-standard naming schemes, "black-box" (i.e. unknown or unpublished) pre- and post-processing techniques, etc.

1.4 Variation and Robustness Testing

It has long been recognized in the world of quantitative imaging that the sources of variation described above make widespread quantitative imaging challenging. In the particular case of emphysema scoring, a large body of work exists attempting to quantify the levels of variation observed in quantitative measures due to common clinical sources of protocol variation. In an early example, Boedeker et al. (2004) [30] looked at the impact of changing reconstruction kernel on density mask scoring. In a cohort of 42 patients, it was found that by changing the reconstruction kernel alone, away from the standard recommended kernel, shifts of up to 15.3% were observed, with a statistically significant average shift of +9.3% for one particular class of kernel. Earlier studies had found a mixture of results, with some finding that lung densitometry was reproducible across scanners and manufacturers [31], and others finding that protocol variations led to strong differences in density mask score [32]. Careful examination of these early study results however do not disagree with [30]: density mask scores were found to be reproducible utilizing phantom measurements, performing custom air calibrations on each scanner, and finally utilizing a near-standardized protocol (up to differences in manufacturer implementations). This level of control and standardization is nearly impossible to achieve in a realistic clinical setting.

More recent studies have further built upon the findings of [30] examining the impacts of acquisition dose, reconstruction parameters, and reconstruction algorithm selection on quantitative emphysema scoring. In particular, since there is concern regarding the ionizing radiation dose associated with CT, ensuring that it is kept “as low as reasonably achievable” and with the introduction of more standardized low-dose protocol recommendations, such as lung cancer screening, reduction in radiation dose has been of particular concern. In 2005, Gierada et al. [33] found that differences in RA-950 score in a cohort of 56 subjects was minimal between “standard dose” scans (120 kVp, 100-250 mAs effective tube current) and a reduced-dose scan (120 kVp, 30-60 mAs effective tube current). This indicates that “precise consistency of exposure factors is not necessary when CT scans are used for comparative studies of emphysema… in contrast to differences related to slice thickness and reconstruction filter," [33] although as will be shown later in this work, this is only true up to a point.

In 2007, Trotta et al. [34] utilized phantom measurements to evaluate the impacts of tube current reduction and reconstructed slice thickness on several measurements of emphysema, including mean value and standard deviation (no longer employed for emphysema evaluation), PERC15, and RA-910 (note that RA-950 is the more commonly used value today). While there is substantial depth in their investigation that will not be discussed here, a key result is that there existed threshold tube current after which measurements of RA-910 would begin to deviate from the expected value, which was known exactly since a lung phantom with calibrated materials was employed. As tube current continued to be reduced further, the deviations became larger. Additionally, they highlighted that there were regional differences between the apices and base of the lungs. These results are highly informative about changes in density mask scoring one might potentially see as a result of dose reduction or slice thickness changes, however they provide little insight into the realities of emphysema scoring in a clinical setting with a real patient, or possible pathways forward to improving emphysema quantification. Finally, phantom studies typically do not account for realistic population features, such as patient size, differences in disease state, breath hold, etc.

Many other studies have taken similar approaches to testing the impacts of dose reduction and reconstruction technique on quantitative emphysema scoring: a common source of protocol variation (e.g. dose, reconstruction kernel, etc.) is selected and evaluated at two or more values, and change in quantitative emphysema scores are measured. Dose reduction is perhaps the most commonly investigated [34]–[37], followed closely by slice thickness [7], [21] and reconstruction kernel [30], [35]. More recently, several investigations have explored the impacts of clinically available iterative reconstruction approaches on quantitative emphysema scoring [36]–[39]. Results of these investigations have been promising, suggesting that iterative reconstruction in many cases does not have a very large impact on quantitative emphysemas scores, however the results are often of limited utility when attempting to translate them to the broader clinical world. For example, in [37], "standard dose" CT scans (~10mGy CTDIvol) were compared against "ultra-low-dose" CT scans (~0.5mGy) with and without iterative reconstruction (Toshiba AIDR 3D). One single fixed set of reconstruction parameters was employed, for one scanner. While AIDR 3D largely had the effect of restoring the ultra-low-dose RA-950 scores to their standard dose values, it is difficult to infer the reasons for this due to the number of data points investigated (only two: standard and ultra-low dose). Additionally, the reader is left wondering if these same results would hold up under changes in slice thickness, or more moderate dose reduction.

The majority of the studies above highlight two key gaps in the current body of work regarding the robustness of quantitative emphysema scores: (1) the number of data points investigated for a given test parameter and (2) the lack of consideration for possible interactions between different acquisition and reconstruction parameters. The exception is one study conducted in 2010 by Gierada et al. [35] investigating the impacts of slice thickness and reconstruction kernel combination on emphysema scores. Five reconstruction kernels were tested at four slice thicknesses for a total of 20 reconstructions per subject (N=21). By investigating two reconstruction parameters simultaneously and at a number of different settings, they were able to demonstrate the behavior of RA-950 scoring over a realistic clinical range of those two parameters. Because of the careful, multi-parameter approach utilized in [35], and the fact that their subjects had broad range of levels of emphysema, it was further observed by the researchers that the amount of emphysema present in a patient's lungs strongly impacted how susceptible their RA-950 score was to parameter changes; subjects with high levels of emphysema displayed substantially less variation in emphysema score as a result of kernel or slice thickness change when compared to subjects with little or no emphysema. This fact was not taken into account in any of the previously discussed studies, further complicating the interpretation or translation of those results. Finally, while Gierada et al. [35] did not highlight it in their discussion, their results illustrate that there are ranges of acceptable parameters that result in equivalent, or nearly equivalent emphysema scores. This concept of "regions" of CT parameter space that produce equivalent emphysema scores (or equivalent for diagnostic purposes) is one that will be explored at length in this dissertation.

Aggregating and interpreting all of the existing emphysema-scoring robustness studies is challenging: there are a large number of studies investigating similar topics in slightly different manners all claiming different levels of success in finding parameter configurations that make emphysema scoring "robust." However, there are fortunately some key lessons to be learned from all of these studies. First, there is an intense desire to use emphysema scoring clinically, and as a result make it more robust to acquisition and reconstruction parameter change. Second, most studies were able to identify at least some combination of acquisition and reconstruction parameters that resulted in emphysema scores matching a reference score, which means theoretically that for most image datasets there could exist a means to map image data or emphysema scores to their true values. The next challenge facing quantitative emphysema scoring is identifying a correct approach to develop such a mapping for image data or scores, and understanding the methods and potential limitations of that approach.

1.5 Potential Pathways to Improving Emphysema Scoring

1.5.1 Standardization

One of the more obvious possible solutions to reliable quantitative imaging would be to simply recommend or require a standardized protocol when acquiring and reconstructing images for the purposes of quantitative emphysema scoring. While this is an interesting theoretical exercise, practically, it is almost impossible. One of the first real attempts at multicenter protocol standardization for the purposes of evaluating the feasibility of CT for a given task was developed for the National Lung Screening Trial [40]. In [41], Cagnon et al. detail the protocols, carefully specified across all manufacturers and scanner models likely to be utilized in the study, and additionally the required quality-assurance program that was developed to ensure that sites, scanners, and imaging protocols met the inclusion requirements for the study. However, even in a well-funded, well-supported study it is noted that ongoing cooperation typically requires a “local site champion” to help maintain compliance and interest in the trial. It is also discussed how, over the course of the lung screening trial, CT technology developed from 4- and 8-slice scanners into predominantly 16- and 64-slice CT scanners, and standards and protocols must be ready to adapt to the ever-changing world of CT technology. Unfortunately, much of what was described by [41] lies well outside of the normal support structures and budget available for the purposes of routine clinical imaging. Recommendations, similar to those in Table 1 of [41], can be made, however it is difficult to guarantee that sites are aware of them and using them.

There are two large-scale projects that have provided recommended standardized protocols for quantitative imaging as part of their efforts: the COPDGene Project [13] and SPIROMICS [12], [42]. The efforts of COPDGene are to identify genetic factors associated with COPD, and one of the methods of phenotyping COPD with CT is the measurement of RA-950. The protocols recommended in the appendix of [13] could be regarded as a good starting point for quantitative emphysema scoring, and for patients enrolled in the COPDGene study scores are likely to be reliable means of tracking disease state and possible progression. The SPIROMICS recommended protocols closely resemble those of COPDGene, which is to be expected given similar quantitative imaging end points, and thus is also a very reasonable choice for a starting point for protocol standardization. The SPIROMICS protocols are also somewhat more recent and includes recommendations for more modern scanners. A key component of both studies however, as with the NLST's study design, is the quality control portion, which ensures that a scan being used for quantitative evaluation fits within the recommended protocol specification. Despite the fact that these larger multicenter trials have achieved reasonable success with standardizing quantitative CT protocols, the question remains however, in routine clinical practice (i.e. the absence of a well-funded, tightly-controlled quality assurance program), is standardization a reasonable approach to reliable quantitative imaging?

Standardization without quality assurance as an approach to reliable quantitative imaging is realistically not a viable option for a number of reasons, however it is worth noting that several larger organizing bodies have worked to provide some level of standardization/recommendation in the CT domain outside of the above multicenter trials. Since 2010, the American Association of Physicists in Medicine (AAPM) has provided general recommendations for routine scans on a variety of scanners, and the American College of Radiology provides the CT Accreditation Program [43], which is required for reimbursement of scans billed to Medicare. The goals of these recommendations however are not to standardize image quality or ensure that reliable quantitative imaging is performed. In the case of the ACR, it ensures that minimum acceptable imaging standards are being met (testing routine protocols for minimum contrast to noise ratio, acceptable attenuation values for a calibrated phantom, etc.), and the AAPM's efforts are an attempt to provide recommendations that balance the needs of image quality, dose, and are "reasonable and appropriate for a specific diagnostic task" [44]. Despite these efforts however massive variability still exists.

One common challenge, even within controlled multicenter trials, is technological evolution [41], [42]. Over time, protocols must be reevaluated to incorporate new scanners, reconstruction approaches, etc. Features common to newer scanners are sometimes unavailable on older scanners and could cause substantive differences in image quality and impact the usability of the scan for quantitative imaging. Examples include nearly all modern iterative reconstruction algorithms as well as detector technology such as the Siemens Stellar Detector [45]. This fragmentation of technology poses significant challenges when trying to specify a standard protocol, since patients visiting the same doctor often end up being imaged on different scanners. Another significant challenge to standardization in the clinical environment is protocol errors: patients are occasionally scanned using imperfect protocol settings. If standardization of the protocol is the requirement for quantitative imaging using the scan, this would cause the patient to perhaps miss a key time point or require the patient to be rescanned (often not an option due to radiation dose concerns). Thus, strict standardization cannot be the only criteria for usability of a scan for quantitative imaging. Furthermore, a standardizing body for every quantitative imaging task would be required that has extensive knowledge of most modern imaging equipment and the task to which they are assigned, making such organizing bodies unlikely in the near future.

1.5.2 Image Post-Processing and Normalization

Accepting that there will be imperfect or minimal standardization of CT protocols in a broad sense, other approaches to achieving reliable quantitative emphysema scoring would be (1) enumerating and testing all protocols yielding acceptable values, which is similar to standardization, and largely infeasible for many of the same reasons, (2) developing an understanding of reasonable parameter “regions” that yield acceptable quantitative imaging results, or (3) developing image processing methods that could normalize or restore image data to a minimum or standard level suitable for quantitative evaluation. The concept of developing acceptable parameter regions has never been fully realized due to limitations of the existing available research infrastructure for quantitative imaging (discussed in more detail below). Most approaches published to date attempt to correct for known, confounding features of CT imaging that cause problems for density masks, such as image noise or too sharp of a reconstruction kernel.

In an early example, corrections were applied to account for fluid pooling in the back of the lung, an overly noisy image due to either dose reduction or reconstruction kernel, and lung motion artifacts [46]. Also proposed in 2001, was a method based on clustering of voxel groups within the density mask, which accounts of the biological concept that emphysema occurs in "pockets" and isolated low attenuation voxels are unlikely to actually reflect true emphysema [47]. More recent approaches have incorporated the concept of image "normalization," meaning transforming image data to more closely resemble a defined standard. In [48], Gallardo-Estrella et al. proposed an image processing technique based on frequency decomposition to alter image noise expression; essentially a method to correct for the variation of emphysema score due to kernel selection, such as that observed in [30]. In their work, the target noise expression should resemble that of a Siemens B31f reconstruction kernel. Combining several of these ideas into one complete normalization approach, Gallardo-Estrado et al. have recently (December 2017) built on their previous work proposing "normES," ("normalized emphysema score") a combination of their preprocessing technique and the standard RA-950 density mask approach [49]. This refined method includes resampling of slices into non-overlapping 3mm slices, the image normalization described in [48], and finally the minimum cluster size requirement, similar to that of [47]. While all tests [46]–[49] showed improvements in the measured end-points, the work described in [49] showed extremely strong results for the proposed normES as a biomarker for mortality (both from lung cancer, and all-cause) in multicenter lung screening populations (i.e. the NLST dataset). Furthermore, the researchers theorize that previous studies finding that emphysema score was a weak predictor of lung cancer risk or all-cause mortality [50], [51] was due primarily to variation in acquisition and reconstruction protocol, which the proposed normalization process helps correct for. Thus, image post-processing and normalization present an extremely promising pathway forward for improved quantification of emphysema.

1.5.3 Regions of "Stability" For Emphysema Quantification

The final possibility towards a pathway for more widespread clinical quantitative emphysema scoring would be to determine if there are regions of parameter space that produce consistent results. For example, such a study would seek to provide a given range of reconstruction kernels, slice thicknesses, and doses that will result in emphysema scores that would not impact a subject's emphysema diagnosis. At its core, this type of study is a more rigorous extension of the previously discussed robustness studies, where, instead of investigating one CT acquisition or reconstruction parameter at 1-3 potential values, multivariate parameter space (two or more parameters) would be systematically explored at many combinations of values, such as in [35]. If such regions of stability (i.e. stable output emphysema score) exist, then recommendations can be made regarding the usability of an image dataset for the purposes of emphysema quantification. It is also possible that from such a study any potential interactions between parameters can be investigated, and further understanding of the physical and mathematical properties of CT scanning that produce a stable quantitative emphysema result can be determined.

To date, this type of exploration has largely been intractable due to the existing means of obtaining clinical CT image data. While this topic will be explored in more detail in Chapter

2, obtaining large numbers of clinical CT image datasets at present typically requires some tradeoff between cohort size, number of CT imaging parameters investigated, and/or a tightly-controlled dataset, stemming from the fact that most datasets are built using clinical scanners. Additionally, CT has the added challenge of radiation exposure, meaning patients can seldom be scanned multiple times, precluding the rigorous investigation of some parameters (such as dose or beam energy). While building datasets using the clinical scanner is possible, it is extremely labor intensive and time consuming since the workflow of the scanner is built for clinical use and not research. Automation of reconstruction tasks and quantitative image tests with the scanner is not typically possible, and conducting them using the interfaces provided by the scanner requires the full attention of a researcher over the course of weeks or months.

At present, a further obstacle to building custom datasets is obtaining access to the raw projection data. This data is not readily available for export on most scanners, is cleared from the scanner typically after a short period of time. If accessible, the proprietary file formats must be decoded via reader libraries or scripts provided by the manufacturer or reverse engineered. With access however, new modes of study are possible, such as the dose reduction simulation utilized in this work, sinogram domain preprocessing, or even research into new reconstruction algorithm design. If raw data access is obtained, however the scanner is still utilized for reconstructions, data must still be imported and exported to the scanner if any processing outside of what is offered on-scanner is to be done, adding another layer of complexity and challenge to research workflows that leverage the clinical scanner; this limits the number of quantitative tasks that can be investigated, as well as cohort size, and number of parameters that can be investigated.

1.6 Review and Discussion

While significant headway has been made on the path to quantitative emphysema scoring, there is still substantial work that needs to be done prior to its widespread clinical use. Some key areas of success have been highlighted over the course of this chapter as well as some of the key limitations. First, when quantitative emphysema scoring with CT is supported with clear protocol designations and a thorough quality assurance program, such as in the COPDGene and SPIROMICS studies, it has been found to be reproducible, however few clinical sites have the funding and support for careful day-to-day quality assurance. It is also highly likely that some scans will be acquired outside of recommended protocols, and it is unclear if these are still usable for quantitative evaluation.

Next, substantial investigation has been conducted to date exploring the “robustness” of emphysema scoring to acquisition and reconstruction parameter variations, however these have typically only targeted one parameter at a time, and only a small number of settings of that parameter. As a result, translating the results of these studies into clear implications or guidelines for clinical use of emphysema scoring, has proven challenging or impossible. One of the reasons this has failed to be explored rigorously is the challenge of assembling large-scale, thoroughly controlled, and detailed datasets using the clinical systems. If such investigations could be carried out, then regions of parameter space that result in stable quantitative emphysema scoring could potentially be found, providing insight into the underlying physical and mathematical mechanisms that produce reliable reproducible emphysema scoring results.

Finally, image “normalization” via image post-processing has been found to perhaps be an excellent method of improving the stability of quantitative emphysema scoring results, however has yet to gain widespread utilization. While one recently introduced method [49] has shown exceptionally promising results, there may be other simple and reasonable approaches to image post-processing that could improve the quality of emphysema scoring as well, meriting further investigation. Thus, continued investigation into these areas and improvements to the quantitative imaging research infrastructure stand to provide improved insight into the details of robust quantitative imaging in a clinical setting, and guidance and strategies for its correct use and interpretation.

1.7 Specific Aims

The immediate goal of this work is to characterize and minimize the impacts of CT parameter selection on quantitative emphysema scoring, with a broader goal of providing strategies and guidance for the clinical use of quantitative emphysema scoring. In order to accomplish this goal in a manner that builds upon the existing body of work and provides new insight into problems facing quantitative emphysema scoring, all of the investigations presented here consider combinations of three commonly varied parameters: reconstruction kernel (and its iterative equivalent), slice thickness, and acquisition dose. For each parameter, at least three reasonable settings are utilized, four in the case of acquisition dose. Furthermore, the investigations are conducted in a large cohort of subjects scanned with the lung screening protocol at our institution. By capturing many different parameter configurations, possible interactions between parameters are investigated in a new manner, and regions of parameter space that produce “stable” emphysema scores can be identified. Further extending our sampling of clinical variation, we also evaluate the same sampling of CT parameter space, using a commercially available iterative reconstruction approach, and contrast it to an implementation of more traditional weighted filtered backprojection. This provides an unparalleled starting point for the final evaluation of a potential approach to image post-processing for the purposes of reducing variation due to parameter selection. The size of these datasets however represents a significant data acquisition and processing challenge that must first be addressed prior to quantitative analysis.

The specific hypotheses tested in this work are the following: (1) that there exist regions of CT parameter space that produce emphysema scores close enough to a clinical reference protocol to not change scores more than 5%, and (2) that a denoising approach applied to the image data, here bilateral filtering is employed, will allow for more stable and reliable emphysema scoring under a broader range of acquisition and reconstruction parameters than images without any post-processing. To investigate these hypotheses, the specific aims of this dissertation are the following:

SA1: Develop a high-throughput fully-automated quantitative image analysis platform for CT

Offline reconstruction algorithms, analysis tools, and automation frameworks are developed to enable more rapid and thorough exploration of quantitative imaging topics, namely emphysema scoring for this work.

SA2: Characterize robustness of emphysema scoring to changes in reconstruction kernel, reconstructed slice-thickness, simulated dose reduction

The tools developed in SA1 are used to assess a large number of acquisition and reconstruction conditions in a large cohort of subjects scanned with a lung screening protocol. Parameter configurations that result in stable emphysema scores are enumerated and assessed for both filtered backprojection and an iterative reconstruction algorithm.

SA3: Evaluate the impacts of post-reconstruction denoising on the quantitative imaging tests from SA2

Denoising is applied to the image data using a bilateral filter and resulting levels of change in emphysema scores are quantified.

Chapter 2 - A High-Throughput Reconstruction and Quantitative Analysis Pipeline for CT Image Data

Significant portions of the material for this chapter are adapted from three manuscripts:

(In preparation for submission to Medical Physics) J. Hoffman, N. Emaminejad, M. Wahi-Anwar, G. Kim, S. Young, M. McNitt-Gray. "Technical Note: Design and Implementation of a High Throughput Pipeline for Reconstruction and Quantitative Analysis of CT Image Data."

J. Hoffman, S. Young, F. Noo, and M. McNitt-Gray. "Technical Note: FreeCT_wFBP : A robust, efficient, open-source implementation of weighted filtered backprojection for helical , fan-beam CT." Med. Phys. **43**(3), 10 pp. (2016).

(Submitted, under review, Medical Physics) J. Hoffman, F. Noo, S. Young, M. McNitt-Gray. "Technical Note: FreeCT_ICD: An Open Source Implementation of a Model-Based Iterative Reconstruction Method using Coordinate Descent Optimization of CT Imaging Investigations."

2.1 Introduction

In this chapter, we describe the technological and research infrastructure developments that were carried out to make the experiments described in Chapter 3 and Chapter 4 possible. It is important however to motivate why such infrastructure developments were necessary. One of the key challenges of the work pursued in this dissertation, and one of the larger challenges facing the quantitative imaging community, is that of data availability. In this work, we seek to create datasets involving large number of subject scans (N=142) reconstructed to represent a wide variety of realistic, clinical parameter configurations. Each scan was reconstructed with weighted filtered backprojection (WFBP) using 36 unique combinations of kernel, slice thickness, and acquisition dose. Each scan was additionally reconstructed with reasonably paired iterative reconstruction parameters and the Siemens SAFIRE reconstruction algorithm for a total of 72 unique reconstructions per subject, and ~10,200 image datasets for processing and final analysis. At the outset of the project, no tools existed to automate the creation, organization, or processing of an imaging dataset of this size.

While the creation of a custom dataset was required for this work, it is important to note that there are several alternatives for researchers wishing to conduct quantitative imaging studies. While these are good options, they are not without limitations. Publicly available datasets such as LIDC [52], NLST [40], Maastro NSCLC [53], etc., represent one means to access large-scale datasets. However, these are limited by the number of reconstructions available per patient (typically one), the reconstructions selected (optimized for human readers, not necessarily computer vision or algorithms), and are typically heterogeneous in terms of scanner and the protocols used to acquire the data. While this is a good representation of clinical variability, it is challenging if not impossible to achieve a highly-controlled dataset when reconstruction parameters are to be systematically varied and investigated. Finally, large scale public datasets quickly go out-of-date due to the turnover and updates of CT imaging technology, such as increasing numbers of detector rows, improvements in automatic exposure control, more efficient detectors, and new advances in reconstruction techniques. Another approach is retrospective assembling of datasets from PACS. This allows for more control over the reconstructions acquired, however is still limited to reconstructions optimized for/selected by human readers, and requires substantial time and effort to assemble datasets large enough for some studies.

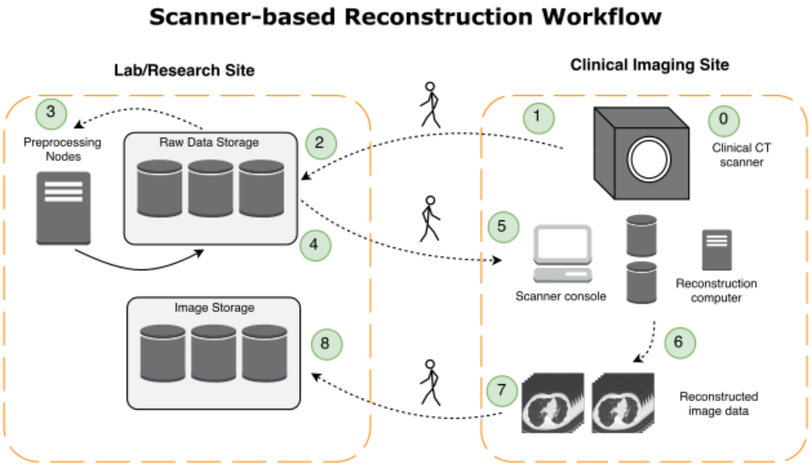


Figure 2-1 Illustration of workflow using clinical scanner reconstruction. (0) A clinical study is performed and the raw projection data is temporarily stored in the scanner database. (1) Raw projection data is exported to an encrypted hard drive, returned to the lab and then (2) loaded into a secure raw data storage. If preprocessing (3), such as reduced-dose acquisition simulation, is desired raw data is copied to a network node, preprocessing is performed, and the modified raw file is pushed back to network storage. When data is ready to be reconstructed for an experiment it is (4) loaded back onto an encrypted external hard drive, carried back to the scanner, and (5) uploaded back into the scanner database. (6) Reconstructions are manually configured and performed across all cases and desired reconstruction parameters. (7) All reconstruction image data is exported from the scanner back onto the encrypted hard drive, raw data is deleted from the scanner, and finally, image data is returned to the lab-based network storage location for reconstructed image data. Significant human intervention is required at each step of the seven-part workflow, indicated with dashed arrows.

A promising approach pursued in our work has been the collection and storage of raw projection data from CT scanners, and then subsequently returning to the scanner at a later date to perform reconstructions. This has been employed in several studies from our group [54]–[56] to great effect, and furthermore this allows for preprocessing of the raw projection data, such as simulated noise addition/dose reduction and projection domain denoising, opening up new research pathways not previously possible. However, the workflow of collecting, processing and returning raw data to the scanner (illustrated in

Figure 2-1) is not without substantial logistical limitations. Raw projection data is not widely available and access to that data typically requires some cooperation from the manufacturer (including permission and information about the file format which is typically considered to be proprietary). Even when it is available, the raw projection data must be re-imported to the scanner, the scanner reconstructions cannot be operated in “batch-mode,” and reconstructions different than the clinical protocols often cannot be preprogrammed, and must be manually configured via a graphical user interface. Finally, all image data must be exported and returned to the lab site, uploaded to a secure network share and organized for storage and future use. When cohort sizes are small and only a few reconstructions per subject required, this is a viable approach, however it quickly becomes burdensome. For example, to assemble the dataset used in Young et al. 2017 [55] required six months (3 reconstructions per patient for 481 patients). Increasing the number of parameters investigated, or investigating an additional source of variation (e.g. acquisition dose, reconstruction kernel, or slice thickness) dramatically increases the time and labor required to generate datasets used in QI analysis. If timely, large scale quantitative imaging research is to be conducted, an improved approach is required.

In this chapter, we develop and detail the construction of such an automated system (referred to as “the pipeline") including the reconstruction portion, illustrated in Figure 2-2, and analysis components. In addition to the pipeline itself, we also detail the development of open-source CT image reconstruction software that enables the automated reconstruction portion of the pipeline, FreeCT_wFBP and FreeCT_ICD. FreeCT_wFBP is utilized extensively in the rest of the dissertation for all filtered backprojection reconstructions; FreeCT_ICD was not directly employed in this dissertation however it represents an important extension of the pipeline framework to cover a relevant topic in modern quantitative CT research: model-based iterative reconstruction.

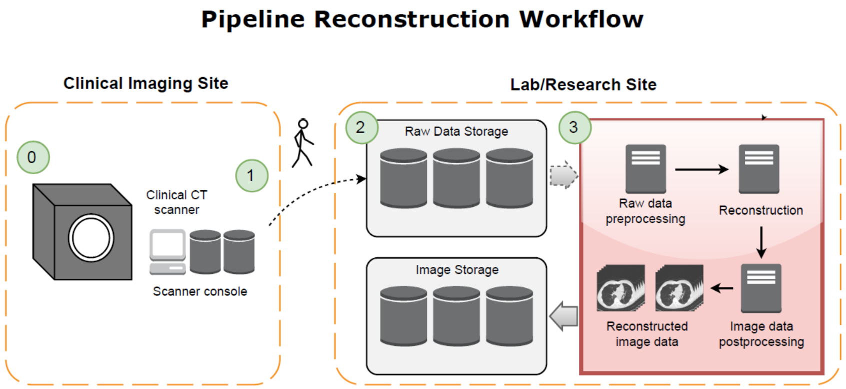


Figure 2-2 Illustration of the reconstruction workflow utilizing the pipeline. (0) Clinical scans are acquired and raw projection data is temporarily stored in the scanner database. (1) Raw projection data is identified and exported to an encrypted hard drive, returned to the lab, and (2) uploaded to network-based network storage for raw data. When data is ready to be reconstructed, (3) the pipeline is set up with one configuration file or via the GUI interface. The pipeline then manages all data-fetching, preprocessing, reconstruction, and uploading to the network-based image storage. No human interaction is required after the pipeline configuration. Image post-processing (prior to storage and subsequent analysis) may include, for example, denoising or other image-domain enhancement technique.

It will be shown that the developed pipeline, leveraging the open-source reconstruction software, achieves the following: (1) allows for a wide range of acquisition and reconstruction parameters to be configured and applied to raw CT projection data, (2) performs the high-throughput reconstruction of large data sets, (3) automatically organizes the resulting reconstructed volumes for archiving and QI analysis, (4) allows for highly configurable post-processing and analysis to be applied to the reconstructions, (5) produces QI results in a manner to facilitate easy, rapid statistical analysis, and finally (6) functions as an automated tool that requires minimal human intervention after initial configuration. To illustrate the utility and performance of the pipeline in the setting of quantitative imaging, a cohort (N=142) of low dose lung cancer screening exams with a wide range of acquisition and reconstruction conditions (36 combinations of slice thickness, reconstruction kernel and simulated acquisition dose) was created for analysis in Chapter 3 and Chapter 4.

2.2 Pipeline Overview

The pipeline is a collection of compiled programs and Python scripts designed to carry out reconstruction and quantitative imaging analysis. While the pipeline should be thought of as a generalizable framework for high-throughput imaging work, it has been developed thus far with the specific application of robustness testing of quantitative imaging metrics for diagnostic CT, which involves the evaluation of a quantitative imaging test across a range of different acquisition and reconstruction parameters such as slice thickness, reconstruction kernel, and acquisition dose. This application however gives an excellent example of the different, more general components of the pipeline. For this work, the pipeline has been specifically configured to test the robustness of quantitative emphysema scoring approaches using CT image data.

The pipeline workflow, illustrated in Figure 2-3, is roughly the following (1) reading of the raw projection data (2) raw data preprocessing (3) reconstruction (4) image data processing (5) analysis and (6) final results. Initially raw projection data must be parsed into a format readable by the reconstruction software. At present, the freely-available, open-source version of the pipeline accepts a binary format as well as an open-format, vendor independent DICOM format [57]. A customized version is employed in our lab that is able to read directly from Siemens raw projection data files. Thus, the pipeline is easily extensible to work on non-standard, proprietary data via the programming of a small raw data reading module that converts into either of these two open formats. After reading the raw data, and desired preprocessing is applied. In the case of robustness testing, a calibrated noise addition module [54], [58] is used at this stage to simulate reduced-dose scans, however this could also be other processing steps such as projection-domain filtering, or denoising algorithms (e.g. [59], [60]).

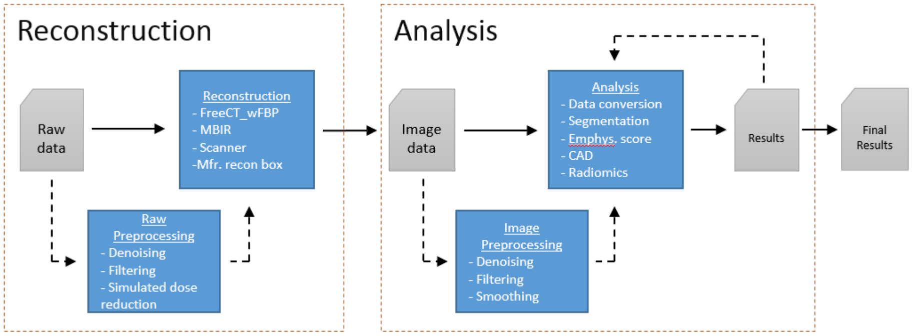


FIGURE 2-3 BLOCK DIAGRAM OF PIPELINE WORKFLOW FROM RAW PROJECTION DATA TO FINAL QUANTITATIVE IMAGING DATA. EACH BLOCK REPRESENTS A SELF-CONTAINED TASK THAT CAN BE ENCAPSULATED IN ONE OR MORE "MODULES." DASHED LINES REPRESENT OPTIONAL PROCESSING PATHWAYS. MODULES CAN BE PROGRAMMED BY THE USER AND INCORPORATED DIRECTLY INTO THE PIPELINE AUTOMATION FRAMEWORK OR CAN EXIST OUTSIDE OF THE HIGH-THROUGHPUT FRAMEWORK AS NEEDED, SUCH AS THE RECONSTRUCTIONS COMING FROM THE SCANNER OR SOME OTHER ALTERNATIVE SOURCE. EXAMPLES OF THE TYPES OF MODULES ARE GIVEN IN EACH BLOCK.

Reconstruction of the raw data is performed next. While FreeCT_wFBP [61] is utilized at present, the pipeline can accept image data from most sources, including other open-source reconstruction software, and image datasets reconstructed at the clinical scanner. This is made possible via a data conversion module that accepts image data in multiple formats (including DICOM, NIfTI, mhd, binary data and others) and converts to the format utilized by the analysis modules. This also allows for image-domain processing (e.g. denoising, smoothing, etc.) to be applied after reconstruction if desired. In the case of robustness testing, image denoising algorithms are being tested to assess their ability to “stabilize” quantitative measures (i.e. reduce the variation caused by changing reconstruction and acquisition parameters).

Finally, analysis is carried out through a series of modules that perform tasks required to produce the final result. In this work, the analysis performed is emphysema scoring and requires the modules that perform image conversion, segmentation, calculation of a lung histogram, and finally the emphysema scoring and aggregation of final results. This is discussed more in sec. 3.D. analysis modules. The pipeline is designed so that each module can be replaced based on the requirements for a given experiment, enabling future experiments to leverage the underlying high-throughput design and framework to automate and accelerate imaging data generation and analysis. In order to fully realize the pipeline concept, several key software developments were necessary, including FreeCT_wFBP, FreeCT_ICD, and a purpose-built GPU queuing framework to schedule and carry out reconstructions in an automated manner.

2.3 FreeCT_wFBP

The pipeline as described above was first conceived of following the work in [55]. It was recognized that the interest in advanced uses of CT that involve quantitative imaging, radiomics, and CAD for lung screening and other applications [62]–[69] is growing, however the ability to generate data utilizing the clinical scanners was limiting researchers’ abilities to fully investigate the topics. To ensure that these applications are robust, they need be tested across a wide variety of scanner platforms, acquisition conditions, and reconstruction parameters [54], [70]. Additionally, studies often have been limited to dozens of patients, rather the larger cohorts often needed to establish statistical power (e.g. [35],[37]). Few resources exist to support this type of endeavor. The most time-consuming aspect of [55] was returning to the scanner to perform reconstructions, discussed in the previous section.

Relying on the scanner for reconstructions presents additional concerns as well. First, the scanners are typically only available for research activities outside of clinical operation hours, reducing the time available for large-scale reconstruction projects. Second, scanner work-flows are optimized for clinical work and not the "batch mode," high-throughput reconstruction typically required in quantitative imaging research, reducing the number of reconstructions that can be accomplished in any allotted time. Third, the clinical setup typically requires manual changes to many parameters prior to each reconstruction, requiring the constant attention of a researcher, and increasing the likelihood of errors. And finally, if a site upgrades a scanner, a researcher may lose the ability to reconstruct "legacy" raw data associated with that scanner model. Therefore, it was identified that there was a need to develop customizable tools that allow efficient, large-scale reconstruction of diagnostic CT images independent of the acquisition scanner.

While there are many available options for reconstructing cone-beam CT data using third-party open source software libraries (e.g. RTK [71], CONRAD [72], OSCaR [73]), the options for reconstructing helical, diagnostic CT data are significantly more limited. The only alternative to on-board reconstruction that could be identified, is a standalone reconstruction computer from manufacturers, but these have limited availability and are not widely deployed. Furthermore, the work-flow of these “recon boxes” is typically ill-suited to large-scale reconstruction and rarely customizable to a researcher's needs. Thus, the standalone reconstruction computer does not represent a satisfactory solution to the problems facing researchers hoping to perform large numbers of reconstructions, and no other alternatives were available.

As a result, FreeCT_wFBP was developed which is a free and open-source implementation of a commonly used reconstruction concept – specifically weighted filtered backprojection (wFBP) – for third-generation, helical, fan-beam CT in an effort to overcome some of the limitations of the currently available tools. The software is highly flexible, with features such as user-configurable scanner geometries, user-modifiable reconstruction kernels, CPU and GPU implementations, and support for data acquired using sampling techniques such as flying focal spots and quarter-detector offsets. FreeCT_wFBP is a command-line program providing flexible and fast reconstruction of helical, diagnostic CT data using a GPU or CPU. It is written in C and utilizes the NVIDIA CUDA framework for GPU-specific code.

2.3.1 The FreeCT Algorithm

FreeCT_wFBP is an implementation of weighted filtered backprojection (wFBP), a widely used approach for helical CT reconstruction that offers a good trade-off between computational effort, accuracy, and flexibility [74]–[77]. While wFBP is a relatively simple reconstruction approach to implement, there are many possible variations such as where and how to handle slice-thickness settings, weighting function choice, and weighting “tuning” parameters. FreeCT_wFBP is, specifically, an implementation of weighted filtered backprojection as described in [75]. It is suitable for reconstructing any third-generation, helical CT data (without gantry tilt), including clinical or simulated raw projection data from any manufacturer or software tool, so long as the user can extract that raw data into an appropriate format readable by FreeCT_wFBP and the scanner geometry is configured properly. A full discussion of implementations details and choices including equations and pseudo-code can be found in the FreeCT_wFBP documentation [78].

FreeCT_wFBP reconstructs helical data from third-generation multi-detector CT scanners, currently the most widely employed geometry in clinical diagnostic CT. Third-generation CT scanners utilize a detector with circular curvature in the axial (XY) plane, and no curvature in the longitudinal (Z) direction. For this work, the Siemens Definition AS 64 was specifically targeted. FreeCT_wFBP does not currently reconstruct axial scans, nor does it reconstruct helical scans acquired with gantry tilt.

2.3.2 Rebinning, Flying Focal Spots, and Quarter Detector Offsets

Weighted filtered backprojection, as described in [75], utilizes a row-wise fan-to-parallel rebinning process prior to filtering and backprojection. While this requires an extra set of interpolations, it has been shown to have negligible effect on image quality [79] while at the same time providing several benefits: (1) simplified geometry for backprojection, (2) artifact reduction (mitigation of cone-beam artifacts) during filtering since the data is recast along the spiral tangent [75], and (3) straight-forward accounting for changes in geometry caused by sampling techniques such as flying focal spots.

Flying focal spots are a technique employed by some CT scanners to improve sampling in the axial and/or longitudinal directions by periodically deflecting the electron beam to different locations on the x-ray tube anode between detector readouts [80], [81]; a depiction of this periodic motion can be found in figure 1 of [81] and figures 1, 2 and 5 of [80]. Use of the in-plane flying focal spot (called the "Phi" flying focal spot) improves in-plane spatial resolution in axial images. Use of the longitudinal flying focal spot (called the "Z" flying focal spot) improves spatial resolution in the longitudinal direction and also reduces windmill artifacts observed in the axial plane near the edges of high-contrast objects [80], [81]. Depending on scan configuration (namely rotation time and collimation), a scanner equipped with flying focal spots may use both Z and Phi, Z only, Phi only or no flying focal spots to acquire data. Incorporation of the flying focal spot rebinning routines allows the software to reconstruct projection data from a much larger subset of scanners than would otherwise be possible.

It is worth remarking that no other freely-available, open-source software supports flying focal spots to the best of our knowledge, however this is a critical feature to support if clinical data is to be reconstructed. With the exception of an extremely limited number of protocols, every clinical acquisition made at UCLA utilizes a flying focal spot; in particular for the lung screening protocol utilized in this work, the Z-only flying focal spot setting is used.

Another sampling technique employed clinically is the quarter- or eighth-detector offset. Similar to the flying focal spot, all regularly employed protocols acquired clinically at UCLA employ a quarter detector offset. The quarter-detector offset (QDO) shifts the detector center by a quarter of the detector width relative to true detector "center" (the point at which a ray traced from the focal spot through isocenter would intersect the detector plane). While the quarter detector offset is most beneficial in axial CT scans (doubled in-plane sampling, improved in-plane resolution, and reduction of in-plane aliasing artifacts) its use in helical CT is also very common [80], [82]. FreeCT_wFBP is capable of reconstructing with and without the QDO, as well as with an eighth-detector offset which occurs when the QDO is used in conjunction with the in-plane flying focal spot [80].

2.3.3 Reconstruction Kernels

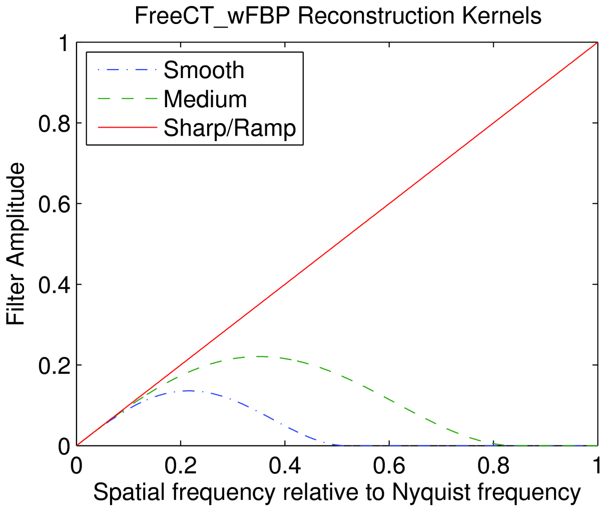


FIGURE 2-4 SMOOTH, MEDIUM AND SHARP/RAMP RECONSTRUCTION KERNELS PROVIDED WITH THE FREECT_WFBP SOFTWARE PACKAGE, PLOTTED IN THE FOURIER/SPATIAL FREQUENCY DOMAIN.

Reconstruction kernels are stored as binary files of single-precision, floating-point data (a vector representing the kernel profile in the spatial domain) and are read in at program runtime making it easy for a user to create and utilize their own filters without needing to recompile source code. Full details on filter creation and installation can be found in the documentation [78]. FreeCT_wFBP comes with three reconstruction kernels ready for use: smooth, medium and sharp/ramp, plotted in the Fourier domain in Figure 2-4. While these filters should not be directly compared to those offered by Siemens, and a rigorous comparison has not been done, noise values in a 32 cm CTDI phantom compare roughly as follows: FreeCT_wFBP's smooth kernel is similar to a B10 or B20 kernel; the FreeCT_wFBP medium kernel has similar noise to a Siemens B40-B45; and finally, the FreeCT_wFBP sharp kernel has approximately the same noise magnitude as a Siemens B50-B60. These comparisons however do not capture or represent noise "texture" which can play a substantial in detection tasks, and they are intended only as a rough guideline for the reader. Sample reconstructions of the ACR phantom utilizing each of the FreeCT_wFBP reconstruction kernels can be found in Figure 2-5, Figure 2-6, and Figure 2-7. These are the reconstruction kernels employed in Chapter 3 and Chapter 4 for the wFBP reconstructions.

2.3.4 GPU and CPU Implementations

Due to the computational demands of CT image reconstruction and the need to automate large numbers of reconstructions in an efficient manner, FreeCT_wFBP is first and foremost a GPU-based software package. To extend the functionality and accessibility of the software, a single-threaded CPU implementation has also been created. The GPU implementation is significantly faster and will be most useful for researchers looking to process large numbers of reconstructions, however the CPU implementation is well suited to running large numbers of reconstructions on distributed clusters that may not have GPUs available. It should be noted that the performance on a single CPU is not expected to be fast enough for large-scale reconstruction projects. The utilization of GPUs to accelerate FreeCT_wFBP additionally motivates the development of the GPU queueing framework discussed later in this chapter.

2.3.5 Evaluation of FreeCT_wFBP - Methods

Image quality and accuracy were evaluated according to the current ACR CT Accreditation Program (CTAP) criteria for CT number evaluation, CT number uniformity, and contrast-to-noise ratio (CNR) using the methods and formulas described in [43]. The ACR phantom was scanned under the “Phi and Z” flying focal spot (FFS) conditions described in Table 2-1, using a routine adult abdomen protocol, and the central slice of each module of the ACR phantom was reconstructed to a thickness of 5mm using the included smooth, medium, and sharp/ramp kernels from the FreeCT_wFBP package. The reconstructed slices of each module were then evaluated to see if they fell within ACR-acceptable ranges.

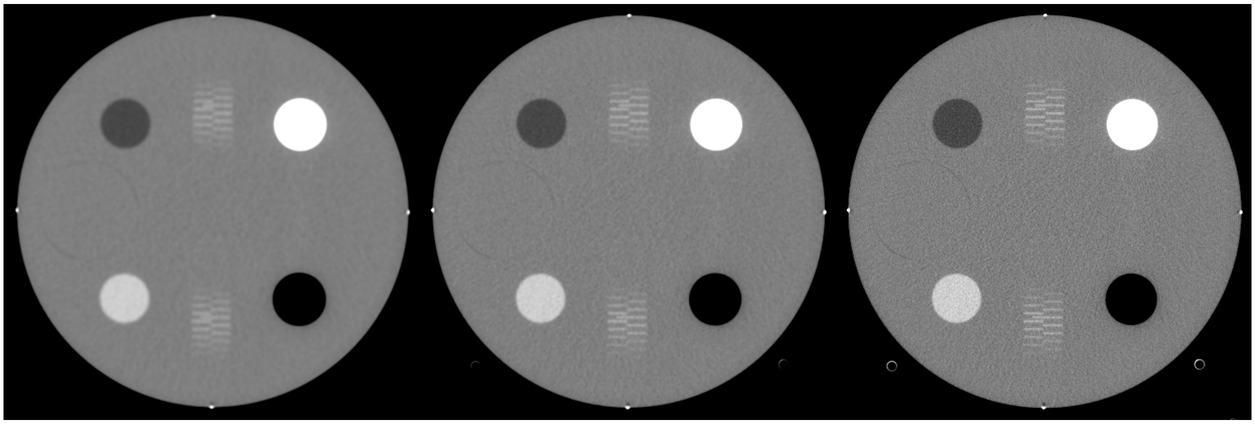


Figure 2-5 CT number module of the ACR phantom. Reconstructed with FreeCT_wFBP's smooth (left), medium (middle), and sharp (right) reconstruction kernels. Shown with a window/level of 400/0 HU.

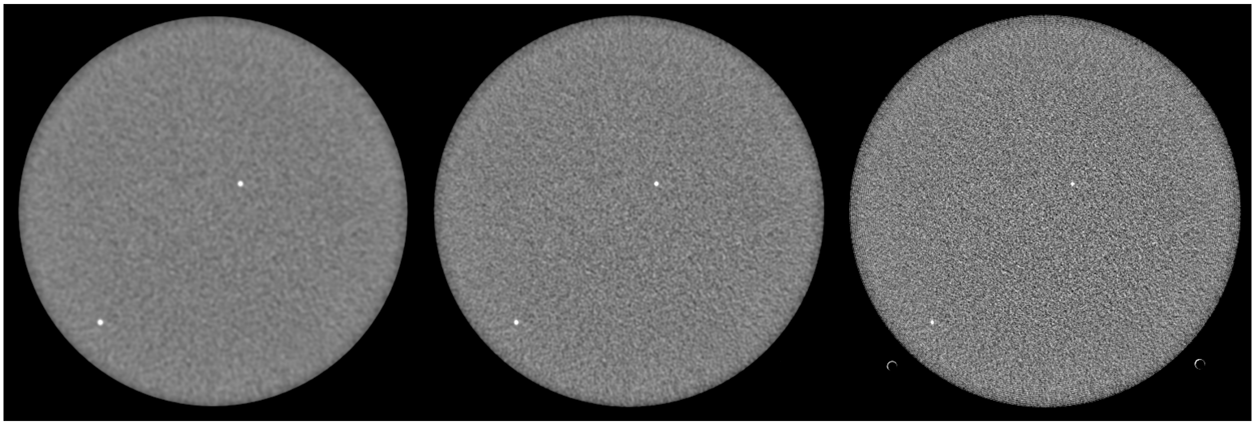


Figure 2-6 Uniformity module of the ACR phantom. Reconstructed with FreeCT_wFBP's smooth (left), medium (middle), and sharp (right) reconstruction kernels. Shown with a window/level of 100/0 HU.

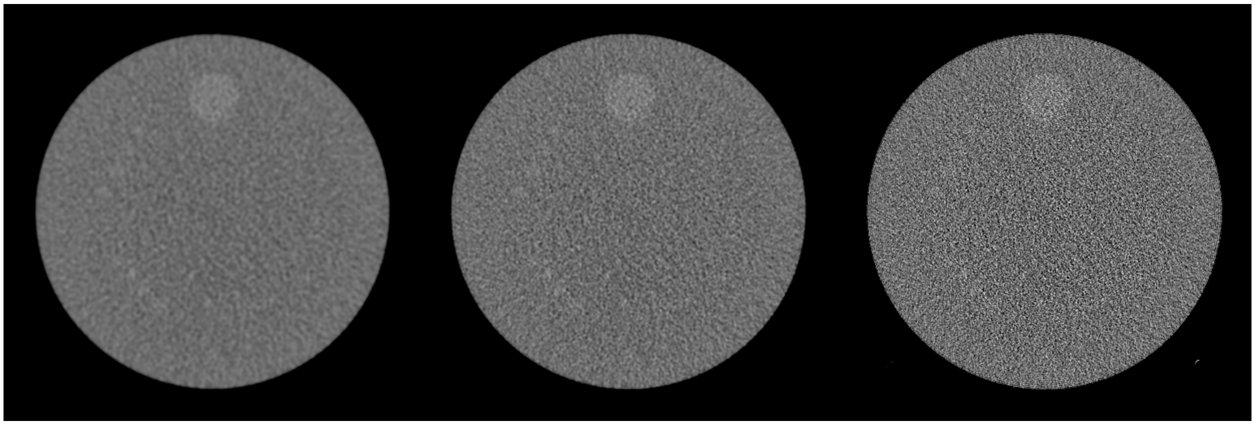


Figure 2-7 Low contrast module of the ACR phantom. Reconstructed with FreeCT_wFBP's smooth (left), medium (middle), and sharp (right) reconstruction kernels. Shown with a window/level of 100/0 HU.

TABLE 2-1 SUMMARY OF SCAN PARAMETERS FOR SPEED PROFILING SCANS. THE NUMBER OF REBINNED PROJECTIONS REQUIRED FOR EACH RECONSTRUCTION ARE SHOWN AND REFLECT THE EFFECTS OF COLLIMATION AND FLYING FOCAL SPOT SETTINGS, WHICH ARE INFLUENCED BY THE RECONSTRUCTED VOXEL SIZE RELATIVE TO THE SIZE OF THE EFFECTIVE DETECTOR THICKNESS (E.G. 16X1.2MM).

FFS	Collimation (mm)	Rotation time (s)	Recon slice thickness (mm)	Rebined projections for 32 slices
No FFS	16 × 1.2	0.5	1.2	3840
Phi-only	16 × 1.2	1	1.2	3712
Z-only	64 × 0.6	0.5	0.6	2624
Phi and Z	64 × 0.6	1	0.6	2528

Pipeline performance was evaluated using timing benchmarks for each step of the reconstruction process for a given set of reconstruction/acquisition conditions. Reconstruction speed (i.e. computational performance) is dependent on many factors, including but not limited to collimation, flying focal spot configuration, and slice thickness; column 5 of Table 2-1 (“Rebinned Projections for 32 Slices”) is included to highlight how collimation and flying focal spot settings in particular can have a large impact on parameters the user does not directly control (i.e. number of projections required to fully reconstruct a volume), but do affect reconstruction speed. FreeCT_wFBP reconstruction speed was evaluated by reconstructing 512x512x32 voxel volumes from scans of the ACR accreditation phantom (Model 464, Gammex, Middleton, WI) performed on a 3rd generation CT scanner (Definition AS 64, Siemens Healthcare, Forchheim, Germany) under all flying focal spot (FFS) combinations available on that scanner: (a) no FFS, (b) Phi FFS only, (c) Z FFS only and (d) both Z and Phi FFS as described in Table 2-1. Slices were reconstructed to thicknesses matching detector collimation. Code profiling for both the CPU and GPU implementations was performed using the NVIDIA Visual Profiler included with the CUDA toolkit.

All reconstruction and evaluation was performed on an Alienware Aurora R4 computer with an Intel i7-4960X CPU (3.6 GHz, 15 MB L3 cache), 32 GB of RAM, and an Nvidia GeForce GTX 780 GPU with 3 GB of global memory. GPU reconstructions were acquired by running the software with the standard settings (auto-detection and use of GPU resources), and CPU reconstructions were acquired using a "--no-gpu" command line option, forcing all reconstruction to take place on the CPU.

While spatial resolution is no longer evaluated as part of the ACR CTAP, the spatial resolution module of the phantom was used to ensure proper implementation of flying focal spot rebinning. A slice through the middle of the spatial resolution module from each of the scans listed in table 1 was reconstructed to a 1.2mm slice thickness and a 100mm field of view centered on the 9 lp/cm bar pattern. A sharp/ramp filter was used to maximize spatial resolution. Each reconstructed image was evaluated for changes in in-plane resolution and changes in windmill artifacts. If implemented correctly, an improvement in in-plane resolution with activation of the Phi flying focal spot, and a reduction of windmill artifacts with the activation of the Z flying focal spot, should be observed.

Finally, reconstruction accuracy was evaluated using a reconstruction of a simulated FORBILD thorax phantom [83] for which attenuation values were known exactly. The FORBILD thorax phantom data was generated with a simulated, 80keV, monochromatic beam without flying focal spots, and reconstructed using the included sharp/ramp kernel. The attenuation value of water at 80keV (mass attenuation coefficient of 0.0183 mm^2/g) was used to create the phantom. Without scaling the reconstructed image to Hounsfield units (HU), an ROI was placed over a region of simulated water and the mean value (mass attenuation coefficient) was compared to the value used in the simulation.

2.3.6 Evaluation of FreeCT_wFBP - Results

Tables Table 2-2 and Table 2-3 summarize the GPU and CPU reconstruction times, respectively. Rebinning and filtering are a hybrid process in the FreeCT_wFBP implementation and thus are combined into one step for timing purposes. Note that GPU reconstruction times are in seconds and CPU reconstruction times are in minutes. Table 2-4, Table 2-5, and Table 2-6 summarize the imaging performance of the smooth, medium, and sharp reconstruction kernels using the ACR testing protocols. Figure 2-5, Figure 2-6, and Figure 2-7 show the reconstructed slices that were used for evaluation, all windowed and leveled to the values recommended in the ACR testing protocols.

Table 2-2 Sample speed results for GPU reconstruction for different flying focal spot (FFS) configurations. Note that times are given in seconds.

	GPU		
FFS	Rebin and filter (s)	Backprojection (s)	Total (s)
No FFS	2.7	13.8	16.5
Phi FFS	2.9	13.8	16.7
Z FFS	7.9	9.9	17.8
Phi and *Z* FFS	8.8	9.9	18.7

Table 2-3 Sample speed results for CPU reconstruction for different flying focal spot (FFS) configurations. Note that times are given in minutes.

	CPU		
FFS	Rebin and filter (min)	Backprojection (min)	Total (min)
No FFS	0.5	93.5	94.0
Phi FFS	3.0	93.4	96.4
Z FFS	4.6	64.8	69.4
Phi and *Z* FFS	8.2	64.9	73.1

Table 2-4 Summary of CT number performance for each reconstruction kernel provided with FreeCT_wFBP. All values are within acceptable ACR ranges. All values are in HU.

	Polyethylene	Bone	Acrylic	Air	Water
Smooth	−88.3	869.3	123.6	−989.1	−0.4
Medium	−88.3	869.0	123.5	−988.7	−0.4
Sharp	−88.1	866.6	123.4	−988.1	−0.3
Acceptable range	−107 to −84	850 to 970	110 to 135	−1005 to −970	−7 to 7

Table 2-5 Summary of uniformity measurements for each reconstruction kernel. All values are in HU and well within the -5 to +5 HU range specified by the ACR.

	Uniformity (maximum difference from center)
Smooth	−0.3
Medium	0.6
Sharp	0.9
Acceptable range	−5 to 5

Table 2-6 Summary of CNR values for FreeCT_wFBP. The smooth and medium reconstructions are well above the ACR limit of 1.0.

	CNR
Smooth	2.6
Medium	1.3
Sharp	0.4
Acceptable range	>1.0

The effects of flying focal spot usage on reconstruction quality are shown in figures Figure 2-8 and Figure 2-9. In Figure 2-8, an improvement in spatial resolution is observed when the Phi flying focal spot is utilized (Figure 2-8, (b) and (d)) allowing the 9 lp/cm bar pattern to be clearly resolved. When the Phi flying focal spot is not used (Figure 2-8, (a) and (c)), the 9 lp/cm bar pattern can no longer be precisely resolved.

In Figure 2-9, the image is windowed and leveled to highlight the impacts of Z flying focal spot usage. Usage of the Z flying focal spot manifests itself in the axial plane as a reduction of windmill artifacts. In Figure 2-9, a high contrast bead produces windmill artifacts when the Z flying focal spot is not utilized (top row, highlighted with the larger red arrow), which then disappear with the activation of the Z flying focal spot (highlighted with the smaller, black arrows). The higher frequency noise in the right column is due to the increased in-plane resolution with the Phi flying focal spot combined with the ramp filter reconstruction.

Finally, Figure 2-10 shows the unscaled axial, sagittal, and coronal FreeCT_wFBP reconstruction of the FORBILD thorax phantom for which all attenuation values are known exactly. The ROI's mean value of 0.0183 mm^2/g agrees to within 0.0001 (0.5%) of the value used to simulate the data.

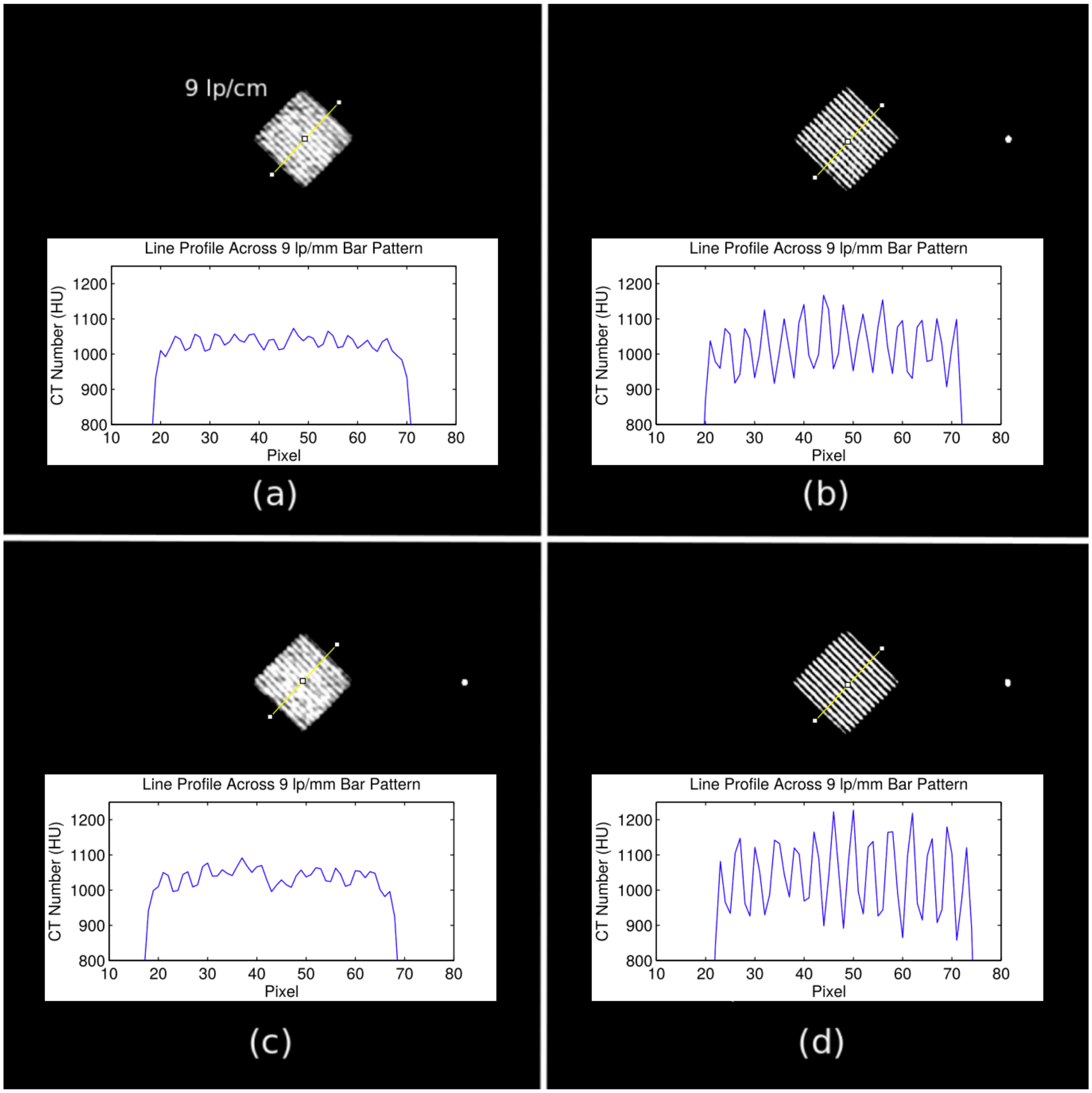


FIGURE 2-8 SPATIAL RESOLUTION RECONSTRUCTIONS USING (A) NO FLYING FOCAL SPOTS, (B) PHI FLYING FOCAL SPOT, (C) Z FLYING FOCAL SPOT, AND (D) Z AND PHI FLYING FOCAL SPOT. IN (B) AND (D) THE PHI FLYING FOCAL SPOT IS ACTIVE AND SPATIAL RESOLUTION IS QUALITATIVELY IMPROVED AND LINE PROFILES ACROSS THE BAR PATTERN FURTHER INDICATE THE IMPROVED ABILITY TO DISTINGUISH FINE DETAIL. ALL IMAGES ARE OF THE 9 LP/CM BAR PATTERN OF THE ACR PHANTOM SPATIAL RESOLUTION MODULE. SHOWN WITH A WINDOW/LEVEL OF 100/1000 HU.

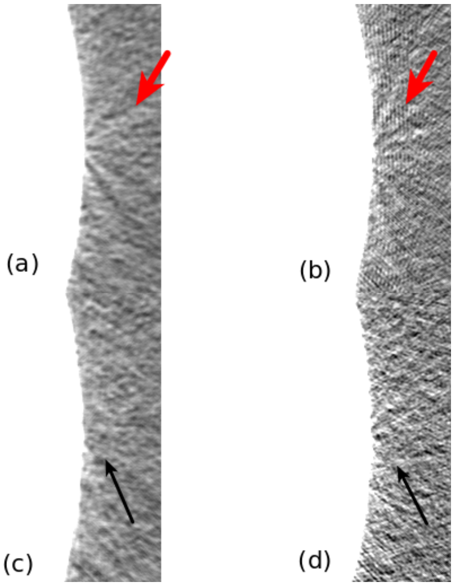


FIGURE 2-9 Z FLYING FOCAL SPOT COMPARISON USING (A) NO FLYING FOCAL SPOTS, (B) PHI FLYING FOCAL SPOT, (C) Z FLYING FOCAL SPOT, AND (D) Z AND PHI FLYING FOCAL SPOT. LARGE, RED ARROWS HIGHLIGHT WINDMILL ARTIFACTS OFF OF A HIGH CONTRAST CENTERING BEAD, WHILE SMALLER BLACK ARROWS HIGHLIGHT THEIR ABSENCE IN (C) AND (D) WHEN THE Z FLYING FOCAL SPOT IS ACTIVE. SHOWN WITH A WINDOW/LEVEL OF 125/-1000 HU. HIGH FREQUENCY ARTIFACTS (MOST PRONOUNCED IN TOP RIGHT IMAGE) ARE DUE TO THE RAMP FILTER RECONSTRUCTION COMBINED WITH THE HIGH, IN-PLANE RESOLUTION OFFERED WITH THE PHI FLYING FOCAL SPOT; NOTE THEIR DISAPPEARANCE WHEN THE PHI AND Z FLYING FOCAL SPOTS ARE UTILIZED TOGETHER.

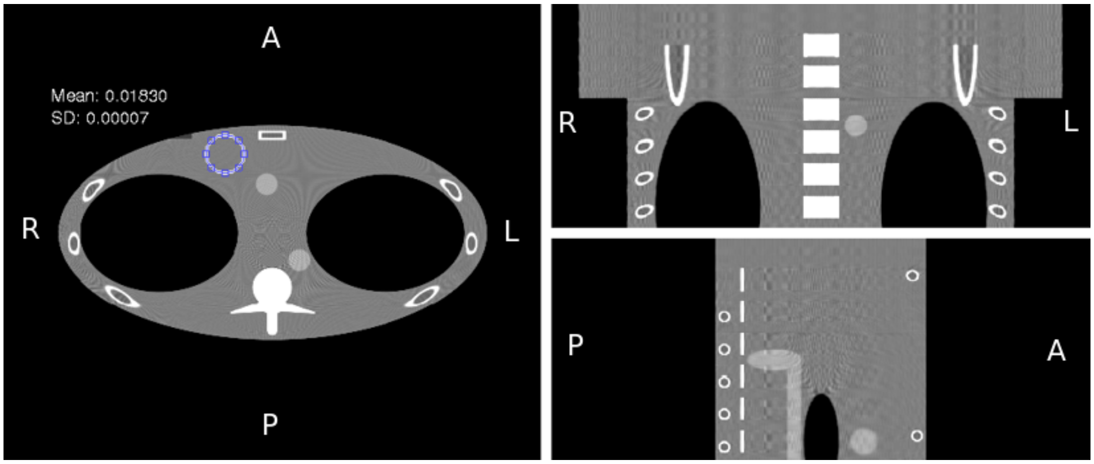


FIGURE 2-10 SAMPLE AXIAL (LEFT), CORONAL (TOP RIGHT), AND SAGITTAL (BOTTOM RIGHT) RECONSTRUCTIONS OF A SIMULATED FORBILD THORAX PHANTOM (MONOCHROMATIC, 80 KEV BEAM ENERGY) ARE SHOWN. AN ELLIPTICAL ROI IS PLACED OF A REGION OF SIMULATED WATER SHOWING A RECONSTRUCTED ATTENUATION VALUE OF 0.0183 MM^2/G. SHOWN WITH A WINDOW/LEVEL OF 0.005/0.0183 MM^2/G, WHICH CORRESPONDS TO A WINDOW/LEVEL OF APPROXIMATELY 272/0 HU. RESIDUAL ALIASING ARTIFACTS CAUSED BY HIGH-FREQUENCY COMPONENTS IN THE NOISE-FREE DATA ARE VISIBLE DUE TO THE USE OF A RAMP KERNEL.

2.3.7 Discussion of FreeCT_wFBP

FreeCT_wFBP not only offers software dedicated diagnostic CT reconstruction, but does so in a highly flexible and configurable package capable of handling complex imaging setups including detector offsets, flying focal spots, and multiple scanner geometries. While this package may not represent the exact reconstruction algorithms employed by clinical CT scanners, an initial assessment demonstrated that FreeCT_wFBP can provide acceptable performance on the ACR phantom as well as accurate reconstruction of attenuation values. FreeCT_wFBP's GPU implementation allows for fast reconstruction of clinical CT data and is well suited to large-scale explorations of reconstruction parameter space.

To provide an initial verification that the software was yielding acceptable results, reconstructions of the ACR phantom were performed and analyzed. These reconstructions (Figure 2-5, Figure 2-6, Figure 2-7) were performed using the smooth and medium reconstruction kernels and met or exceeded all of the ACR accreditation standards (Table 2-4, Table 2-5, Table 2-6). The CNR of the sharp/ramp kernel did not pass the adult abdomen standard (>1.0), however ramp kernels are not used clinically due to the fact that they over-enhance high-frequency noise. The sharp/ramp kernel met all other ACR accreditation standards, and provided the highest spatial resolution. Using a reconstruction of a simulated scan (80keV monochromatic beam) of a FORBILD thorax phantom (Figure 2-10), it was shown that the software reconstructs accurate attenuation values. Therefore, while FreeCT_wFBP may not represent the exact reconstruction algorithms employed in clinical scanners, ACR-acceptable performance, and accurate reconstruction of physical attenuation values in a known phantom indicate its readiness for use in scientific research, such as [84] where the software was used to perform 10,000 volumetric reconstructions to assess lesion detectability in lung screening scans.

FreeCT_wFBP is utilized extensively in Chapter 3 and Chapter 4 for all weighted filtered backprojection images. The results presented demonstrate the correct implementation of clinically important details such as flying focal spots and accurate attenuation value reconstruction, critical for the use of FreeCT_wFBP to perform reconstructions usable for quantitative imaging.

2.4 FreeCT_ICD

FreeCT_ICD represents the fully-3D, model-based iterative complement to FreeCT_wFBP, enabling fully-automated, offline reconstruction of clinical, 3rd generation, helical CT datasets. While not directly utilized in the quantitative evaluation experiments presented here[1], it is an important extension of the pipeline to address modern CT technology. Here, we provide an overview of the algorithm implementation details and some sample results for completeness.

2.4.1 Introduction

The purpose of the broader FreeCT project is to provide a set of tools for offline reconstruction of raw projection data (i.e. sinogram data) for CT imaging research. These tools have enabled the development of the reconstruction pipeline that is not dependent on the availability of clinical CT scanners, can be configured to operate in high throughput

[1] FreeCT_ICD was not utilized here since the scanner-based SAFIRE reconstruction algorithm was utilized. Additionally, FreeCT_ICD is computationally intensive and run times are currently much longer than FreeCT_wFBP. Future work will concentrate on accelerating FreeCT_ICD.

batch modes, can incorporate simulated dose reduction techniques [54], [58] and therefore can produce a large collection of image datasets that represent a wide range of acquisition and reconstruction settings such as different slice thicknesses and reconstruction kernels used in wFBP. The results of the reconstruction pipeline has contributed to the growing list of investigations evaluating the robustness of quantitative imaging, radiomics and CAD methods across a range of scanner platforms, acquisition conditions and reconstruction parameters [54]–[56], [64], [65], [85], [86].

FreeCT_wFBP represents an important contribution to this work, since all modern scanners still typically offer a version of filtered backprojection reconstruction and many protocols still utilize it exclusively (e.g. [13], [42]). Modern scanners however offer advanced image reconstruction techniques that use some form of statistical or iterative reconstruction. These advanced image reconstruction methods are an important clinical technique for reducing radiation doses in CT, but require further investigation for their effects on quantitative imaging, radiomics and CAD performance. Model-based iterative reconstruction offers the potential for substantial radiation dose reduction [87], but comes with a challenging computational burden. Part of this burden lies in the size of the system matrix, which can be 1000 times larger than system memory for a typical CT scan. While open-source reconstruction packages now exist for FBP, to the authors' knowledge there are no open-source, freely-available packages that can directly reconstruct clinical datasets using model-based iterative reconstruction methods. The aim of this work is to fill this gap with software, FreeCT_ICD, that can provide this capability. A complementary initiative is underway to provide the community with freely available raw data from clinical scanners [57].

2.4.2 Algorithm Description and Software Features

System Matrix Definition

FreeCT_ICD employs a stored system matrix, in contrast to standard approaches, which avoid storing the system matrix by focusing on the evaluation of matrix-vector products on the fly. This "on-the-fly" approach however limits the choice of system matrix (or forward projection model) to that which can be quickly computed. Model-based iterative reconstruction depends on the accuracy of the CT system model, and more detailed models of the x-ray source or detector responses may lead to improved resolution and image quality. Storing the system matrix offers the potential for modeling these higher-order effects, obviating the need to re-calculate them on the fly at each iteration, at high computational expense. The trade-off for this approach however is extremely high memory requirements, which are often infeasible to store in a computer's memory, or much slower traversing of full-size matrix stored to disk. To make the size of this matrix practical, we have adopted the concept of Xu et al. [88], which exploits the helical symmetry through the use of rotating slices, resulting in a system matrix size that is substantially smaller, can be kept in memory on high-performance systems, and is functional and fast on lower-memory systems that require it to be saved to disk.

Within this concept, there exist many ways to define the system matrix elements. We have used a method based on Joseph's method [89] that can be viewed as a natural 3D extension of the bilinear method employed by Hahn et al. [90] however with the refinement of reducing bilinear interpolating to Joseph's method. This yields similar image quality with smaller system matrix sizes. Our system matrix approach provides results that are very similar to those one might expect using Joseph's method in 3D on a conventional Cartesian grid. We have verified this aspect using computer simulation with the FORBILD head phantom (for brevity, this is not reported here). We expect the blob approach of Xu et al. [88] can provide results with fewer discretization errors, particularly in the absence of a regularizer, but this advantage comes with a much higher memory requirement. Compared with the intersection length based approach of Guo and Gao [91], the opposite effect is expected: fewer discretization errors at the cost of an increase in memory requirement. We did not consider employing a Siddon-based approach as our experience in 2D fan-beam tomography is that Siddon's approach is suboptimal for practical CT geometries.

Implementation

The program includes a penalty term in the objective function as a regularizer, with two choices for the potential function: quadratic or Fair (edge-preserving) potential. In the quadratic case, the single coordinate optimization problem is solved analytically; in the Fair potential case, it is solved via the bisection method. The program sequentially iterates along the axial direction first, followed by the transaxial direction, so that the elements of the stored system matrix need only be accessed once per iteration. Eight transaxial neighbors are used to calculate the penalty term. Iterative coordinate descent does not lend itself easily to GPU parallelization, so the system matrix calculations and iterations are performed on a normal desktop CPU architecture. However, individual iterations are accelerated with multi-core CPU OpenMP libraries [92], which produces up to a factor of 5 speed-up.

One key offering of FreeCT_ICD is that it can be initialized from a filtered backprojection reconstruction using FreeCT_wFBP, which dramatically reduces the number of iterations required to achieve a converged solution. To take advantage of this, users will have to install FreeCT_wFBP and have a suitable GPU. For the requirements of FreeCT_wFBP, readers are referred to the FreeCT_wFBP technical note [93] and the FreeCT_wFBP documentation [78].

FreeCT_ICD is coded in C++ and was developed on Linux (Ubuntu 14.04LTS, Canonical, Ltd, London, UK) and should compile and run on all modern Linux distributions with little to no modification. Only two major external dependencies for building and running the software are required: (1) the Boost uBLAS C++ library (http://www.boost.org) and (2) the "yaml-cpp" library (https://github.com/jbeder/yaml-cpp). The Boost libraries come preinstalled on most Linux systems and/or are easily available through the distribution's package manager along with the YAML-cpp library.

2.4.3 Sample Results

In this section, we report on reconstructions of clinical datasets that were carried out to evaluate image quality. The datasets used raw projection (sinogram) data acquired on a clinical scanner (Definition AS, Siemens Healthineers, Forchheim, Germany). The latter involved both the phantom from the American College of Radiology (ACR) CT Accreditation Program and a pediatric thoracic scan. In both cases, the scans were performed on the clinical scanner, the raw projection data was collected from the scanner and then image data was reconstructed using FreeCT_ICD. These are described below.

ACR CT Accreditation Phantom

The ACR CT accreditation phantom was scanned on the clinical scanner using a helical scan protocol with acquisition parameters described in Table 2-7. The raw data was captured from the scanner and reconstructed using both wFBP and the ICD algorithm using wFBP initialization. The reconstruction parameters used are also described in Table 2-7. Figure 2-11 shows images through the reconstructed ACR phantom from the ICD reconstruction.

TABLE 2-7 ACQUISITION AND RECONSTRUCTION PARAMETERS FOR BOTH THE ACR PHANTOM SCAN AND THE PEDIATRIC THORACIC SCAN (INCLUDING PENALTY TERM AND EDGE-PRESERVING PARAMETERS).

Scan	ACR Phantom	Pediatric Chest
Acquisition Parameters		
Tube voltage [kV]	120	100
CareDose4D	Off	On
Quality Reference mAs	---	180
Effective mAs	100	73
Collimation	16 x 1.2 mm	16 x 1.2 mm
Flying focal spot	Off	Off
Rotation time [s]	0.33	0.33
Reconstruction (ICD) Parameters	wFBP initialization	wFBP initialization
Voxel grid dimensions	512 x 512 x 132	512 x 512 x 163
Voxel size [mm]	0.58 x 0.58 x 1.5	0.98 x 0.98 x 1.5
FOV radius [mm]	300	500
Edge-preserving parameter	0.005	0.005
Penalty term parameter	0.1	0.1
Matrix size [GB]	8.5	14.6
Iterations	50	50

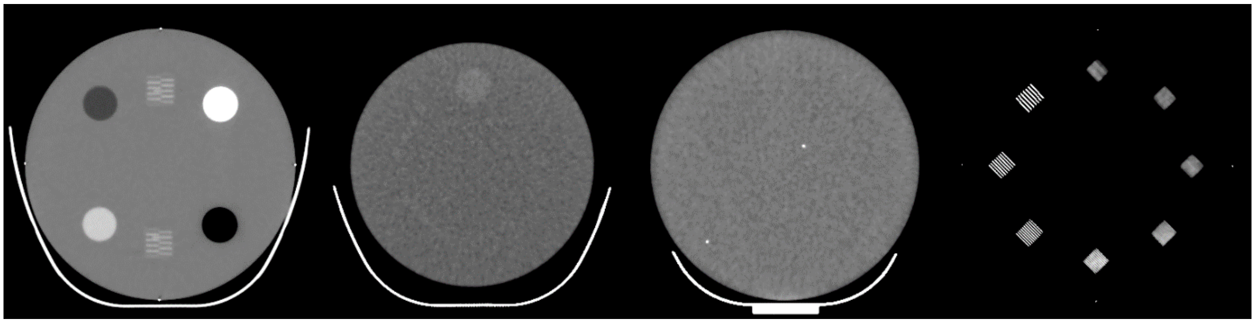


FIGURE 2-11 IMAGES FROM THE FREECT_ICD RECONSTRUCTIONS OF THE ACR ACCREDITATION PHANTOM USING THE ACQUISITION AND RECONSTRUCTION PARAMETERS DESCRIBED IN TABLE 2-7 WITH THE EDGE PRESERVING PENALTY FUNCTION. THIS FIGURE SHOWS FROM LEFT TO RIGHT: THE CT NUMBER MODULE; THE LOW CONTRAST MODULE; THE UNIFORMITY MODULE AND THE SPATIAL RESOLUTION MODULE. EACH IMAGE HAS BEEN WINDOWED AND LEVELED TO THE VALUES RECOMMENDED BY THE ACR [94].

Using the image shown in Figure 2-11, the CT number of all materials were evaluated according to ACR CT Accreditation Program instructions[94]. The results are shown in Table 2-8. All reconstructed CT number values were within the acceptable ranges as defined in the accreditation instructions.

TABLE 2-8 RESULTS FROM CT NUMBER EVALUATIONS OF THE ACR CT ACCREDITATION PHANTOM SHOWN IN FIGURE 2-11.

Material	Acceptable range [HU]	Reconstructed value [HU]
Polyethylene	-107 to -84	-89
Bone	850 to 970	864
Water	-7 to 7	-2
Acrylic	110 to 135	123
Air	-1005 to -970	-988

The low-contrast module reconstruction gave a CNR of 3.83, primarily due to a very low standard deviation value of 1.73. For adult abdomen protocols, the accreditation program guidelines specify that the CNR should be > 1.0, so this value is acceptable (it should be noted that there is no CNR specification for a routine chest protocol). It should be noted that the CNR is affected by the reconstruction kernel in FBP, and by a number of parameters in the iterative reconstruction algorithm. Depending on the selection of these parameters, the apparent CNR can be increased or decreased.

For the uniformity module, the maximum difference from center was 1.1 HU, indicating acceptable uniformity in the reconstructed image. The ACR specifies a range of +/- 5 HU as acceptable.

For the resolution module, the image in Figure 2-12 indicates a resolution of 8 lp/cm was achieved with our reconstruction. The ACR no longer requires this evaluation, but does require resolution evaluation as part of annual QC testing. While there is no limiting resolution value stated for adult head protocols, the adult abdomen protocol limiting resolution is 6 lp/cm. Therefore, this resolution can be judged to be acceptable.

To emphasize the differences between the conventional wFBP reconstruction (which served as the initial condition to the ICD) and the ICD reconstruction, Figure 2-12 shows some additional images of module 4 which evaluates spatial resolution. This figure shows the resolution section reconstructed from both wFBP with a smooth filter as well as with the ICD algorithm with the edge-preserving penalty term. These images are shown at both the window and level the ACR recommends for evaluation of the bar patterns (approx. L=1100/W=100) as well as a window/level setting that is closer to a soft tissue window, which allows us to evaluate both the noise level (evaluated through standard deviations in each image) as well as the reduced streaking artifacts observed in the ICD image. In addition, a difference image is provided to demonstrate the improved resolution that ICD provides. Thus, these images demonstrate that ICD is indeed providing improved resolution at slightly reduced noise compared to wFBP. These images are meant to be illustrative and not definitive of the advantages of ICD over wFBP in all cases; it is recognized that the parameters selected (including those for wFBP) will have significant bearing on any comparisons of resolution, noise and image quality in general.

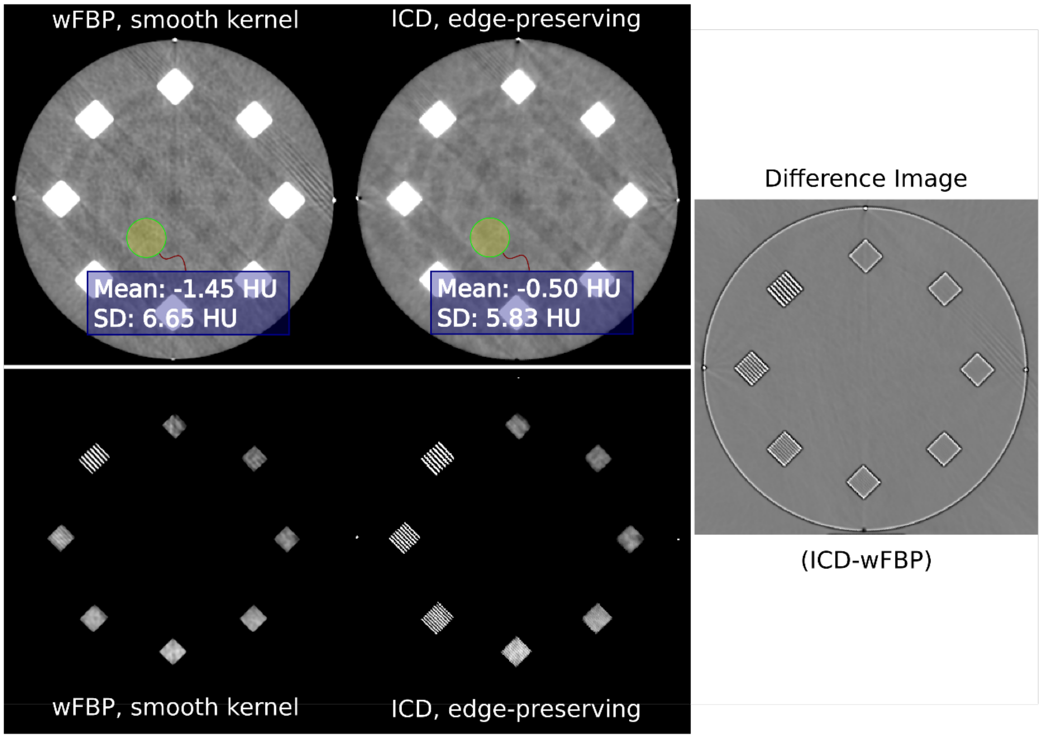


FIGURE 2-12 MODULE 4 (RESOLUTION MODULE) OF THE ACR PHANTOM RECONSTRUCTED WITH BOTH WFBP (SMOOTH KERNEL) AND ICD (EDGE PRESERVING PENALTY TERM). THE TOP ROW DISPLAYS THE IMAGES AT THE WINDOW AND LEVEL SETTINGS SIMILAR TO SOFT TISSUE WINDOWS TO DEMONSTRATE THE REDUCED STREAK ARTIFACTS AND SLIGHTLY REDUCED STANDARD DEVIATION RESULTING FROM ICD RECONSTRUCTIONS. THE BOTTOM ROW DISPLAYS THE IMAGES AT THE WINDOW AND LEVEL SETTINGS SUGGESTED BY THE ACR TO EVALUATE IMAGES FROM THIS MODULE[94] AND SHOWS THE IMPROVED RESOLUTION PROVIDED BY ICD (EVEN AT REDUCED NOISE LEVEL). THE IMAGE ON THE RIGHT IS A DIFFERENCE IMAGE WHICH CLEARLY DEMONSTRATES THE IMPROVED RESOLUTION FROM ICD IN THIS CASE.

Initialization with Weighted Filtered Backprojection

One of the key contributions of FreeCT_ICD is that the iterative reconstructions can be initilized using automatically-configured weighted filtered backprojection reconstructions from FreeCT_wFBP. While the final reconstruction provided is the same, the number of required iterations is substantially fewer when initialized with FreeCT_wFBP. This is demonstrated in

Figure 2-13. Initializing the reconstruction volume with the wFBP scan tends to achieve convergence after approximately 25 iterations, while initializing from an empty volume requires between 50 to 100. This amounts to an hours-long reduction in computing time for one reconstruction, making FreeCT_ICD substantially more viable for large scale investigations.

Pediatric Thoracic Scan

A clinically indicated thoracic scan was performed on a pediatric (7-year-old) patient on the same multidetector CT (Definition AS 64, Siemens Healthineers, Forchheim, Germany). The raw projection data was obtained and anonymized under IRB approval at our institution. Our pediatric chest scans are performed with very low doses (CTDIvol for the 32cm phantom of for this scan was 2.5 mGy). Figure 2-14 represents a coronal image reconstructed from this pediatric thoracic scan using conventional wFBP (with a smooth reconstruction filter) as well as ICD using first a quadratic penalty term and then ICD using an edge preserving penalty term. In both ICD cases, the wFBP was used as the initialization. However, each ICD image can be shown to provide more detail than the wFBP (with smooth kernel) as evidenced by the clearer representation of fine details such as fissures and vascular markings. In this figure, the quadratic penalty term results in noisier images (higher standard deviation) than the edge preserving penalty term, although in general the resolution and noise characteristics depend on the specific parameters used in the regularizer.

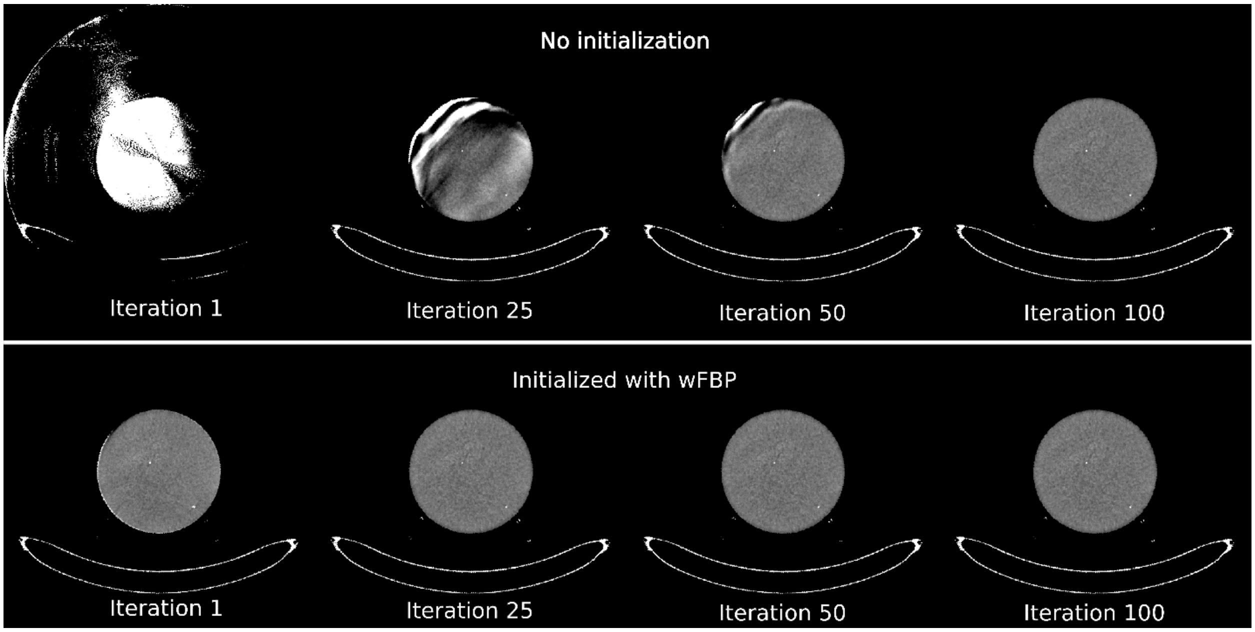


FIGURE 2-13 FIGURE DEMONSTRATING THE EFFECTS OF INITIALIZATION USING WFBP IMAGE DATA USING THE ACR UNIFORMITY MODULE (MODULE 3). THE TOP ROW SHOWS RECONSTRUCTIONS OF THIS MODULE WHEN NO INITIALIZATION IS USED. THE BOTTOM ROW SHOWS RECONSTRUCTIONS OF THIS MODULE WHEN THE WFBP IMAGE DATA IS USED AS THE INITIAL CONDITION AND HOW MUCH FASTER THE IMAGE CONVERGES TO THE EXPECTED ANSWER.

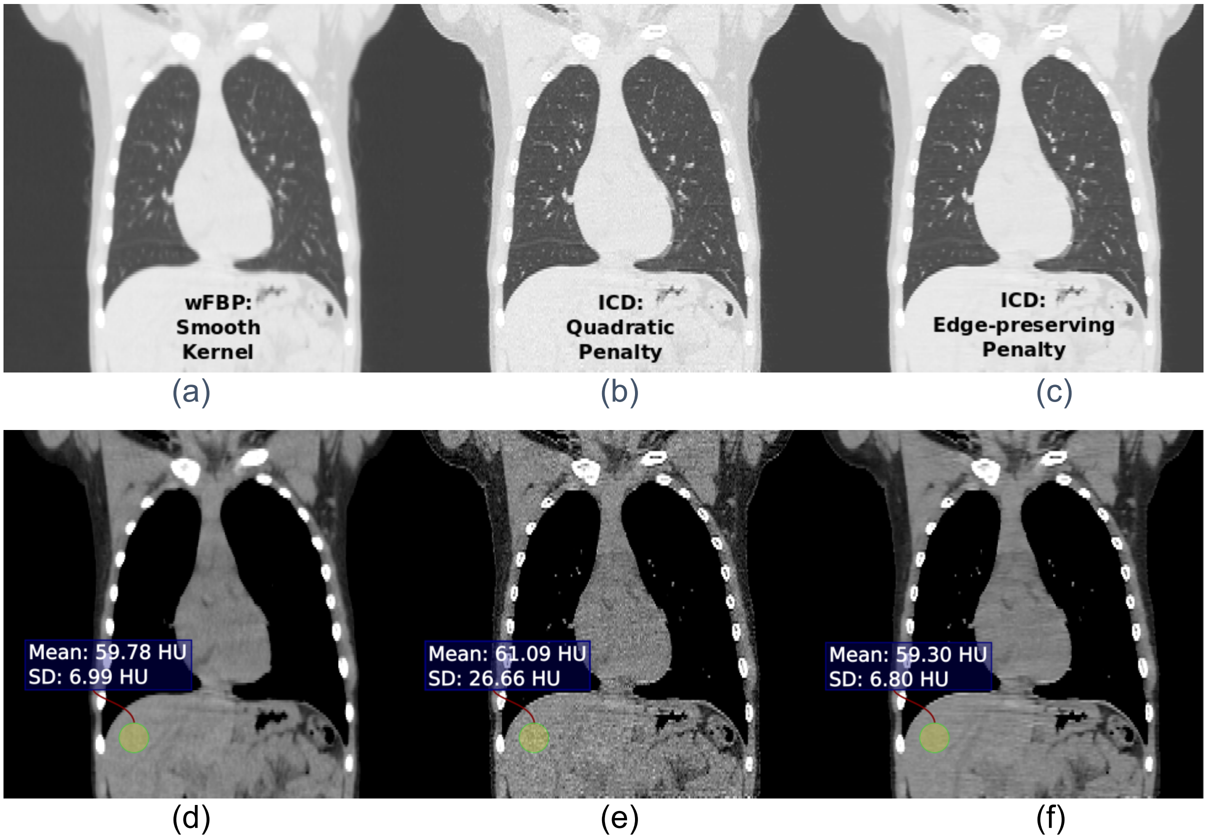


Figure 2-14 Coronal reformat image of a pediatric thoracic CT exam from the same raw projection data to illustrate the differences in reconstructions. The top row shows images displayed at lung windows for: (a) wFBP using a smooth reconstruction kernel, (b) ICD using a quadratic penalty term and (c) ICD using an edge preserving penalty term. The bottom row shows the same images but displayed at soft tissue windows and with a region of interest within a homogeneous area in the liver which demonstrates the similarity in mean values across reconstructions as well as differences in standard deviation values across reconstructions. The order of images is the same as the row above: (d) wFBP using a smooth reconstruction kernel, (e) ICD using a quadratic penalty term and (f) ICD using an edge preserving penalty term.

2.4.4 Discussion

FreeCT_ICD is model-based iterative reconstruction software for helical CT images that uses an iterative coordinate descent approach. This method represents a reasonable tradeoff in computation time and memory requirements for practical implementation of off-line reconstructions. This tool was designed to facilitate CT imaging research such as investigations into the effects of radiation dose reduction and reconstruction method and parameter selection on CT image quality, quantitative imaging and CAD performance. The offline (i.e. away from the clinical scanner) capabilities provided, coupled with standard representation formats for raw projection (sinogram) data [57], may provide advantages in terms of the breadth and depth of investigations that can be performed. This tool was intended as a complement to the weighted filtered backprojection tool already developed and made available [93] via the FreeCT website. It is hoped FreeCT_ICD be a useful addition for the medical physics community and the broader research community.

In the context of this dissertation, FreeCT_ICD represents a key extension of the pipeline infrastructure allowing it to address a broader range of clinical reconstruction configurations. By offering model-based iterative reconstruction (MBIR) in addition to weighted filtered backprojection (wFBP), a more thorough investigation of CT parameter space can be conducted; additionally, because both FreeCT_wFBP and FreeCT_ICD are released under the GNU General Public License version 2.0, the software can be audited in detail, tuned to more closely match clinically used version of wFBP and MBIR if desired. While neither tool is exactly the reconstruction done on the clinical scanners, both demonstrate clinically-similar performance, and pass the ACR CT accreditation standards, and thus represent a very reasonable choice for use in research. Future investigations will focus on accelerating FreeCT_ICD making it more viable for large scale investigations such as those performed in Chapter 3 and Chapter 4 using FreeCT_wFBP.

FreeCT_ICD differs from other iterative reconstruction approaches in that it stores the system matrix directly. In order to fit within memory, it is necessary to employ a rotating grid. The combination of stored system matrix and rotating grid has also been analyzed and studied in [88]. Our work differs from that work in that it has been verified with experimental and clinical data, uses ICD for optimization rather than the alternating direction method of multipliers (ADMM), uses a different representation of the system matrix, and will be released as an open-source, free package. In our implementation, we have mostly used a CPU approach to make the software available to a wider community. The final stored system matrix size is strongly influenced by reconstruction and acquisition parameters (e.g. collimation, reconstructed field of view, pitch, etc.). System matrix sizes for this work fell roughly between 10GB and 20GB using the modified Joseph's method described above. Taking into account the reconstruction and acquisition parameters, our matrix sizes were larger than those achieved by Guo et al. [91] (roughly 1-10GB, Siddon-based), however smaller than those achieved by Xu et al. [88] (roughly 27GB, Blob-based approach). Based on the size of our stored system matrix and the recent analysis provided by Matenine et al. [95], an efficient GPU refinement of our code may be possible in the near future for high end GPU cards. This warrants further investigation and development.

FreeCT_ICD at release is one of only two known open-source, free software packages for CT reconstruction that explicitly supports flying focal spots (the other being FreeCT_wFBP). While full support is still under development, FreeCT_ICD supports the in-plane or "phi" flying focal spot with support for the Z flying focal spot due in a future release. This dramatically extends the immediate usability of the software package since many clinical scans are acquired using a flying focal spot.

These investigations provide a basis for continuing work including improvements in both computational performance as well as image quality improvement. Specific future developments will include investigations into the utility of extending the regularization into the third (longitudinal or "z") dimension, which may include incorporating a longitudinal direction penalty term as well as ensuring that interpolated values are aligned in the longitudinal direction.

2.5 The Pipeline GPU Queuing Framework

While FreeCT_wFBP and FreeCT_ICD represent pathways to accomplish individual reconstructions away from the scanner, an approach was needed to systematically sample a wide range of reconstruction and acquisition parameters with minimal user intervention or input. In addition, to achieve high-throughput reconstruction, a custom GPU framework was developed. While Figure 2-3 illustrates *what* is happening in each stage of the pipeline, the following subsections cover *how* the primary components of the computing framework achieve this. The primary components of the framework are: the "launcher" which starts the pipeline; the "daemon", which dispatches individual jobs and ensures system resources are optimally utilized; and the "queue item" script, which processes an individual reconstruction from start to finish. A schematic flowchart view of the components discussed below can be found in Figure 2-15.

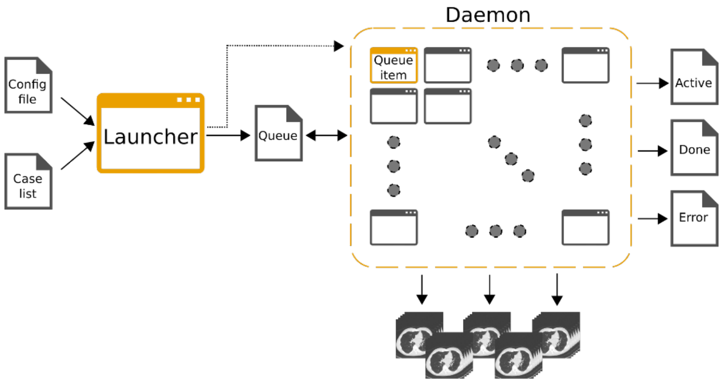


FIGURE 2-15 FLOW DIAGRAM OF PIPELINE RECONSTRUCTION COMPUTATION. PRIMARY PROGRAMMATIC COMPONENTS ARE HIGHLIGHTED IN ORANGE (I.E. THE LAUNCHER, DAEMON, AND QUEUE ITEM). SUPPORTING FILES (E.G. CONFIG FILE, CASE LIST, QUEUE, ETC.) ARE NOT PROGRAMMATIC COMPONENTS THEMSELVES, HOWEVER ARE CRITICAL TO THE PROPER FUNCTIONING OF THE PIPELINE.

2.5.1 Launcher and Configuration Files

The launcher script is the first of the three primary programmatic components and serves to parse a simple configuration file into a job-list, update the pipeline queue, and subsequently start full pipeline execution managed by the “daemon.” After the pipeline launcher has started the daemon, the launcher exits and is not utilized further for a given set of cases. After the work of Young et al. in 2017 [55], it was recognized that the largest expenditure of researcher time was in the manual configuration of reconstructions on the scanner, requiring constant human intervention and attention. The launcher script completely eliminates this requirement, automatically configuring an arbitrary number of reconstructions in seconds based only on the simple input configuration file.

Configuration files are written in YAML (http://yaml.org), a simple, human-readable "data serialization" format that is well supported across many platforms and programming languages, in particular Python. A sample configuration file is given in Listing 2-1. Users can request any number of doses, N_{dose}, and any number of slice thicknesses, $N_{s.t.}$, to explore; the pipeline is capable of assessing any number of reconstruction kernels (N_{kern}), however is limited to the offerings of the reconstruction software. The three used in this work are currently the only three offered with FreeCT_wFBP. In the pipeline's present implementation, the total number of reconstructions per patient will thus be $N_{dose} * N_{s.t.} * N_{kern}$.

```
library:           /data/DefAS_Full/library
case_list:         /data/DefAS_Full/case_list.txt
doses:             [100,50,25,10]
kernels:           [1,2,3]
slice_thicknesses: [0.6, 1.0, 2.0]
```

LISTING 1: Configuration file use in this experiment for the reconstruction of the cases listed in "case_list.txt." Cases will be reconstructed with 4 simulated dose levels (100%, 50%, 25%, and 10%), three reconstruction kernels (smooth, medium, and sharp), and three slice thicknesses (0.6mm, 1.0mm, and 2.0mm). The "library" parameter specifies the location where all reconstructed data will be stored, and where subsequent QI analysis will take place.

The launcher script's handling of job list creation, and spawning of all further processes simplifies the role of the researcher in experimental setup which increases throughput and reduces the risk of possible configuration errors.

2.5.2 Pipeline Daemon

The daemon is a management script that runs in the background and ensures that system resources are utilized continuously and efficiently. The daemon performs three primary functions: (1) poll the GPU resources of the current system and detect when they are available, (2) spawn "queue_items" which handle the processing of individual reconstructions (see below) when GPU resources are available, and (3) evaluate the exit status of the queue items for logging/debugging purposes.

The daemon runs continuously once the pipeline is launched until there are no more items in the current queue. Checking for available GPUs is done via the polling of a directory of lock files every five seconds. If an available GPU is detected (i.e. no lock file is present) the daemon removes the next item from the job queue, assigns it to the GPU, generates a corresponding lock file, and spawns a new queue item. This ensures that multiple jobs do not compete for the same resources and that all of a system's available GPU resources are used continuously to maximize throughput.

2.5.3 Queue Items

The queue item script handles the processing of an individual reconstruction from start to finish. After receiving its instructions from the daemon in regard to which reconstruction to compute, and which GPU to utilize, the primary steps of this process are data-fetch (i.e. retrieval of raw data from network storage), simulation of reduced dose data if required, and reconstruction according to requested parameters. In addition to these tasks, the queue item also manages data organization for the given reconstruction, generating a specific directory structure inside of the project library (a directory specified in the configuration file) that prepares the case for quantitative imaging analysis.

2.5.4 Quality Assurance

A critical challenge of the pipeline is to ensure that reconstructions and analysis tasks were performed correctly. To accomplish this on the large scale required for the pipeline, slice visualizations are automatically generated and presented to researchers in structured HTML documents allowing for the rapid review of the thousands of image volumes generated. Sample visualizations are shown in Figure 2-16. A sample HTML quality assurance document is show in Figure 2-17 highlighting the ease with which errors can be detected using this approach. Since the generation of QA documents is a key component of each analysis module, this approach can scale with dataset size and allows researchers to easily review and correct problems.

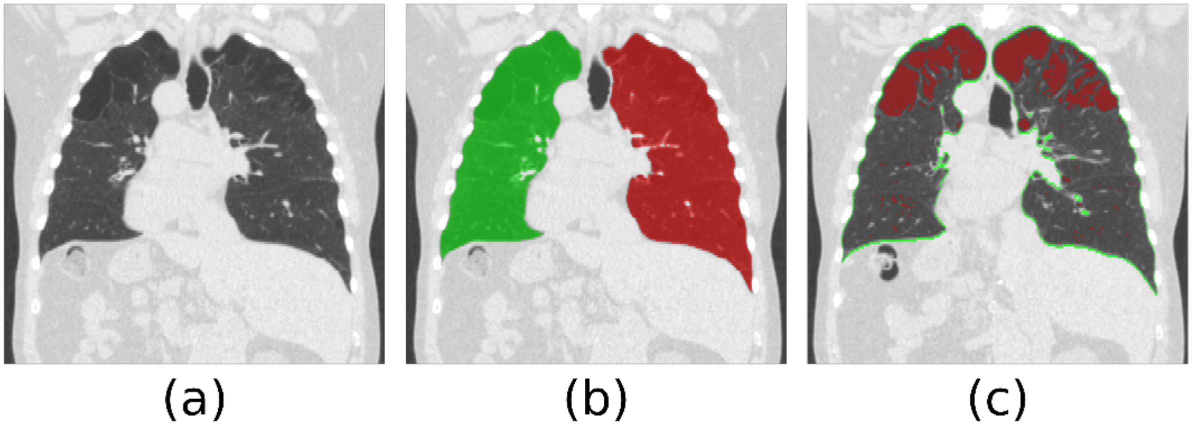
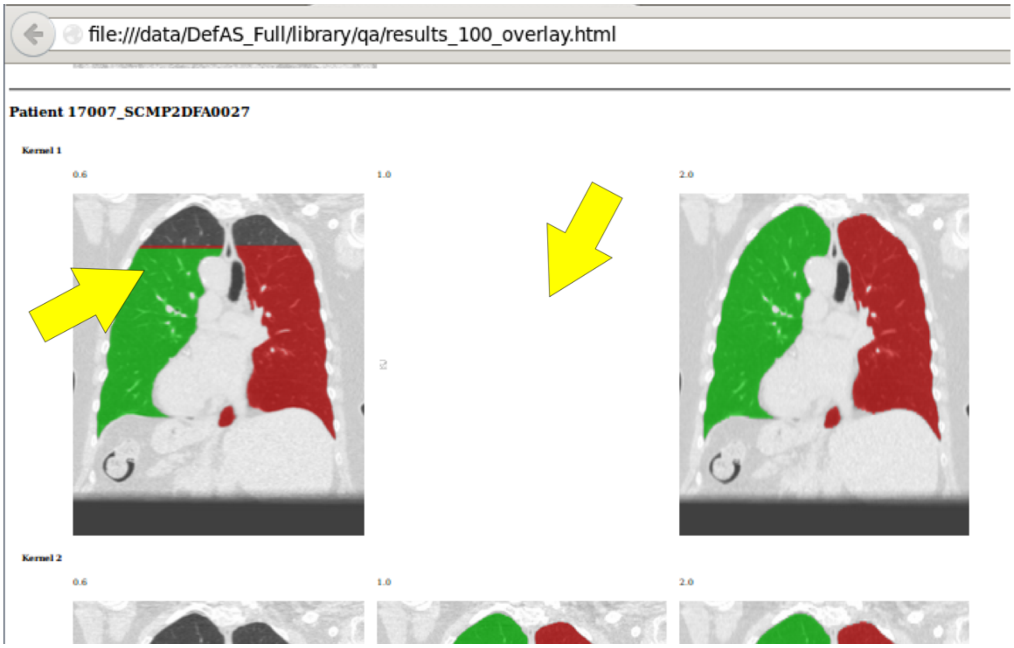


FIGURE 2-16 : SAMPLE QA IMAGES UTILIZED TO VERIFY RECONSTRUCTION QUALITY AND THAT QUANTITATIVE TESTS ARE BEING CORRECTLY APPLIED. (A) RECONSTRUCTION, (B) RECONSTRUCTIONS AND LUNG SEGMENTATIONS, (C) QUANTITATIVE EMPHYSEMA SCORING.



file:///data/DefAS_Full/library/qa/results_100_overlay.html

Patient 17007_SCMP2DFA0027

FIGURE 2-17 SAMPLE HTML DOCUMENT, VIEWABLE IN A STANDARD WEB BROWSER. TWO ERRORS ARE CAUGHT (HIGHLIGHTED WITH ARROWS). A FAULTY SEGMENTATION IS SHOWN (TOP OF THE LUNGS IS TRUNCATED IN THE 0.6 MM SLICE) AND A MISSING RECONSTRUCTION OR SEGMENTATION FILE IS CAUGHT VIA THE IMAGE MISSING FROM THE GRID.

2.5.5 Data Organization

The pipeline creates a unique directory structure designed specifically for quantitative imaging analysis and stores imaging data and study metadata directly into the structure for further analysis. The elements of the directory structure are described in Table 2-9 and Table 2-10. Because the directory structure is standardized across all pipeline runs, post-processing and analysis tasks are simple to apply across all image volumes in a pipeline library, and all analysis data is stored with its respective image data making aggregation for statistical analysis efficient and straightforward. Furthermore, multiple QI analyses can be performed on the same dataset using different analysis modules.

TABLE 2-9 PARENT DIRECTORY STRUCTURE FOR A PIPELINE LIBRARY. THE TOP-LEVEL DIRECTORIES HOLD INFORMATION FOR THE ENTIRE LIBRARY, SUCH AS AGGREGATED ANALYSIS RESULTS AND HIGH-LEVEL LOGGING INFORMATION.

Directory element	Purpose
case_list.txt	Stores original file paths to each raw data file used in the current library, and a "hash" value of the raw data file. The hash serves as a unique identifier, and helps to ensure it is not duplicated unnecessarily.
Eval/	Directory containing final, aggregated quantitative imaging data ready for statistical analysis and interpretation.
Log/	Directory containing copies of all pipeline logging data including the daemon log and logs from individual queue items.
Qa/	Directory containing auto-generated structured documents to assist with quality assurance
Recon/	Directory containing all of the reconstructed image data and results from individual queue items (note that there is further directory organization discussed in table 2)
Recons.csv	Contains a record of all reconstructions present in the library including data such as the source raw data file (and its unique hash value), parameter configuration information, and filepath to the image data.

An organization scheme such as this one is critical for the efficient use of data on the scale of that output by the pipeline. Manual management, such as that provided by dragging and dropping folders and files, does not typically scale to thousands of image datasets, and any speed improvements gained from fast, efficient, automated reconstruction would likely be lost.

TABLE 2-10 INDIVIDUAL RECONSTRUCTION DIRECTORY STRUCTURE (ALL STORED IN THE "RECON" DIRECTORY FROM TABLE 1). DIRECTORIES IN THIS CASE STORE DATA ONLY FOR ONE RECONSTRUCTION. THIS INCLUDES RAW QUANTITATIVE ANALYSIS DATA SUCH AS SEGMENTATIONS, AND INDIVIDUAL TEST RESULTS.

Directory element	Purpose
Eval/	Directory containing "compiled" quantitative imaging data (e.g. complete multi-score results for the emphysema module)
Img/	Directory containing all image data and metadata for the current reconstruction
Log/	Directory containing all logging data for the current reconstruction as well as stdout and stderr output.
Qa/	Directory for the storage of reconstruction-specific data used for quality assurance, for instance, a PNG visualization of overlay of the segmentation on top of the reconstruction
Qi_raw/	Directory containing "raw" quantitative imaging results, such as a histogram of voxel values inside of an ROI, computer automated detection reports/results, etc.
Ref/	Directory containing non-pipeline data specific to the reconstruction (i.e. "reference" data). E.g. A clinical reconstruction being used for comparison against with the pipeline data; human-reader markings being used for CAD comparison
Seg/	Directory containing and segmentations for the current dataset.

2.5.6 Code Availability

The pipeline GPU queuing framework source code as well as the analysis cluster framework code is being made available under the free, open-source GNU General Public License version 3.0 to encourage usage in research and quantitative imaging. Full details can be found on the Github page [96], however in brief this means that users are free to copy, distribute and modify the software provided changes are identified and dated in the source code and any modifications are made freely available under the same license. The reconstruction software, both FreeCT_wFBP and FreeCT_ICD, has also previously been made freely available [61]. At present, the code for the calibrated noise addition and the analysis modules (segmentation, internal data format conversion, etc.) cannot be released due to proprietary code and research agreements, however free, open-source versions would work within the frameworks being released, and future efforts from our research group may result in the release of some of this code.

2.6 Demonstration of Pipeline Performance

2.6.1 Methods

To illustrate the utility and performance of the pipeline, datasets were created for a project in which lung cancer screening CT datasets representing a wide range of acquisition and reconstruction parameters were created. This dataset used the raw data from 142 subjects to create image datasets that represented 4 dose levels (original and 3 simulated reduced dose levels), 3 reconstruction kernels and 3 slice thicknesses. This resulted in a total of 36 reconstructions per subject, 5112 unique image datasets in total. Total size of the dataset was approximately two terabytes.

The reconstruction portion of the pipeline was run on an Alienware Aurora R4 computer with an Intel i7-4960X CPU (3.6 GHz, 15 MB L3 cache), 32 GB of RAM, and two Nvidia GeForce GTX 780 GPU (2304 cores, base clock speed of 863 MHz) with 3.2 GB of global memory each. Analysis of the reconstructed volumes was performed on an in-house computing cluster built with HTCondor cluster software with 15-25 computers in use at a time.

Log files generated by the pipeline were mined for timing data using a Python script that is part of the pipeline software package [96]. Start and stop times for each major processing step were recorded for each individual reconstruction (i.e. data fetch time, simulated dose reduction, and image reconstruction), and elapsed times were computed for both the individual steps as well as the overall execution time for the job queue items (times for all individual steps plus any pipeline overhead). Average times across all 5,112 reconstructed image datasets were computed and compared with previous similar experiments conducted by our research group.

2.6.2 Reconstruction

To explore robustness of emphysema scoring to protocol variation, a range of parameters were selected to capture the possible variability one might see clinically, and additionally some configurations that would push the limits of study “acceptability” for diagnosis, in particular with respect to dose reduction. For this study, four doses were explored: 100%, 50%, 25%, and 10% of the original “low dose” scans (approximately 2.0, 1.0, 0.5 and 0.1 mGy CTDIvol), three reconstructions kernels: smooth, medium and sharp (corresponding roughly to Siemens B10f/B20f, B40f/B50f, and B60f respectively), and three slice thicknesses: 0.6, 1.0, and 2.0mm. Thus, each study was evaluated under 36 different parameter configurations and sample reconstructions with each parameter configuration are shown in Figure 2-18.

Simulated dose reduction was performed on the raw data with a noise model [58], an implementation of which has been validated and utilized for similar, previous experiments[54], [55]. The model adds noise to individual projections considering quantum and electronic noise. Electronic noise is an important consideration since the starting dose of CT lung cancer screening is already low; samples of electronic noise were acquired directly from the scanner on which all patients were scanned. Additionally, a realistic attenuation model of the bowtie was generated using measurements from the scanner. For the pipeline, a GPU implementation of the noise model has been developed that achieves an acceleration of roughly 12x. More discussion of the noise model will be provided in Chapter 3.

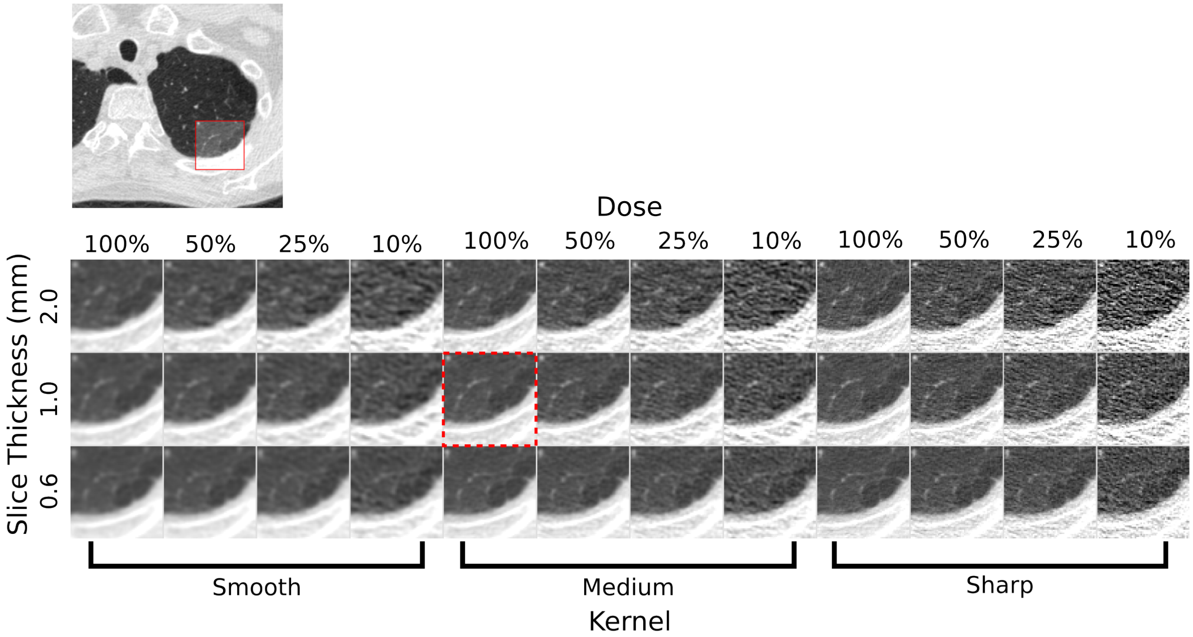


FIGURE 2-18 SAMPLE RECONSTRUCTIONS OF AN ROI IN THE LUNGS ACROSS THE PARAMETERS UTILIZED FOR THIS EXPERIMENT. ROI INCLUDES A SMALL POCKET OF EMPHYSEMA (RIGHT SIDE, AGAINST LUNG WALL). THE APPEARANCE AND CONTRAST OF THE EMPHYSEMA POCKET RELATIVE TO THE LUNG PARENCHYMA CHANGES SUBSTANTIALLY WITH PARAMETER SELECTION. THE SCAN MOST SIMILAR TO WHAT IS PERFORMED CLINICALLY AT OUR INSTITUTION IS HIGHLIGHTED WITH A RED, DASHED RECTANGLE.

All reconstructions were performed using FreeCT_wFBP, designed to be similar to the clinical weighted filtered backprojection algorithms utilized on Siemens scanners. While not precisely the same algorithm, when applied to raw data from the scanner utilized, it has been shown to meet the criteria specified by the ACR CT accreditation protocol [43], and produce clinically-similar reconstructions [61].

2.6.3 Analysis

Threshold- and histogram-based emphysema scoring was chosen as an example task on which to test the pipeline because they are established approaches and have been the subject of much research to date [30], [33], [35], [36], [38], [39]. Four analysis modules were utilized to carry out the analysis: (1) data format conversion (2) lung segmentation (3) calculation of the lung histogram, and (4) emphysema scoring (shown in Figure 2-19). The data format conversion reads the reconstructed image data and image metadata and converts it to a custom format suitable for use with the automated segmentation tool. The lung segmentation module reads the converted image data and runs previously-published, fully-automated lung segmentation software [67]. The histogram calculation module then utilizes both the converted image data and the generated segmentation to create a histogram of the lung voxels. Finally, the emphysema scoring module pulls on all of the previously generated data to achieve final scoring values for the various executed tests. The metrics calculated for this experiment were density mask metrics [6], calculated by evaluating the number of voxels in the lung below the given threshold (e.g. -950HU), and percentile metrics, calculated by evaluating the location of the N^th percentile of the lung parenchyma histogram (most commonly the 15th

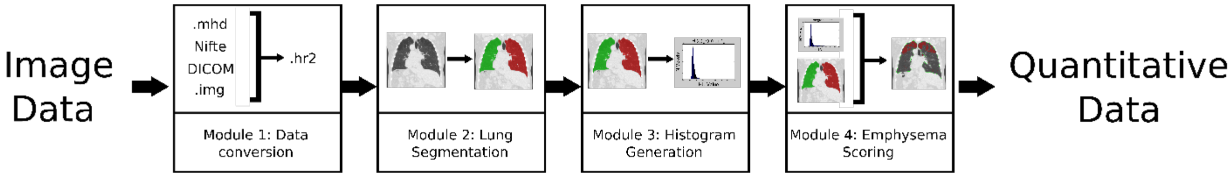


Image Data
.mhd
Nifte
DICOM
.img
.hr2
Module 1: Data conversion
Module 2: Lung Segmentation
Module 3: Histogram Generation
Module 4: Emphysema Scoring
Quantitative Data

FIGURE 2-19 ANALYSIS MODULES USED TO GENERATE QUANTITATIVE RESULTS FOR EMPHYSEMA SCORING APPROACHES ASSESSED IN THIS STUDY. MODULE 1 CONVERTS FROM STANDARD IMAGE OUTPUT FILE TYPES TO THE FILE TYPE USED BY THE OTHER ANALYSIS MODULES. MODULE 2 ACCEPTS CONVERTED IMAGE DATA AND PERFORMS AUTOMATED SEGMENTATION OF THE LUNGS. MODULE 3 THEN LEVERAGES THE OUTPUT OF THE SEGMENTATION MODULE AS WELL AS THE IMAGE DATA TO CREATE A HISTOGRAM OF THE LUNGS. FINALLY, MODULE 4 UTILIZES THE RESULTS OF ALL PREVIOUS THREE MODULES TO EVALUATE THE DIFFERENT EMPHYSEMA SCORES FOR THIS EXPERIMENT (SEE TAB. 3 FOR A COMPLETE LIST). EACH ANALYSIS MODULE IS DESIGNED IN SUCH A WAY THAT IT CAN BE USED FOR FUTURE EXPERIMENTS.

percentile). Additionally, mean, median and volume of the lung was computed. Density mask metrics were computed using thresholds from -900HU to -970HU in increments of 10HU, and 10th, 15th and 20th percentiles were computed.

2.6.4 Results

Table 2-11 summarizes the timing results for the experiment conducted and the average times required for each processing step of the pipeline run. For the data-generation portion of the pipeline (reduced dose simulation, image reconstruction, and post-reconstruction data handling), the most time-consuming step is reconstruction requiring on average approximately 4.4 minutes, while simulated dose reduction and data handling requires less than ten seconds on average.

Table 2-11 Timing results for pipeline run. While speedup is notable by itself, it is also important to consider that no researcher involvement beyond initial configuration is required during the 8.75 days of run time, while substantial time and attention was required for Young et al. 2017 [55]. "Queue item time" considers any computational or data organization overhead, in addition to the time required to perform data fetch, dose reduction, and reconstruction. Total time is dependent on the system used to run the pipeline. Modern GPUs coupled with a greater number of GPUs in a system will substantially reduce the total run time since the individual reconstructions will run faster, and a greater number of reconstructions will be processed concurrently.

Mean data fetch time	1.74 s
Mean dose reduction time	8.81 s
Mean reconstruction time	4.40 min
Mean queue item time	5.57 min
Total time, full dataset (2 GPUs)	8.75 days
Approximate speedup over Young et al. 2017 [55]	**72x**

In general, the GPU implementation of simulated dose reduction requires approximately 1.5 minutes to process a full case and the data-fetch requires approximately 30 seconds when these steps are required, however if the required raw data file is already in the library, or the required dose reduction has already been simulated, the pipeline does not re-compute them. Thus, most scans end up not requiring a separate data fetch or dose reduction step, reducing the average time required for these steps dramatically.

The total size for the reconstructed dataset consisting of 5,112 reconstructions and 568 raw data files (142 x 4 dose levels) required approximately 2 terabytes of storage. Raw data files were on average approximately 2 gigabytes and reconstructed image data ranged from 200-600 megabytes per reconstruction.

Thirty-one reconstructions did not succeed on the first try, and had to be re-queued and re-processed. The failures were likely due to GPU memory conflicts since one of the graphics cards was being used to drive a computer display (which would not occur on a dedicated system); all succeeded on the second try. The pipeline software package provides a script that automatically identifies failed reconstructions and adds them back to the job queue making this “clean up” step simple and fast to perform for the researcher.

Additionally, automated segmentation is known to be imperfect, and the structured QA documents allowed for the fast identification of segmentations with errors. 30 subjects were identified as having one or more segmentations that needed revision. Most errors occurred on the 0.6mm slice thickness. Criteria for error identification included substantial airway inclusion in the segmentation, and/or “truncation” of the upper lung which were easily visible in the segmentation QA document. Once the failed cases were identified, segmentations for them were manually edited to correct errors. Quality assurance of all results required less than one day.

2.6.5 Discussion

In total, the pipeline required slightly over one week for data generation and analysis for 142 lung screening patients, assessing 36 unique reconstruction configurations of each scan. The total number of reconstructions analyzed was 5,112 and 15 quantitative imaging metrics were computed for each reconstruction (all 15 were histogram based). The pipeline allowed for this experiment to be conducted approximately 80 times faster, and with substantially less researcher involvement required than the most comparable experiment conducted by our research group, which required approximately six months for data generation (i.e. simulated dose reduction and reconstruction) and more for quality assurance and analysis[55]. While a larger cohort was assessed in [5] (N=481), only one "dimension" of CT parameter space (i.e. acquisition dose) was assessed, and only three data points per patient were tested (i.e. 100%, 50%, 25% of clinical dose) for a total number of reconstructions of 1,443.

The reconstruction portion of the pipeline is highly optimized and performed extremely well. Reconstructions were generated as expected with few failures. Failures were easily identified and reprocessed to complete the dataset which helps make the pipeline be an efficient and robust tool for large-scale quantitative imaging work.

The pipeline is programmed to automatically scale to the system on which it is running. Improved GPU hardware and a greater number of GPUs on the pipeline system will further accelerate the pipeline without any code modifications required. "Deep learning" workstations, such as one recently acquired by our research group, are becoming more common in research groups and typically contain four, state-of-the-art GPUs. Preliminary tests suggest that the new workstation's Nvidia GTX 1080Ti GPUs reconstruct cases 5x faster than the GTX 780 GPUs utilized for this work due to a faster clock speed and a greater number of computing cores. With four GPUs in the new machine, this suggests that all of the data for this work could have been generated in approximately one day.

While faster than any previous alternatives, it was observed that the current implementation of the analysis modules could be improved. Namely, the computational overhead required for the cluster node configuration script was a substantial burden requiring more time than the actual quantitative image processing steps. The simplest potential solution would be to revise the node configuration script and optimize for only the specific resources required for a given experiment (a "general" version giving access to all resources was used in this work). This will be done for the next experiment using the pipeline, however it is not clear that this will yield substantial improvements in performance. Another potential pathway to improve this would be to let a single cluster node process all modules for a given reconstruction. While this is a promising route forward, it is somewhat less easily adapted for general quantitative imaging than the current modular implementation due to the dependence of some processing steps on data from other reconstructions (e.g. in this case, segmentations from the 100% cases were utilized for the reduced dose cases, instead of attempting to segment the low-dose cases). Implementation of modules would be experiment specific, and a parent "analysis" script (analogous to the reconstruction "queue items") would be required. These scripts would also be required to build in protection for race conditions and shared resources to ensure accurate results and reduce node latency.

2.7 Review and Discussion of the Pipeline

Over the course of the previous chapter we have presented key developments enabling the construction of a high-throughput pipeline for the reconstruction and quantitative analysis of CT image data, with a specific focus on support for clinical, 3rd-generation, diagnostic, helical CT systems. The developments were: (1) FreeCT_wFBP, GPU-accelerated weighted filtered backprojection reconstruction software; (2) FreeCT_ICD, multicore, fully-3d, model-based iterative reconstruction with quadratic and edge preserving penalty functions; and (3) an automated reconstruction and analysis framework, leveraging modern GPU computing technology to accelerate, and improve high-throughput quantitative imaging studies. We have demonstrated the clinically-acceptable and clinically-similar performance of both reconstruction algorithms, and illustrated the performance gains possible for quantitative imaging studies through the utilization of the automated framework.

While the highest throughput is achieved when using the pipeline for both reconstruction and analysis, the pipeline workflow has been constructed to accept image input from multiple sources. In particular, there may be instances where it is beneficial to perform quantitative evaluation of image data from a clinical scanner or other devices (such as a manufacturer "recon box"), which is fully achievable under the data paradigm utilized. Even non-standard or proprietary image formats are usable; however, a small data conversion tool or script will likely need to be added to the data conversion module to translate into a format directly usable by the pipeline. This functionality will be utilized in Chapter 3 to conduct the analysis of reconstructed image datasets obtained on the scanner utilizing the Siemens SAFIRE reconstruction algorithm.

In order to take full advantage of the pipeline, some prerequisites need to be met. First, users must have access to raw projection data, be it from the clinical scanner or from a simulation. This can be a challenge if physical access to the scanner console is not available, or if manufacturers do not allow the exporting of the projection data. Second, users must be able to decode the raw projection data if it is stored in a proprietary format. This has been recognized as a challenging problem for the generalization of these approaches to the broader diagnostic CT community, and effort has been made to overcome this such as the introduction of a DICOM-based vendor-independent CT raw data format [57]. Finally, to achieve the calibrated noise reduction discussed in this work, a "noise model" for each scanner must be developed which involves characterizing the scanner's bowtie filter as well as the detector's electronic noise and gain. This requires physical access to the scanner as well as a means of acquiring a zero tube current scan (i.e. no x-ray production), typically via service mode.

Future improvements to the pipeline are planned, namely a more robust, integrated interface that combines reconstruction process monitoring, quality assurance, and analysis module configuration into one application. New reconstruction algorithms are under development, namely FreeCT_ICD, will be added as a configuration option to the pipeline allowed users to select from weighted filtered backprojection or iterative reconstruction [97] and more fully capture the broad variety of clinically-realistic reconstructions. Lastly, while the pipeline's current implementation is intended for CT imaging, the GPU job queuing framework developed here is generalizable beyond just CT reconstructions and would be an important tool allowing researchers to leverage multi-

GPU workstations and servers in a manner not currently available without custom scripts and substantial programming investment.

2.8 Conclusions

The demonstrated reduction of the time required to go from experiment conception to finalized quantitative results was critical to the work presented in this dissertation, however it is an important advancement for the future investigation of quantitative imaging. Since perfect standardization of CT systems will not be realized in the near future, new and existing quantitative imaging metrics and evaluation methods must be tested for robustness to the myriad conditions that occur in day-to-day clinical CT imaging. The pipeline represents an impressive amount of computing power, however its most important developments are the new studies it can enable that were previously intractable due to logistical overhead of data acquisition, and the new scientific insights possible with such a data paradigm.

While previous investigations into the impacts of quantitative imaging have been performed (e.g. [30], [33], [35]–[39]), they have not been comprehensive enough to establish robust confidence for widespread clinical use. One potential application of the pipeline is the exhaustive search of clinical parameters to establish “acceptable” conditions under which a given quantitative test can be performed reliably. From these conditions, future CT protocols can be designed in which clinicians can confidently use quantitative image tests to assess their patients. Such a study will be presented in Chapter 3 for quantitative emphysema scoring. The pipeline is the first tool of which the authors are aware that allows for the thorough investigation of such interplay between CT imaging and reconstruction parameters and represents an exciting new pathway towards new experiments.

Further applications also exist, such as a test-bed for image post-processing and analysis techniques, or as a data-generation tool for machine-learning and deep-learning applications. By providing all of the tools described in this chapter as free, open-source software, we hope to help accelerate the improvement and clinical translation of quantitative CT imaging.

Appendix: Validation of the Pipeline Infrastructure

Many individual components comprise the pipeline infrastructure described in this chapter and validation of each component is a critical step in the development process to ensure that accurate, clinically-relevant data and results are obtained. The components requiring validation for the work in this dissertation were the following: (1) the noise addition model, (2) reconstruction, (3) segmentation, (4) quantitation, and (5) data handling. While some elements were developed from scratch (reconstruction and data handling) and required extensive validation, others were developed and validated for other works (noise addition and segmentation [54], [67], respectively) and the implementations were utilized without modification to the underlying approach beyond a "wrapper" script to automate calls to the programs. The quantitation step that was re-implemented for this work was carefully checked for exact agreement using an existing software package (namely the in-house QIA software package). The remainder of this appendix will focus on the steps used to validate the newly developed individual components (reconstruction and data handling) as well as the entire pipeline to ensure the integrity of the results.

For reconstruction, FreeCT_wFBP and FreeCT_ICD were validated using a series of tests that reflect an approach commonly employed for the development and implementation of CT reconstruction algorithms. Initial implementations of the reconstruction were developed using simulated, exactly-known analytic phantoms, namely the FORBILD head and thorax phantoms [83], [98]. These two phantoms have been employed extensively in the development and testing of other reconstruction algorithms (e.g., [59], [72], [76], [99]–[105]), and provide a large set of objects that were designed to result in artifacts and discretization errors should the algorithm implementation not be correct. These phantoms provided test cases in which the phantom and the simulated system geometry were known exactly. Reconstructed images were reviewed by a CT reconstruction expert (Dr. Frédéric Noo, University of Utah, with extensive experience in evaluating reconstruction algorithms and identifying artifacts and errors in these phantoms) for attenuation value accuracy and artifacts due to implementation errors. After satisfactory initial implementations were reached, further simulation studies were conducted with the inclusion of flying focal spots. Each flying focal spot configuration was tested and debugged individually on the FORBILD simulations and reviewed by the reconstruction expert until a satisfactory implementation (i.e. free of artifacts and producing the expected resolution improvements) was reached.

After all expected geometry conditions were implemented and tested with simulation data, testing moved to evaluation using datasets obtained from the clinical CT systems to be reconstructed in our quantitative imaging work (the Siemens Definition AS 64, Medical Plaza 200, UCLA). The first half of this process was to use the ACR phantom, which has known characteristics designed to evaluate CT system performance. Further testing was performed to ensure that requested reconstructed slice locations agreed with reconstructions performed on the clinical scanners, and that the implementation properly handled changes in pitch and table direction (i.e. into the gantry and out of the gantry).

Only after all of the above were tested and reviewed by the researchers and reconstruction expert was FreeCT employed for the reconstruction and quantitative imaging of clinical subject data. Sections 2.3.6 and 2.4.2 report the results of ACR testing with FreeCT_wFBP and FreeCT_ICD respectively. To summarize validation results overall:

- All CT number evaluations fell within the range deemed acceptable by the ACR CT Accreditation Protocols (Table 2-4)

- All uniformity evaluation met ACR criteria (Table 2-5)

- ACR contrast-to-noise (CNR) criteria were met with the smooth and medium reconstruction kernels. The sharp kernel did not meet the CNR criteria (>1.0 for adult protocols), however this is typical with sharp reconstruction kernels (Table 2-6).

- The use of the in-plane flying focal spot improved spatial resolution and the use of the Z flying focal spot provided a reduction of windmill artifacts, as expected (Figure 2-8 and Figure 2-9).

- FORBILD phantom results demonstrated proper implementation of system geometry and implementation accuracy, reflected by reconstruction of known attenuation values (agreeing to within <0.01%, Figure 2-10), and reconstruction free of artifacts other than the expected aliasing observed due to the noise-free, ramp-kernel reconstruction.

To ensure that the quantitation implementation was correct, a previously-employed (e.g. [56]) software package was initially utilized (i.e. the in-house “QIA” software suite). The final version of RA-950 and the PERC15 scoring was re-implemented using MATLAB to improve computational efficiency. The latter implementation was checked against the results obtained with the QIA-based implementation for agreement by selecting approximately five subject test cases (a volumetric image dataset and a corresponding segmentation) and scoring using both the existing QIA implementation and the newly-developed MATLAB implementation. Both PERC15 and RA-950 scores agreed exactly between the two implementations. As a further check, the histograms of the segmented lung region were found to agree exactly.

Finally, the quality-assurance (QA) images (described in section 2.5.4) serve a dual-purpose: in addition to their quality assurance role, they provide on-the-fly validation of the data handling as the pipeline generates quantitative results. Data handling specifically refers to matching the correct subject and volumetric image dataset with the segmentation and final quantitative results. The quality assurance images illustrate that, for a given reconstruction, the correct subject was identified and matched to the proper segmentation and quantitative results, which is reflected in the fact that the visualizations are generated during the scoring process, therefore the visualizations reflect the actual data utilized to generate the scoring results. Any disagreements in the visualization, which are easily spotted during a QA review, would reflect errors in scoring. Since initial debugging during the pilot studies described below, incorrectly paired images and segmentations have **not** been observed.

Ideally the results of each reconstruction would be reviewed in detail (i.e. slice-by-slice) for both segmentation accuracy and proper scoring of the RA-950 data, however this is intractable given the large-scale datasets being utilized. Instead, a detailed review of a sample of individual cases was performed during a series of pilot studies that were conducted to ensure that the entire pipeline workflow was working properly. In the first study, a cohort of ten subjects was evaluated, and all segmentation and quantitative results were given a detailed review for accuracy [106]. The results of the pilot studies demonstrated that any errors identified using the detailed review were able to also be identified using the QA documents with individual coronal slices. This test was then scaled to thirty subjects in a second pilot study [107] to ensure that results observed on the 10 cohort sample would scale properly without any further software modification. The QA images were employed as the primary means of validation, and detailed review was performed on subjects that had segmentation errors in the QA images, as well as randomly on other subjects. Regarding validation of data handling, the results of this second pilot study were that no data-matching or segmentation errors were found in the randomly selected subjects, and all segmentation errors identified in the QA images were a result of known limitation of the segmentation algorithm utilized (again, not re-implemented or modified for our use beyond the use of a wrapper script for automation), and errors were easily corrected via manual editing of the segmentation. Only after results were verified in the pilot studies was the pipeline infrastructure scaled to evaluate larger cohorts such as those described in Chapter 3 and Chapter 4.

In summary, each component was independently validated when needed (reconstruction, data handling, quantitation), and the entire pipeline itself underwent extensive testing to ensure the integrity of the results of the quantitative imaging measures extracted.

Chapter 3 - Characterizing the Impacts of Slice Thickness, Kernel Selection, and Dose on Quantitative Emphysema Scoring Under Weighted Filtered Backprojection and SAFIRE

Significant portions of the material for this chapter are adapted from the manuscript:

(In preparation for submission to Medical Physics) J. Hoffman, G. Kim, M. Brown, J. Goldin, M. McNitt-Gray. "The Effects of Acquisition and Reconstruction on Quantitative CT imaging: Evaluation of Quantitative Measures of Emphysema in a Large Lung-Screening Cohort."

3.1 Introduction

Emphysema is the degradation of lung alveoli leading to difficulty breathing, shortness of breath, and poor oxygenation of the blood due to limited gas exchange, and is one of the key diseases comprising the larger disease family: chronic obstructive pulmonary disease (COPD). In the United States, roughly 2% of the population is affected by emphysema [108], and worldwide 10.1% of people are affected by COPD [109]. COPD represents the third leading cause of death in the United States, behind heart disease and cancer [17] and can lead to significant degradation of patient quality of life. For these reasons, COPD and emphysema remain important targets in healthcare for improving diagnosis and treatment.

Current gold standard tests for the diagnosis and evaluation of emphysema are primarily functional, consisting of FEV1 ("forced expiratory volume in one second"), and six-minute walking distance. Additional tests include blood gas measurements. While these tests are useful, they are limited in their ability to group patients beyond 3-4 categories or stages (GOLD Classification, [110]), and because diagnosis typically requires spirometry measurement [110], early diagnosis, at which a patient may not yet be symptomatic and the disease may be most treatable, could be very beneficial. Since emphysema is irreversible, early diagnosis is critical for starting treatment to delay onset of severe symptoms, which the functional tests may not be able to provide.

For many years it has been recognized that emphysema is typically highly visible on thoracic CT scans and as such could possibly be quantified using CT [6], [7]. Such tests are able to detect the onset of emphysema earlier and provide more detailed tracking of disease progression than the functional tests. The current most common quantitative CT method utilized is a threshold-based method referred to as “RA-950” (relative area, -950HU), which measures the amount of degraded tissue falling below -950HU relative to the total volume of the lung. Several studies have shown good correlation between this score and the amount of underlying morphological and pathological changes causing or leading to emphysema [6], [7], [21]. Other quantitative tests for emphysema using CT that have been proposed include location of the 15th percentile of the lung histogram, mean value of the lung tissue [20], as well as more modern techniques such as parametric response mapping [27].

Results of quantitative CT for emphysema are promising, however widespread clinical adoption, outside of tightly controlled clinical trials, has not been achieved in the 30 years since its initial introduction. While there are likely a number of contributing factors for this delay, one that is of principle importance is the variation in emphysema scores that is caused by CT acquisition and reconstruction parameters. This has been investigated and reported in a number of different studies. Boedeker et al. [30] reported on observed variations in density mask scores of up to 15% simply by changing the reconstruction kernel. In emphysema scoring with filtered backprojection reconstruction, a clear trend of increasing emphysema score with reduced dose scans has been established [33], [34], [37], [39]. More recently, the impacts of iterative reconstruction algorithms have been explored and found to perhaps be a means to "stabilize" RA-950 scores in the presence of dose reduction [36]–[39], however in some cases they do appear to introduce a small bias in the final metric value and the potential clinical implications of this are somewhat unclear [37]–[39]. Because of this observed variation caused only by changing reconstruction or acquisition parameters (i.e. no change in a subject's underlying disease state), confidence in the widespread usability of emphysema scoring is low.

The use of diagnostic quantitative CT may have improved the early diagnosis of emphysema, however, the number of easily-varied factors (such as slice thickness, reconstruction kernel, dose) impacting the final measured emphysema scores limits its utility in a heterogeneous clinical environment where the same patient may be scanned at different sites, by different technologists, and with different protocols. Because results are not believed to be widely comparable between sites and studies, quantitative imaging of emphysema using CT is thought to only be minimally more useful than the current functional tests [111].

There have been many calls to standardize quantitative imaging protocols (e.g. [11], [111]), however these are unlikely to be achieved in the near future due to costs and practical considerations, such as how to standardize across proprietary manufacturer technology (e.g. x-ray source, filtration, detectors, etc.) and what to do in the event of improper protocol configuration. Because of this, if quantitative imaging for emphysema is to be achieved in any general sense, other approaches are needed. One such alternative approach is to quantitatively evaluate patient scans under a broad range of realistic, clinical acquisition and reconstruction conditions and establish a range or family of parameters choices that produce results comparable to a trusted reference value. This "robustness" testing could lead to increased confidence in the quantitative results, and comfort with clinical interpretation for patients and doctors.

Previous efforts to perform robustness testing for quantitative imaging have typically focused on varying one CT parameter (such as tube current, slice thickness, reconstruction kernel, etc.) and reporting the change in the chosen quantitative test [30], [33], [36]–[38]. This has served well to establish the baseline understanding among the quantitative imaging community that acquisition and reconstruction can be critical factors in quantitative imaging; in other words, *how* we scan can be just as important as *what* we scan. A missing component of nearly all previous studies however has been an investigation of how variation in multiple parameters and their potential interplay can impact quantitative results. The only exception of which the authors are aware has been [35], in which multiple settings of reconstruction kernel and slice thickness were systematically investigated in 20 different combinations (4 slice thickness, 5 reconstruction kernel) for each subject in the study. While very thorough, this study only investigated 21 subjects.

In this study, we build on previous efforts to quantify the robustness of emphysema scoring metrics by leveraging a high-throughput computing platform for quantitative CT (discussed in Chapter 2) to investigate several quantitative emphysema scoring approaches over a large number of reconstruction and acquisition conditions in a large cohort (N=142) of clinical lung screening subjects. Acquisition dose (via dose reduction simulation software, described below), reconstruction kernel selection, and slice thickness (three of the most commonly changing clinical parameters) are varied to create 36 unique weighted filtered backprojection reconstructions of each subject scan to systematically investigate the impact of and interplay between clinical parameters, and gain insight into how their combinations could impact patient diagnosis. Furthermore, an additional set of 36 reconstruction per patient are generated using the Siemens clinical iterative reconstruction approach, SAFIRE ("Sinogram Affirmed Iterative Reconstruction", Siemens Healthineers, Forchheim, Germany). This provided at total of 72 reconstructions of each subject scan with each reconstruction representing a unique point in parameter space.

Finally, from the results, we seek to provide "bounding boxes" of investigated parameters and reconstruction algorithms that would produce reliable quantitative results suitable for diagnosis and evaluation of a subject's emphysema level, in the hopes of providing confidence in quantitative CT results, and a potential pathway for the clinical usage of quantitative emphysema scoring in current medical practice.

3.2 Methods

Patient population

142 lung screening subjects scanned at our institution were selected from an in-house raw data archive. All scans were acquired on the Siemen's Definition AS 64 (Siemens Healthineers, Forchheim, Germany) according to the lung cancer screening protocol of

our institution: 120 kV, Tube Current Modulation (TCM) was on in all scans ("CAREDose4D" is the name for the proprietary Siemens implementation), 25 Quality Reference mAs, pitch of 1.0, and collimation of 64 x 0.6mm (which includes the use of the Z-Flying Focal Spot – ZFFS).

Quantitative scoring methods

The primary emphysema scoring metrics evaluated in this work were RA-950 and the location of the 15th Percentile of the global lung histogram (PERC15). To compute RA-950, the number of voxels in the lung (i.e. identified by the lung segmentation) with HU values below the given threshold were tallied and then divided by the total number of voxels in the lung to yield a fraction of the lung that is below the specified threshold. Percentile location measurements (PERC15) were found by generating a histogram of lung voxel values and identifying the Hounsfield Unit value of the 15th percentile of the histogram; this was done using automated software.

Both emphysema scoring approaches required patients' lungs to be segmented prior to software evaluation. Three lung segmentations were generated for each patient, one per slice thickness, using the segmentation module of our fully automated CAD software [67] on the 100% dose, smooth kernel cases. Segmentations were not run on each reconstruction in order to eliminate variations in quantitative score due only to segmentation differences. Segmentation agreement with patient anatomy was verified by reviewing overlays of the segmentations with patient scans. 22 patients had at least one segmentation "failure" in which large portions of the airways were included, or a portion of the lung was excluded. The bulk of these failures occurred on only the 0.6mm slice thickness (14 patients), while several patients had multiple automated lung segmentation failures (6 subjects).

Pipeline and range of parameters

A high-throughput pipeline was utilized to generate and analyze all of the imaging data for this experiment. This pipeline has been described in detail in Chapter 2, however to summarize, the pipeline is a GPU and cluster computing framework leveraging the free, open-source reconstruction software FreeCT_wFBP (Chapter 2, page 36, and [61]), as well as in-house analysis software for the high throughput creation and analysis of CT imaging data. The pipeline accepts a small configuration file specifying the list of scans to be reconstructed and parameter configurations desired. It then performs the reconstruction and analysis in a nearly fully-automated manner relying on the researcher only for quality-assurance of the results.

In addition to exploring reconstruction parameters, the pipeline is able to simulate dose reduction for clinical CT scans by leveraging a realistic noise model [58] that has previously been employed for similar studies [54], [55], [112]. In addition to considering Poisson counting statistics, the model is built utilizing scanner specific measurements of electronic noise and the bowtie filter producing a highly realistic, simulated reduced-dose scan, even down to the lowest scanner-allowed dose values. Validation of the noise model was described in [54].

The primary aim of this work was to investigate a range of clinically realistic selections of imaging parameters: reconstruction algorithm, slice thickness, reconstruction kernel, and

CT dose. Lung screening studies at our institution are performed at ~2mGy CTDIvol, and reconstructed at 1.0mm slice thickness, using weighted filtered backprojection with a lung kernel (Siemens B40f or similar).

To capture potentially realistic clinical variability the following parameters were selected: dose, slice thickness, reconstruction kernel, and reconstruction algorithm. Values selected and rough clinical equivalents are given in Table 3-1. All combinations of the parameters selected were investigation for a total of 72 reconstructions per subject scan. Figure 3-1 illustrates a sample lung ROI reconstructed under all of the different parameter combinations.

TABLE 3-1 ACQUISITION AND RECONSTRUCTION PARAMETER VALUES INVESTIGATED. (*) INDICATES PARAMETERS USED WERE DIRECTLY FROM THE SCANNER OR DIRECTLY EQUIVALENT TO SCANNER VALUES.

Parameter	Values	Approximate clinical equivalent
Dose	100%, 50%, 25%, 10%	2.0mGy, 1.0mGy, 0.5mGy, 0.2mGy (CTDIvol)
Slice thickness	2.0mm, 1.0mm, 0.6mm	*
Reconstruction algorithm	FreeCT_wFBP, Siemens SAFIRE	Siemens weighted filtered backprojection, *
wFBP reconstruction kernels	Smooth, Medium, Sharp	B10f, B45f, B60f (Siemens)
SAFIRE sharpness settings	I26, I44, I50 (all strength 3)	*

Quantitative metrics used and analyses performed

Quantitative emphysema scores using RA-950 and PERC15 were determined for all reconstructions. Because there is no clear “truth” value defined for each patient, the change in score relative to a patient’s reference reconstruction was utilized as a figure of merit. The reference reconstruction was chosen to be the parameter configuration most comparable to a clinical quantitative imaging protocol [42]: 100% dose, 1.0mm slice thickness, and the smooth reconstruction kernel with weighted filtered backprojection (outlined with a dashed line in Figure 3-1). Summary statistics are provided for the study population in Table 3-2 and Table 3-3. Figure 3-2 provides histograms of emphysema scores at the reference condition for RA-950 and PERC15.

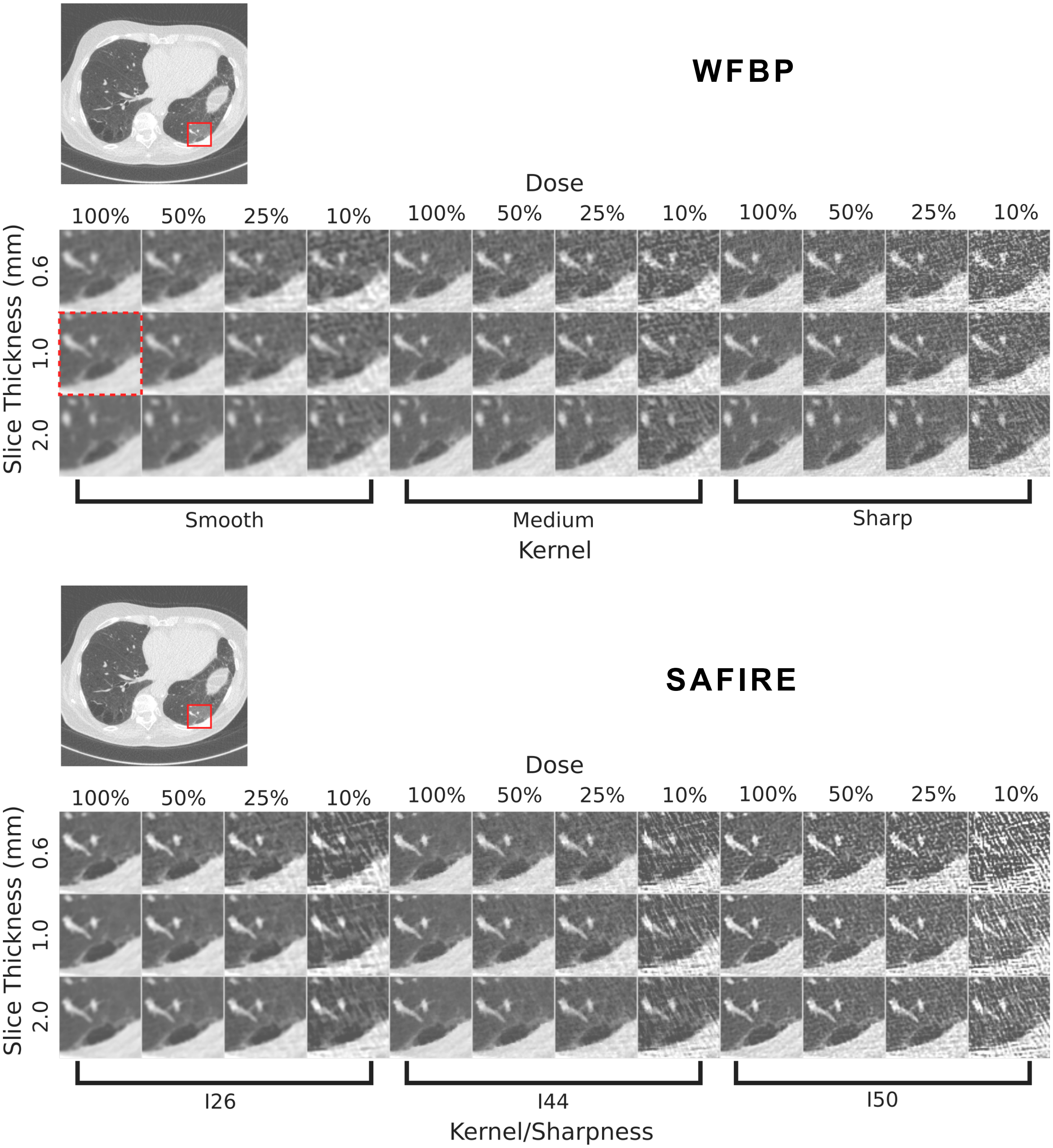


Figure 3-1 Sample region of interest containing a small region of low attenuation. Reconstructed using different combinations of wFBP and iterative (SAFIRE) parameters. Reference reconstruction is outlined with a red, dashed line. All reconstructions are shown with a window/level of 1600/-600.

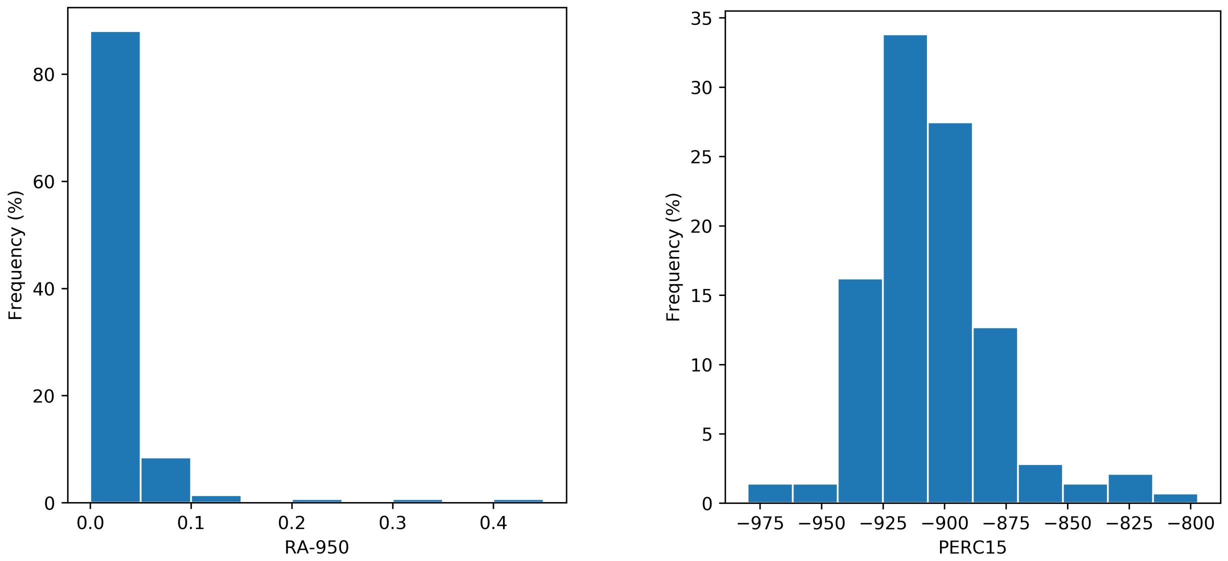


Figure 3-2 Histogram of subject emphysema levels at reference (WFBP, 100% dose, 1.0mm slice thickness, smooth reconstruction kernel).

Table 3-2 Summary statistics of RA-950 and PERC15 scores for study population. Reference condition is 100% dose, 1.0mm slice thickness, and smooth reconstruction kernel. Mean values and standard deviation are provided; minimum and maximum values are giving below in brackets.

Dose	Slice thickness	Kernel	Metric	**WFBP**	**SAFIRE**
100% (~2.0mGy)	1.0	Smooth	**RA-950**	0.024 ± 0.051 [0.000,0.436]	0.050 ± 0.070 [0.001,0.548]
			PERC15	-905.099 ± 26.853 [-980.000,-797.000]	-917.944 ± 26.705 [-987.000,-811.000]

Table 3-3 Summary of emphysema prevalence utilizing RA-950 score at reference. The vast majority of subjects in this study have little to no emphysema.

	RA-950 < 0.05	0.05 ≤ RA-950 < 0.10	RA-950 ≥ 0.10	**Total**
Number of subjects	125	12	5	**142**

To investigate clinically acceptable levels of variation in RA-950 a 5% threshold was utilized (absolute change in score, not percent change); for PERC15, a threshold of 10HU change was evaluated. The change in emphysema scoring metric from reference condition was computed for each patient and parameter configuration, and then averaged across the patient population. From this data, the acquisition and reconstruction parameter configurations producing "acceptable" levels of change in emphysema score were determined and used to provide recommendations for clinical emphysema scoring. A parameter configuration was considered acceptable if the 95% confidence interval determined using a paired t-test of the figure of merit lied within the threshold being evaluated; which for RA-950 scoring was being within 5% of the reference condition and for PERC15 was being within 10 HU of the reference condition.

A key focus of this study was to determine which of the parameters investigated have the largest impact on emphysema score, as well as to determine whether or not any complex interactions between parameters existed. Parameter importance (i.e. large, statistically significant coefficient value) in the model indicates which parameters need to be carefully controlled for clinically (i.e. likely to cause substantial emphysema score change). Additionally, the presence of strong interactions is important to determine since it has implications for establishing guidelines for robust quantitative emphysema scoring. For instance, claims of "safety" for a given parameter would conditionally depend on other parameter selections made, and thus establishing rules or guidelines regarding safe parameter configurations becomes difficult or impossible. Assessing parameter importance and potential interactions was done using regression analysis in two stages: first without interaction terms (i.e. slice thickness, kernel, dose), and second with interaction terms (slice thickness, kernel, dose, and all possible interactions, such as "slice thickness x dose," and "dose x sharp kernel" etc.). A substantially improved R^2 value under the interaction model would indicate the importance of interaction terms to achieve a good fit (i.e. the model closely reflects the data); similar R^2 values suggest that the interaction terms may not necessarily substantially improve or change the model, and coefficient magnitudes should be assessed individually, in particular the interaction terms versus non-interaction terms. If interaction term coefficients are small relative to the non-interaction terms, then the interaction terms do not substantially impact the results, and were ignored in favor of a simpler, non-interaction model. If the interaction term coefficients are on the same order as the non-interaction coefficients, then the interaction results are important to the final model, in particular if the R^2 value is greater for the interaction model. In these cases, the regression analysis including interaction terms was utilized for interpreting results instead of the model without interaction terms.

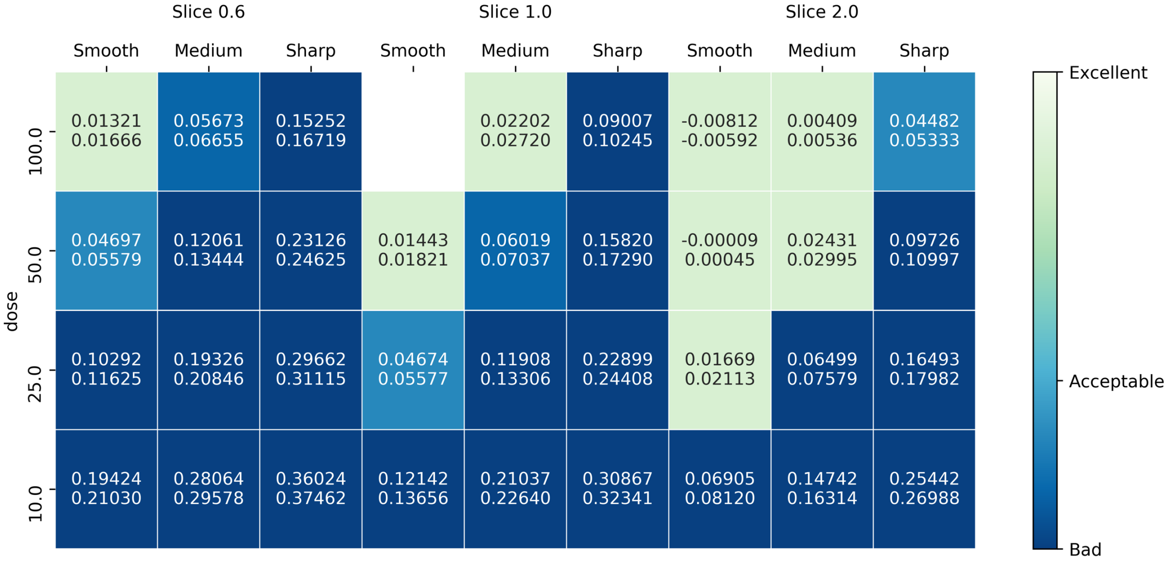
To investigate the potential impacts of overall amount of emphysema present, an effect clearly observed in [35], subgroup-analysis was conducted to determine differences between patients displaying an RA-950 score of greater than 5% (N=17, patients likely having some amount of emphysema present) at the reference condition versus those less than or equal to 5% (N=125, patients having little to no emphysema). Although the number of subjects was small (N=5), an additional comparison was made between the patients with RA-950 at the reference condition of greater than or equal to 10%.

All statistical analysis was conducted using the Python (www.python.org) packages Statsmodels version 0.8.0, and Scipy version 1.0.0.

3.3 Results

3.3.1 Weighted Filtered Backprojection

RA-950



dose	Slice 0.6			Slice 1.0			Slice 2.0		
	Smooth	Medium	Sharp	Smooth	Medium	Sharp	Smooth	Medium	Sharp
100.0	0.01321 0.01666	0.05673 0.06655	0.15252 0.16719		0.02202 0.02720	0.09007 0.10245	-0.00812 -0.00592	0.00409 0.00536	0.04482 0.05333
50.0	0.04697 0.05579	0.12061 0.13444	0.23126 0.24625	0.01443 0.01821	0.06019 0.07037	0.15820 0.17290	-0.00009 0.00045	0.02431 0.02995	0.09726 0.10997
25.0	0.10292 0.11625	0.19326 0.20846	0.29662 0.31115	0.04674 0.05577	0.11908 0.13306	0.22899 0.24408	0.01669 0.02113	0.06499 0.07579	0.16493 0.17982
10.0	0.19424 0.21030	0.28064 0.29578	0.36024 0.37462	0.12142 0.13656	0.21037 0.22640	0.30867 0.32341	0.06905 0.08120	0.14742 0.16314	0.25442 0.26988

FIGURE 3-3 HEAT MAP OF ACCEPTABLE AND UNACCEPTABLE RECONSTRUCTION CONDITIONS FOR RA-950 (I.E. YIELDING A SCORE WITHIN PLUS OR MINUS 0.05 OF THE REFERENCE CONDITION, WHICH IS DISPLAYED AS A WHITE BOX). 95% CONFIDENCE INTERVALS ARE GIVEN INDICATING THE AMOUNT OF CHANGE THAT COULD BE EXPECTED DUE TO A PARAMETER CHANGE. SQUARES RECEIVE GREEN (WHITE, IF FIGURE IS GRAYSCALE) COLORING IF THE CONFIDENCE INTERVALS LIE COMPLETELY WITHIN THE 0.05 THRESHOLD OF ACCEPTABILITY. SQUARES RECEIVE LIGHTER BLUE IF THE CORRESPONDING 95% CONFIDENCE INTERVAL IS PARTIALLY CONTAINED IN, OR EXTREMELY CLOSE (WITHIN 0.01) TO THE 0.05 THRESHOLD. PARAMETER CONFIGURATIONS RESULTING IN SCORE CHANGES BEYOND THIS RECEIVE DARK BLUE COLORING INDICATING "UNACCEPTABLE" CHANGES IN RA-950 SCORE.

Weighted filtered backprojection (wFBP) demonstrated a region of acceptable parameter configurations for RA-950 scoring concentrated around the reference configuration. Figure 3-3 provides a visual representation of the conditions found to be acceptable for the entire pooled population, and as can be seen, most configurations that were only slightly different than the reference yielded acceptable performance, with the exception of the sharp reconstruction kernel. In general, parameter configuration changes that introduce more noise into the image (i.e. moving from a smooth kernel to a sharp kernel, and/or moving from a 1.0mm slice thickness to a 0.6mm slice thickness) typically resulted in a net increase in RA-950 score (i.e. higher apparent emphysema) and greater likelihood of yielding a score that fell outside of the 0.05 threshold of acceptability; parameter configuration changes that would result in less image noise (e.g. thicker slices) had the opposite effect. In some cases, this effect can be leveraged to achieve almost the exact same RA-950 score but with substantively different acquisition and reconstruction parameters, such as 50% dose relative to reference (increases image noise relative to reference), but a slice thickness of 2.0mm (decreases image noise relative to reference), which resulted in a 95% confidence interval of [-0.00009,0.00045] (i.e. essentially the exact same scores as the reference configuration). Figure 3-4 illustrates the same data however in a different manner using line plots.

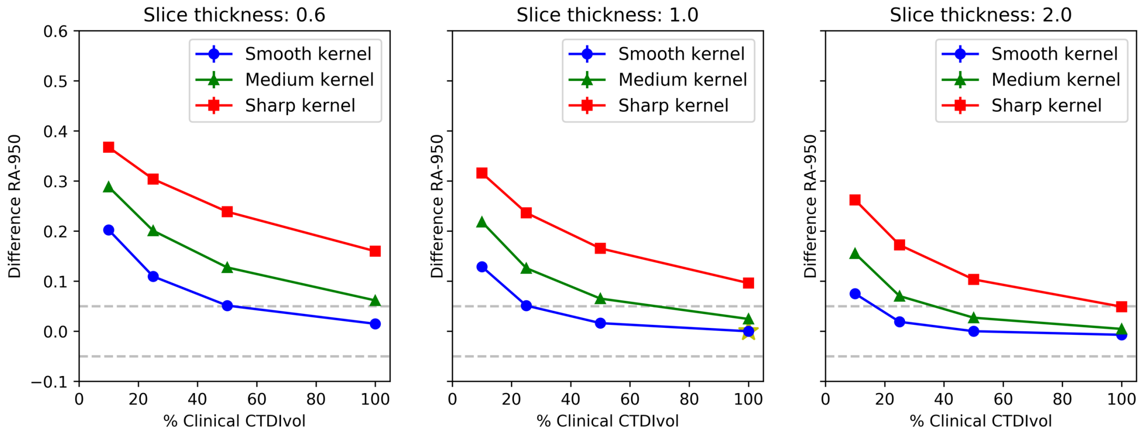


FIGURE 3-4 PLOT OF THE AVERAGE CHANGE IN RA-950 AS A FUNCTION OF PARAMETER SELECTION. AVERAGES WERE COMPUTED ACROSS THE ENTIRE STUDY POPULATION. THE REFERENCE CONDITION IS HIGHLIGHTED WITH A STAR. GRAY, DASHED LINES INDICATE THE THRESHOLD OF ACCEPTABILITY UTILIZED IN THIS EXPERIMENT. ERROR BARS ARE INCLUDED HOWEVER SMALLER THAN THE PLOT MARKERS IN MOST CASES.

Within subpopulations grouped by amount of emphysema at the reference configuration (illustrated in Figure 3-5, page 108), patients with mild and moderate amounts of emphysema were actually somewhat less susceptible to the more “extreme” parameter configurations, such as the sharp kernel and the 10% and 25% dose settings. While these “extreme” configurations still yielded changes in RA-950 score that fell outside of the “acceptable” threshold (for the higher-emphysema subpopulation), these results agree with and support those found in [35], and suggest that patients with at least some emphysema are less likely to receive an incorrect quantitative RA-950 score due to parameter changes than those patients without any emphysema. It appears that this effect becomes stronger the more emphysema a patient has, evidenced by the decrease in variation observed in the ≥ 0.10 group versus the ≥ 0.05 and no emphysema groups. Given the small sample size however for the emphysematous populations, strong conclusions regarding this effect cannot be made from this data.

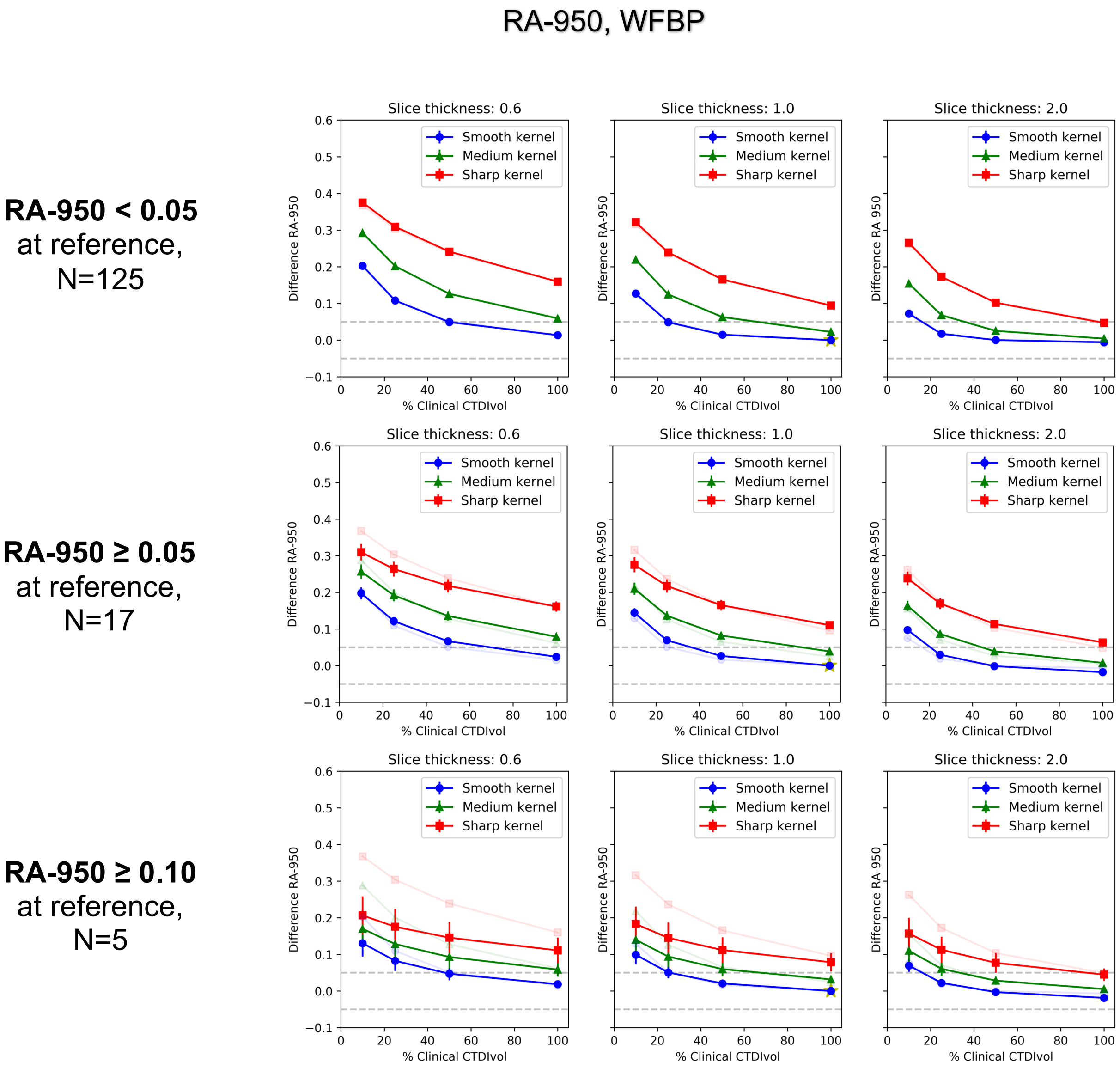
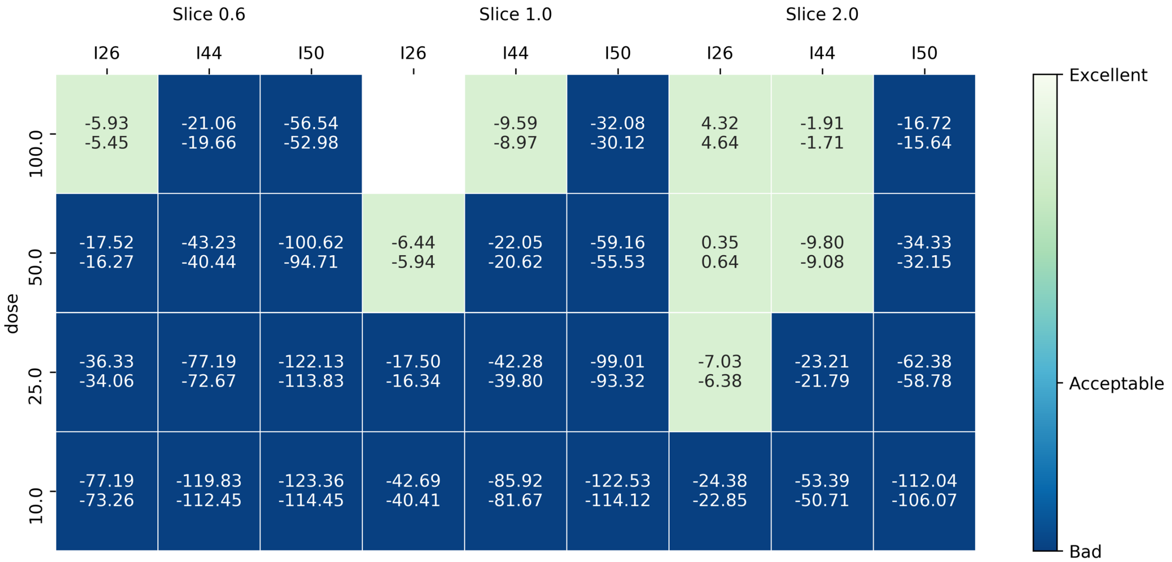


FIGURE 3-5: SUBGROUP PLOTS OF CHANGE IN RA-950 AS A FUNCTION OF PARAMETER CONFIGURATION. TOP ROW REPRESENTS PATIENTS WHO HAVE LITTLE-TO-NO EMPHYSEMA, MIDDLE ROW REPRESENTS PATIENTS THAT HAVE AT LEAST SOME EMPHYSEMA (I.E. A MINIMUM RA-950 SCORE OF 0.05 AT REFERENCE), AND THE BOTTOM ROW SHOWS ONLY PATIENTS WITH MODERATE EMPHYSEMA OR WORSE (I.E. MINIMUM OF 0.10 RA-950 AT REFERENCE). ORIGINAL POOLED PLOTS (FIGURE 3-4) ARE SHOWN WITH REDUCED OPACITY TO HIGHLIGHT HOW THE SUBGROUPS BEHAVE DIFFERENTLY.

PERC15



	Slice 0.6			Slice 1.0			Slice 2.0		
dose	I26	I44	I50	I26	I44	I50	I26	I44	I50
100.0	-5.93 -5.45	-21.06 -19.66	-56.54 -52.98		-9.59 -8.97	-32.08 -30.12	4.32 4.64	-1.91 -1.71	-16.72 -15.64
50.0	-17.52 -16.27	-43.23 -40.44	-100.62 -94.71	-6.44 -5.94	-22.05 -20.62	-59.16 -55.53	0.35 0.64	-9.80 -9.08	-34.33 -32.15
25.0	-36.33 -34.06	-77.19 -72.67	-122.13 -113.83	-17.50 -16.34	-42.28 -39.80	-99.01 -93.32	-7.03 -6.38	-23.21 -21.79	-62.38 -58.78
10.0	-77.19 -73.26	-119.83 -112.45	-123.36 -114.45	-42.69 -40.41	-85.92 -81.67	-122.53 -114.12	-24.38 -22.85	-53.39 -50.71	-112.04 -106.07

FIGURE 3-6 HEAT MAP ILLUSTRATING "ACCEPTABLE" VERSUS UNACCEPTABLE CHANGES IN PERC15 SCORE AS A RESULT OF PARAMETER CHANGES. AN ACCEPTABLE LEVEL OF CHANGE WAS DEFINED TO BE LESS THAN 10HU. 95% CONFIDENCE INTERVALS ARE SHOWN.

Interestingly, PERC15 yielded nearly the exact same set of acceptable conditions (illustrated in Figure 3-6) as RA-950. This is however, perhaps somewhat unsurprising given that both metrics are purely histogram-based, although acceptable regions could be easily changed by where the threshold of acceptability is set. While the acceptable parameter configurations are the same, the behavior of PERC15 is the opposite of RA-950: PERC15 becomes lower with increasing noise in the wFBP images. This is illustrated in Figure 3-7. Furthermore, although the PERC15 response to parameter change is the inverse of RA-950, the same effect on subpopulations is observed: patients with higher levels of emphysema are subject to less change in PERC15 under extreme protocol changes, although no substantive change to the acceptable protocols was observed (Figure 3-8 on page 111).

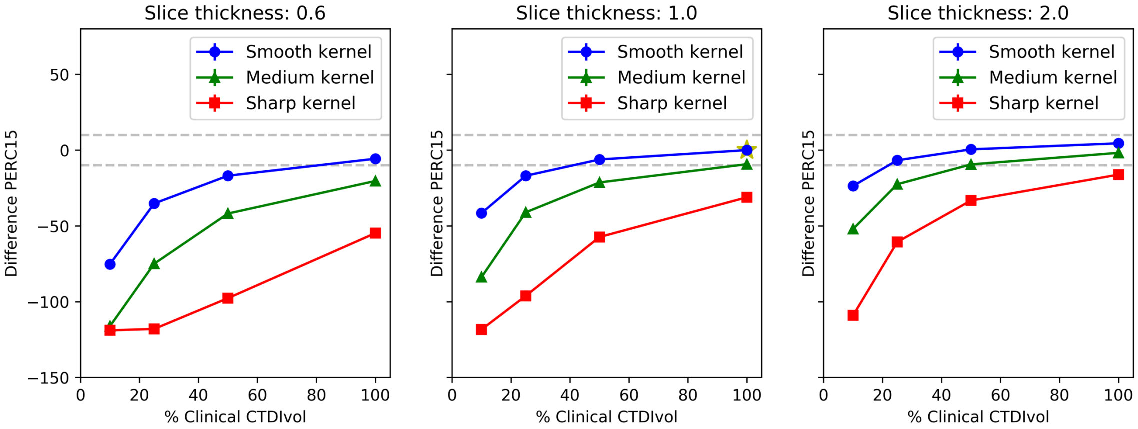


FIGURE 3-7 PLOT OF CHANGE IN PERC15 AS A FUNCTION OF PARAMETER CONFIGURATION FOR FULL STUDY POPULATION. ERROR BARS ARE SHOWN, HOWEVER ARE SMALLER THAN THE MARKER IN ALL CASES. REFERENCE CONDITION IS HIGHLIGHTED WITH A STAR.

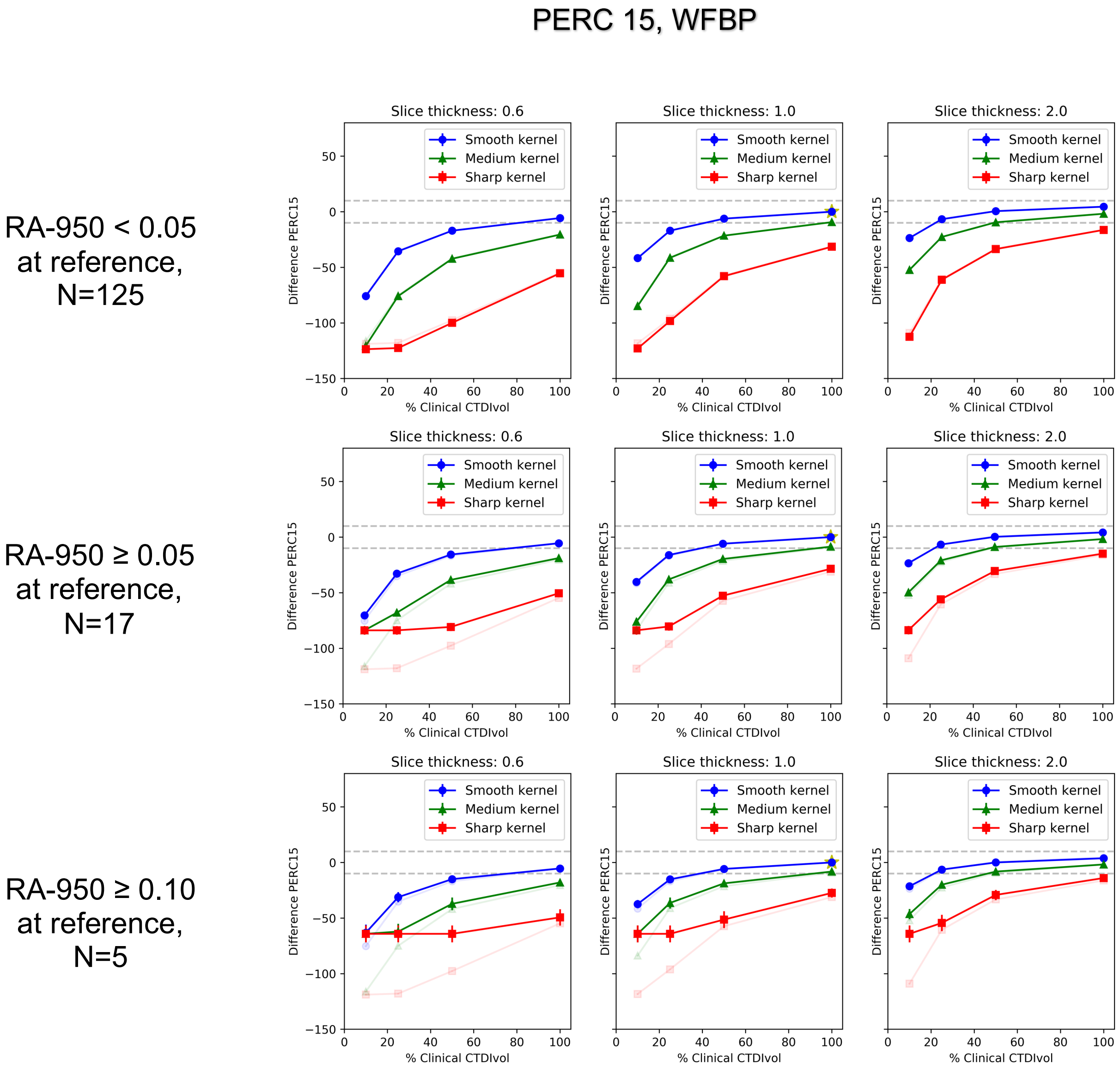


Figure 3-8: Perc15 response in subpopulations of the study cohort, grouped by amount of emphysema at reference (determined using RA-950, same groupings as Figure 3-5). Original pooled results, shown in Figure 3-7, are overlaid with reduced opacity to highlight how trends change with differing levels of emphysema at baseline.

Multiple Linear Regression Modeling of RA-950 Change

Modeling of the data was conducted in two stages: initially without interaction terms between independent variables (i.e. dose, kernel, and slice thickness), and second including interaction terms. Regression results for the no-interaction model are given in Table 3-4 and regression results for the model including interaction terms are given in Table 3-5.

While the model including interaction terms provided a slightly improved fit (R^2 of 0.812, versus 0.783 without interaction terms), and all interaction terms were found to be statistically significant, upon closer inspection the interaction terms are found to be one to two orders of magnitude less important than the individual, non-interaction variables. For example, the strongest interaction term in the model, “Slice Thickness x Kernel: Sharp,” has a coefficient of -0.033, however the sharp kernel alone has a coefficient of 0.235 and slice thickness alone has a coefficient of -0.087. The non-interaction terms are a full order of magnitude “stronger” than the interaction terms. Thus, for wFBP, the non-interaction model is the model selected for further analysis.

Table 3-4 Regression results for RA-950 with wFBP, no interaction terms. All results given are statistically significant at 95% confidence level. Note that the "dose" variable is modeled as continuous and the coefficient reflects a 1% change in dose. Thus, a 50% change in dose would reflect $50 * -0.002 =$ -0.1 $=$ -10%. Slice thickness is also modeled as a continuous variable.

Intercept (Kernel: Smooth)	0.287 (0.002)
Kernel: Medium	0.059 (0.002)
Kernel: Sharp	0.151 (0.002)
Dose	-0.002 (0.000)
Slice Thickness	-0.066 (0.001)
R-squared:	0.783
Adj. R-squared:	0.783
No. Observations:	5,112

Inspecting the individual coefficients of the non-interaction model (Table 3-4), we observed that kernel selection has the largest impact on the change in RA-950 score, in particular the selection of the sharp kernel. This agrees well with the observations regarding over-enhancing kernels found in [30] and [33]. Dose also plays a very important role. Since dose was modeled as a continuous variable however it is important to remember that the coefficient given in the table represents the expected change in RA-950 for a dose change of 1%. Thus, considering a 50% change in dose, the coefficient "adjusts" to -0.10, on the order of coefficients associated with kernel change.

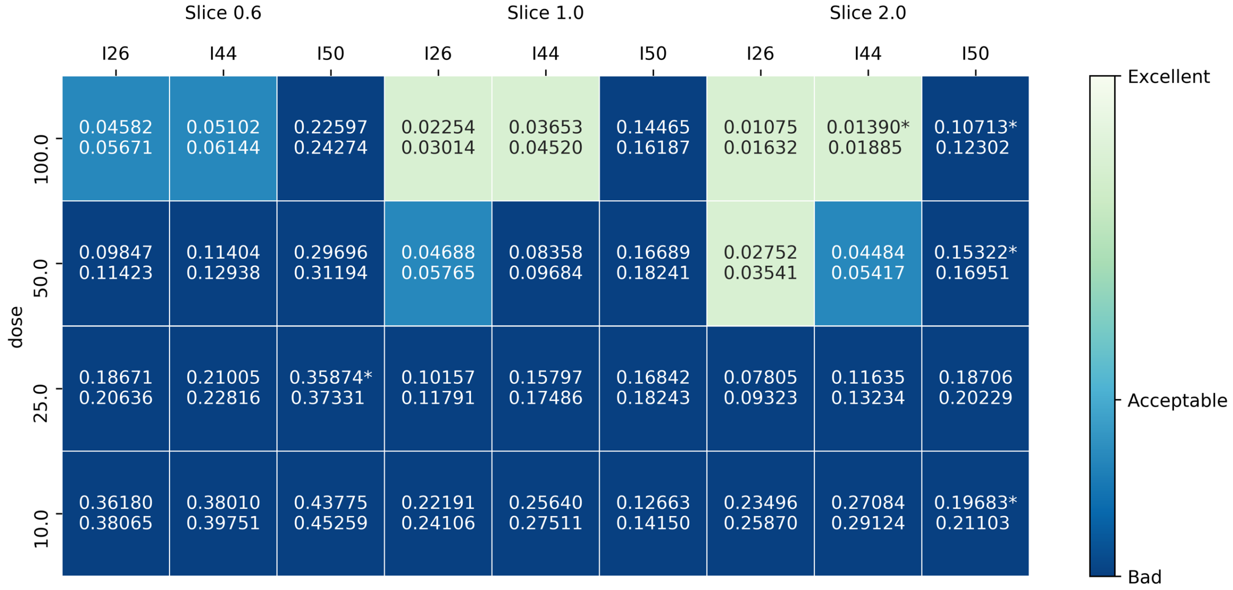
While lots of information regarding emphysema scoring with wFBP can be learned from these models, the key results are the following: no strong interactions were observed, and kernel selection proved to be the parameter having the largest impact on RA-950 change, in particular the sharp kernel. Although kernel seems to have the largest impact, dose and slice thickness play statically and clinically significant roles in the final model. In terms of implications for clinical quantitative imaging, the lack of important interaction terms with wFBP is a good result, implying that parameters can be freely adjusted without risking unexpected changes in RA-950 score (for example, if an “unacceptable” image dataset was acquired, one could thicken the slices or move to a smoother kernel to achieve an acceptable reconstruction). Unfortunately, without any post-processing, every parameter was observed to have important, large impacts on the quantitative imaging, which does not meet the criteria for an “ideal” quantitative imaging test, which only reflects the patient’s underlying disease or biology.

TABLE 3-5 REGRESSION RESULTS FOR RA-950 WITH WFBP, WITH INTERACTION TERMS. ALL RESULTS GIVEN ARE STATISTICALLY SIGNIFICANT AT 95% CONFIDENCE LEVEL. NOTE THAT THE THREE-WAY INTERACTION TERMS WERE REMOVED FROM THE MODEL TO ELIMINATE MULTICOLLINEARITY.

Intercept (Kernel: Smooth)	0.313 (0.003)
Kernel: Medium	0.115 (0.004)
Kernel: Sharp	0.235 (0.004)
Dose	-0.002 (0.000)
Slice Thickness	-0.087 (0.002)
Dose x Slice Thickness	0.001 (0.000)
Dose x Kernel: Medium	-0.001 (0.000)
Dose x Kernel: Sharp	-0.001 (0.000)
Slice Thickness x Kernel: Medium	-0.022 (0.003)
Slice Thickness x Kernel: Sharp	-0.033 (0.003)
R-squared:	0.812
Adj. R-squared:	0.811
No. Observations:	5112

3.3.2 SAFIRE Iterative Reconstruction

RA-950 & PERC15



dose	Slice 0.6			Slice 1.0			Slice 2.0		
	I26	I44	I50	I26	I44	I50	I26	I44	I50
100.0	0.04582 0.05671	0.05102 0.06144	0.22597 0.24274	0.02254 0.03014	0.03653 0.04520	0.14465 0.16187	0.01075 0.01632	0.01390* 0.01885	0.10713* 0.12302
50.0	0.09847 0.11423	0.11404 0.12938	0.29696 0.31194	0.04688 0.05765	0.08358 0.09684	0.16689 0.18241	0.02752 0.03541	0.04484 0.05417	0.15322* 0.16951
25.0	0.18671 0.20636	0.21005 0.22816	0.35874* 0.37331	0.10157 0.11791	0.15797 0.17486	0.16842 0.18243	0.07805 0.09323	0.11635 0.13234	0.18706 0.20229
10.0	0.36180 0.38065	0.38010 0.39751	0.43775 0.45259	0.22191 0.24106	0.25640 0.27511	0.12663 0.14150	0.23496 0.25870	0.27084 0.29124	0.19683* 0.21103

FIGURE 3-9 HEAT MAP OF ACCEPTABLE AND UNACCEPTABLE PARAMETER CONFIGURATIONS FOR RA-950 SCORED ON SAFIRE RECONSTRUCTIONS. NOTE THAT THE REFERENCE CONDITION IS WFBP, 100% DOSE, 1.0MM SLICE THICKNESS, AND SMOOTH RECONSTRUCTION KERNEL. (*) INDICATES ONE MISSING IMAGE DATASET (I.E. SCORES COMPUTED OVER 141 SUBJECTS RATHER THAN 142).

Iterative reconstruction typically marketed as be able to preserve image quality at reduced acquisition dose via noise and artifact reduction, however little has been done to determine whether these denoising approaches impact quantitative imaging. Given the potential to reduce noise in the images, iterative reconstruction could prove immensely valuable for quantitative imaging purposes. While iterative reconstruction approaches vary significantly, Siemens SAFIRE is one current realization that is in use clinically and represents a good candidate for comparison with wFBP. Figure 3-9 illustrates the acceptable conditions for reconstruction. If SAFIRE achieves noise reduction without imparting any other effects on the image data, one would expect to see an increase in the number of “acceptable” parameter configurations when compared to wFBP, however there are noticeably fewer acceptable conditions. The difference in behavior between the SAFIRE I50 setting and sharp wFBP kernel (as well as I26 and I44) becomes most apparent in Figure 3-10. In Figure 3-10 it can be seen that the I26, and I40 iterative settings behave somewhat more similarly to the wFBP kernels than I50 however the behavior of I50 seemingly depends strongly on both slice thickness and acquisition dose. Possible reasons for this are discussed in the next section. While there are similarities with wFBP, SAFIRE reconstruction displayed substantially different results than wFBP reconstruction in particular with regard the I50 setting, the “sharpest” iterative setting investigated.

Subgroups of patients with differing levels of emphysema at reference behaved strikingly different than the subgroups with wFBP. As shown in Figure 3-11, no clear trend in behavior is discernable with increasing levels of patient emphysema other than for subjects with increasing levels of emphysema, fewer parameter configurations fall within the region of acceptability. In particular, under the I50 SAFIRE setting with a 1.0mm slice thickness, robustness to dose appears to get substantially worse in patients with a baseline RA-950 score of ≥0.10. As with wFBP however, sample sizes in the subgroups, with the exception of the no-emphysema subgroup, were too small to make rigorous conclusions.

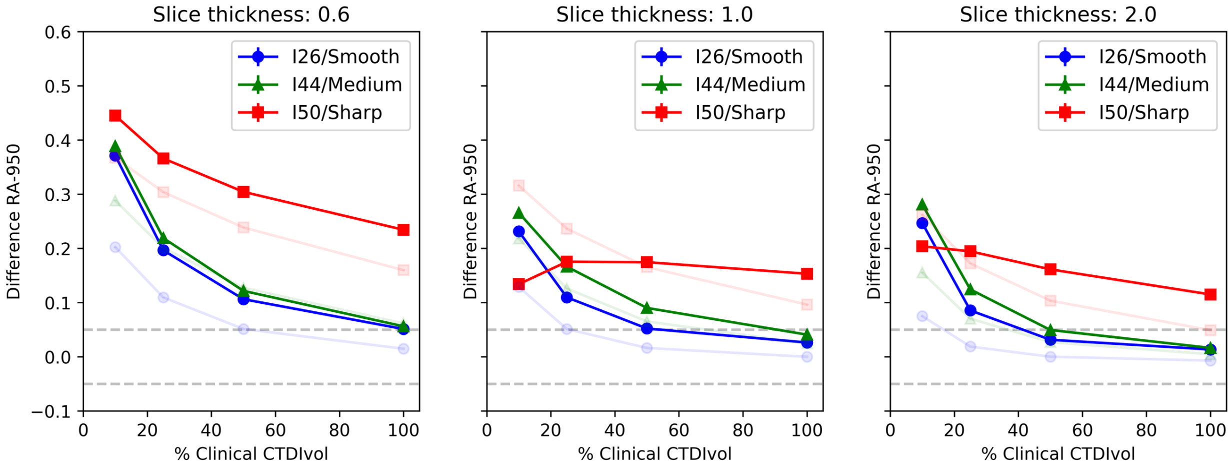


Figure 3-10 Difference in RA-950 under SAFIRE reconstruction using different parameter settings. Reference condition is WFBP reconstruction, 100% dose, 1.0 mm slice thickness and smooth reconstruction kernel. 5% threshold of acceptability is indicated with gray dashed lines. Original WFBP plots are overlaid with reduced opacity for comparison. Strikingly different behavior occurs for the I50 setting compared to all other wFBP kernels and SAFIRE settings.

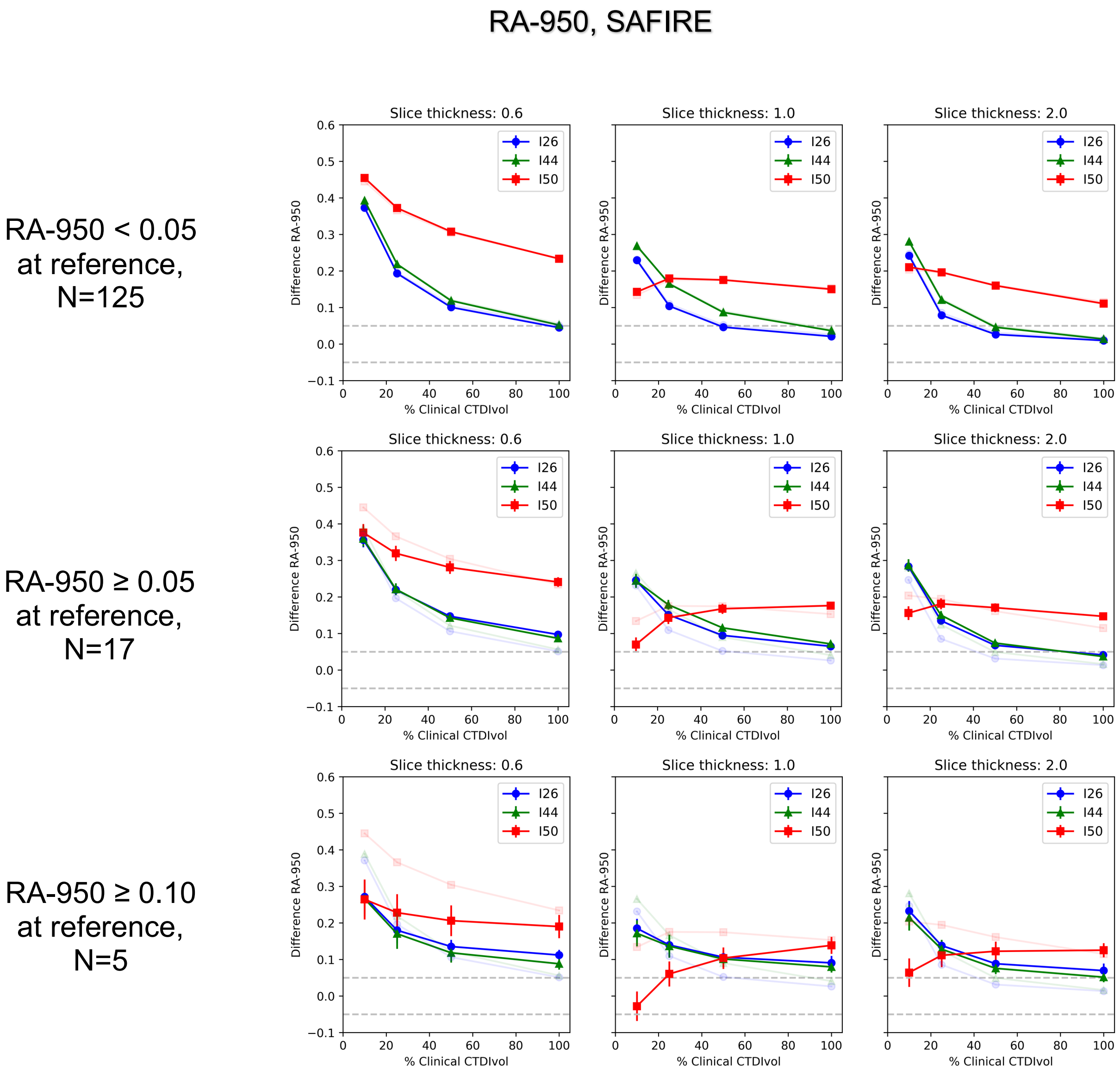


Figure 3-11: RA-950 response in subpopulations of the study cohort, grouped by amount of emphysema at reference (determined using RA-950, same groupings as Figure 3-5). Original pooled results, shown in Figure 3-10, are overlaid with reduced opacity.

While not shown here for brevity, the PERC15 results demonstrate similar trends to those observed in RA-950 scores. Parameter configurations yielding acceptable levels of PERC15 change were similar, albeit with slightly fewer configurations resulting in fully acceptable results (i.e. change ≤10HU). Changes within subgroups were equally unpredictable, seeming to vary depending on all three investigated parameters. Interestingly, PERC15 did seem to behave somewhat more similarly to results observed under wFBP in that there seemed to be less overall change in PERC15 in patients with greater levels of emphysema; RA-950 did not behave this way. For completeness, all results for PERC15 under SAFIRE are included in the appendix at the end of this chapter.

Multiple Linear Regression Modeling of RA-950 Change in SAFIRE

The non-interaction regression model, presented in

Table 3-6, resulted in an R^2 value of 0.547. The regression model including interaction terms (presented in Table 3-7) resulted in an R^2 value of 0.604 indicating a slightly improved fit. While this increase in R^2 is similar to the increase observed for wFBP, it is the relative importance of the interaction terms to the rest of the model that makes the interaction model the better choice for understanding the behavior of RA-950 when scored on the SAFIRE reconstructions.

Table 3-6 Non-interaction model regression results for RA-950 scores computed on SAFIRE reconstructions.

Intercept (Kernel: I26)	0.340 (0.003)
Kernel: I44	0.025 (0.003)
Kernel: I50	0.095 (0.003)
Dose	-0.002 (0.000)
Slice Thickness	-0.066 (0.002)
R-squared:	0.547
Adj. R-squared:	0.546
No. Observations:	5100

Table 3-7 Regression results (with interaction terms) for RA-950 calculated on images reconstructed with SAFIRE. (*) denotes a result that is **not** statistically significant. All other results are significant at the 95% confidence level.

Intercept (Kernel: I26)	0.366 (0.004)
Kernel: I44	0.036 (0.007)
Kernel: I50	0.085 (0.007)
Dose	-0.003 (0.000)
Slice Thickness	-0.087 (0.003)
Dose x Slice Thickness	0.0005 (0.000)
Dose x Kernel: I44	0.000 (0.000)
Dose x Kernel: I50	0.0014 (0.00007)
Slice Thickness x Kernel: I44	0.002 * (0.004)
Slice Thickness x Kernel: I50	-0.044 (0.004)
R-squared:	0.604
Adj. R-squared:	0.604
No. Observations:	5100

With wFBP, interaction terms were statistically significant, however they did not play a clinically significant role in the final model (i.e. coefficients were small relative to the non-interaction term). SAFIRE, however, has interaction terms that are on the same order of magnitude of the non-interaction terms. Of particular note are the interaction terms involving the I50 SAFIRE setting. One can qualitatively observe from Figure 3-10 and Figure 3-11 that change in RA-950 involving I50 depends on both the acquisition dose and slice thickness. Regression results presented in

Table 3-7 support this quantitatively. The coefficient for the interaction term between slice thickness and I50 (-0.0435) is on the same order as the non-interaction slice thickness coefficient (-0.0871) and the non-interaction I50 coefficient (-0.085). Contrasted against the same coefficients for interaction-term wFBP model (-0.033, -0.0869, and 0.2345, respectively), we see that with SAFIRE the interaction term plays nearly as large of a role in the final model as the non-interaction terms, while with wFBP the non-interaction sharp kernel coefficient dominates.

The regression analysis provides further insight into SAFIRE's impact on RA-950 scoring in that the coefficient for the interaction term for dose and I50 (0.0014) has the opposite sign from the coefficient for the non-interaction dose parameter (-0.0026). This reflects that with I50, the change in RA-950 can decrease with decreasing dose (e.g. Figure 3-10, slice thickness 1.0mm); this is the opposite effect of that observed with I26, I40, and all WFBP settings where decreasing dose results in increasing RA-950 change. Thus, the linear regression model analysis establishes and quantifies that there exist non-trivial interactions between SAFIRE setting, slice thickness, and dose when scoring emphysema with RA-950 that were not observed with wFBP.

3.4 Discussion

3.4.1 Weighted Filtered Backprojection

Several key results have emerged from this study regarding emphysema scoring robustness using both wFBP and SAFIRE. In general, parameter settings that produce a smoother image (i.e. less noise with thicker slices, higher dose, or smooth reconstruction kernel) resulted in images that were more likely to agree with the reference condition. Additionally, emphysema scoring on wFBP images did not present interactions between the parameters, although SAFIRE did. It was also observed, that with wFBP, that amount of emphysema present in a patient's lungs impacted the robustness to parameter change of the RA-950. While the sample sizes of patients with emphysema were too small to establish rigorous conclusions, this result was also observed by Gierada et al. [35] and can clearly be seen emerging Figure 3-5. For wFBP, we found that the sharp kernel produced significant change in the RA-950 when compared to reference, and none of the configurations yielded results that were within 5% of the reference value. This agrees well with the findings of Boedeker et al. [30] and Gierada et al. [35] who both identified statistically significant changes as a result of sharp or over-enhancing kernel utilization. Smooth and medium kernels result in some conditions that produce results close the reference value.

Finally, because of the systematic investigation of three parameters concurrently enabled by the pipeline, the linear regression analysis established that there are not typically interactions between wFBP parameters and RA-950, meaning that these parameters can be considered independently of one another in terms of their effects on quantitative emphysema scoring. This lack of complex interactions between parameters has implications for clinical usage of RA-950 scoring that will be discussed in more depth in Chapter 5. This is a new result that has not previously been established in the literature. This was not the case for SAFIRE, discussed below.

PERC15 results in wFBP, in terms of regions of acceptability and impacts on more or less emphysematous patients (i.e. the subgroup analysis), were similar to the results found for RA-950. PERC15 under wFBP reconstructions decreases with increasing image noise, the opposite of RA-950, which increases with increasing noise. Slightly fewer parameter configurations resulted in "acceptable" levels of change for PERC15. This could possibly be attributed to the definition of "acceptable" variation utilized, in that it is difficult to establish that a ±10HU change in PERC15 score is exactly equivalent to a ±0.05 change in RA-950 score; however, given the similarities between the acceptable parameter configurations, and the lack of studies correlating PERC15 with other emphysema metrics (e.g. GOLD status, BODE index, etc.) it represents a good starting point. A more in-depth discussion of defining "acceptable" variation is given below.

3.4.2 SAFIRE Iterative Reconstruction

SAFIRE reconstruction had fewer parameter configurations that resulted in acceptable levels of change for both RA-950 and PERC15. This is somewhat surprising given that iterative reconstruction is generally thought to result in lower image noise at a given dose than wFBP, and lower image noise is thought to produce more "stable" quantitative emphysema scores. This effect was not observed in this study. In addition to fewer acceptable parameter configurations, SAFIRE also displayed both clinically and statistically significant interactions between the I50 setting and both dose and slice thickness. I26 and I44 were also more strongly affected by dose reduction than the wFBP smooth and medium reconstruction kernels (Figure 3-10). While other studies have observed that iterative approaches to reconstruction may represent a means to improve the reliability of quantitative emphysema scoring (e.g. [37][2]) these results suggest that iterative approaches may actually be less reliable for quantitative emphysema scoring than wFBP.

Amount of emphysema in a patient was also observed to have an impact on the robustness of emphysema scoring approaches for SAFIRE, however no clear trends emerged to help understand how these patients are impacted. With I50, amount of emphysema seemed to cause more variation (i.e. less robustness). I26 and I44 had less variation, however the 100% dose scores moved further from the acceptable range of RA-950 change, while RA-950 scores at lower doses moved closer to the acceptable threshold.

While this study did not directly investigate why this complex behavior occurs with SAFIRE, one possible explanation for the difficult-to-predict behavior with regard to quantitative emphysema has to do with the nature of iterative reconstruction algorithms. Siemens SAFIRE algorithm is a hybrid combination an image-domain based iterative

[2] Investigated RA-950 scoring under dose reduction with a similar study design, however using Toshiba's AIDR 3D reconstruction, and found that AIDR 3D dramatically improved robustness

denoising approach, and an iterative projection domain scheme to achieve some level of artifact reduction [113]. Depending on selection of image denoising approach, optimizer, and cost function, different image features can be enhanced (such as edges, or contrast, or noise reduction). It is conceivable that SAFIRE, in particular the I50 setting, is optimized to enhance local contrast differences to assist human readers, whose visual system is much more sensitive to contrast than absolute gray values. However, local contrast enhancement would likely confound quantitative imaging approaches that rely on predictable density values, such as RA-950 scoring, since regions of emphysema could have different HU values depending on the surrounding area.

Finally, while 10% dose levels (~0.2mGy) were investigated and scored quantitatively, in nearly all cases, the image quality is poor enough that these images would likely not be considered usable for any clinical purpose.

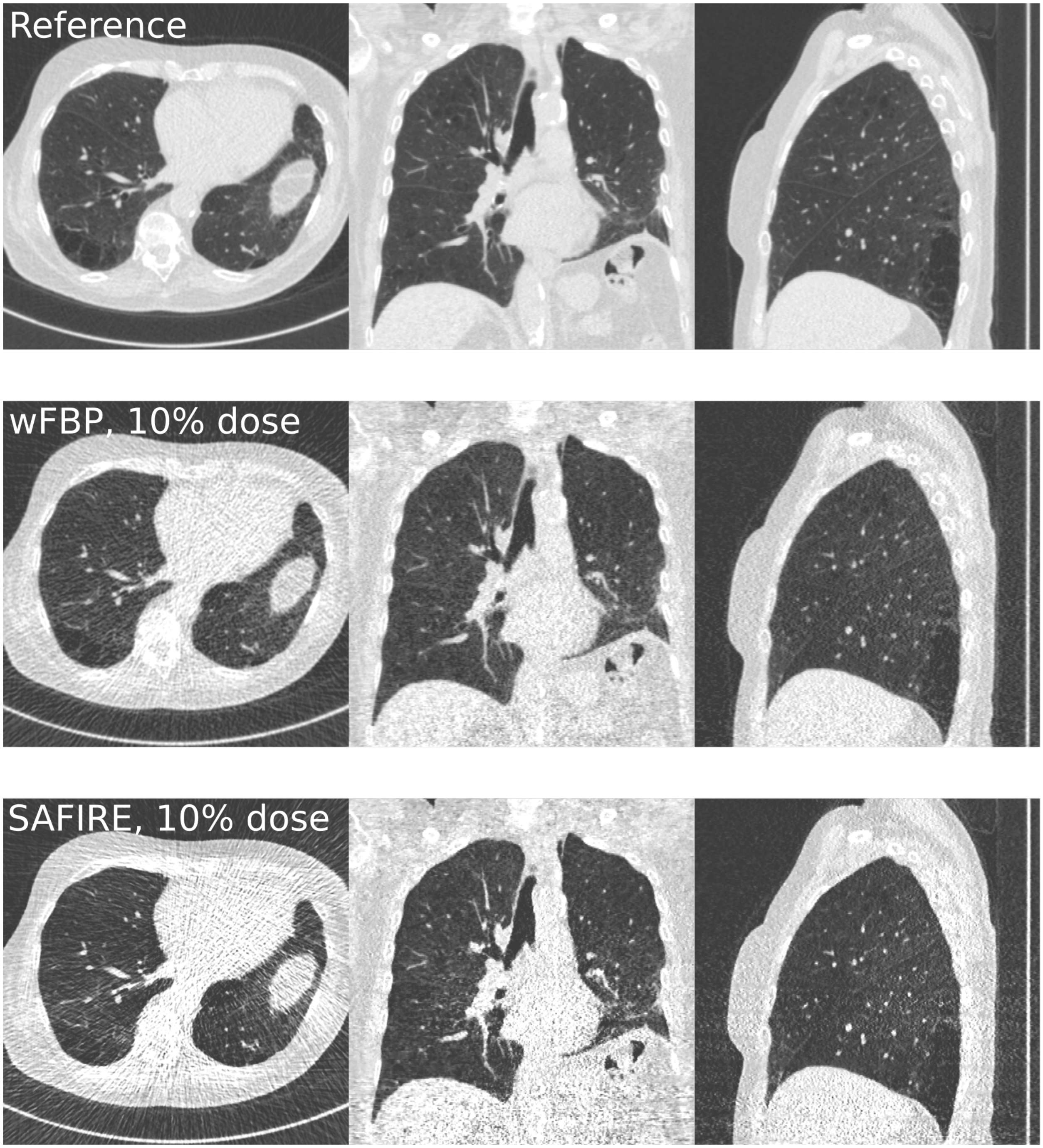


Figure 3-12 Sample reconstructions using the 10% dose setting. Top row: illustrates the reference condition (wFBP, 100% dose, smooth reconstruction kernel, 1.0mm slice thickness). middle row: wFBP, 10% dose, smooth reconstruction kernel, 1.0mm slice thickness. Bottom row: SAFIRE, 10% dose, I26 sharpness setting, 1.0mm slice thickness.

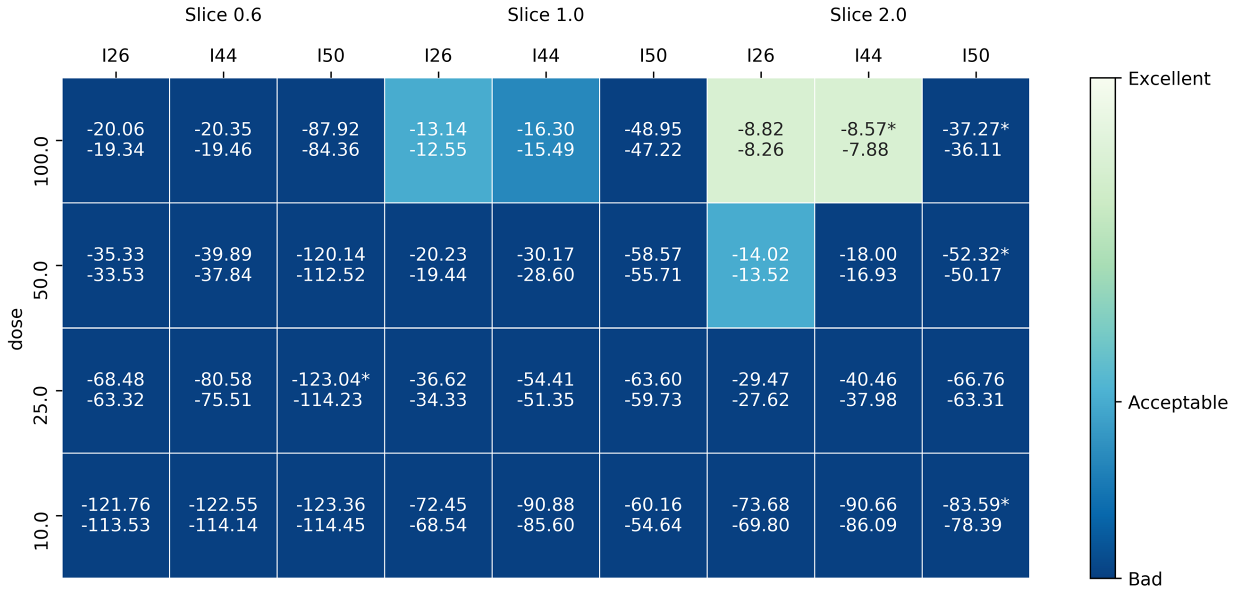
3.5 Conclusions

For the purposes of reliable, robust clinical quantitative RA-950 and PERC15 scoring on lung screening studies, this work supports wFBP as the better choice of reconstruction algorithm. WFBP presented a wider range of acceptable parameter configurations than SAFIRE, more predictable behavior over the range of parameters investigated, and no complex interactions between the investigated parameters: slice thickness, kernel/sharpness selection, and acquisition dose. While this is not to say that all iterative reconstruction approaches should be excluded from quantitative imaging, it highlights one of the difficulties with the broad umbrella term of "iterative reconstruction" in that it rarely clarifies the underpinnings of the algorithm. Some studies have found that "iterative" approaches improved quantitative emphysema scoring results (e.g. [37]) while others, like this study, highlight that care should be taken when attempting quantitation of a CT scan reconstructed with an "iterative" algorithm (e.g. [39]). This conclusion to use wFBP over iterative approaches agrees with the protocol recommendations given for SPIROMICS [42].

With wFBP, there were no major differences between the robustness of RA-950 and PERC15. Other studies have found that PERC15 may be slightly more repeatable for the purposes of longitudinal studies [23], [24] and these results do not disagree with that conclusion (i.e. this was not a repeatability study). For this study, based on the behavior of RA-950 and PERC15 under different parameter configurations, as well as within subgroups of different emphysema levels, it is likely that both scoring approaches are fundamentally extremely similar and could likely be used interchangeably when wFBP is utilized as the reconstruction approach. RA-950 however has the added benefit of localizing pockets of emphysema for visual review, and a recently proposed “normalization” scheme [49] along with other image post-processing techniques may make it more robust and reliable in the long term. This will be addressed further in the next chapter.

Finally, in terms of direct recommendations for clinical use of quantitative emphysema scoring in a lung screening population: quantitative emphysema scoring can be done. However, care should still be taken to ensure that an acceptable protocol been used. Weighted filtered backprojection should be utilized, quantitative scoring should not be performed with sharp reconstruction kernels (no acceptable configurations were found), and scoring should not be conducted directly on 0.6mm slices (only one acceptable parameter configuration). Emphysema scoring using RA-950 and/or PERC15 should be reasonably robust down to 50% of the clinically recommended dose (~1.0mGy for a standard-size patient) when a smooth or a medium reconstruction kernel (roughly Siemens B10, B45 or between) is utilized, and at least a 1.0mm slice thickness is used.

Appendix: Complete PERC15 Results for SAFIRE



	Slice 0.6			Slice 1.0			Slice 2.0		
dose	I26	I44	I50	I26	I44	I50	I26	I44	I50
100.0	-20.06 -19.34	-20.35 -19.46	-87.92 -84.36	-13.14 -12.55	-16.30 -15.49	-48.95 -47.22	-8.82 -8.26	-8.57* -7.88	-37.27* -36.11
50.0	-35.33 -33.53	-39.89 -37.84	-120.14 -112.52	-20.23 -19.44	-30.17 -28.60	-58.57 -55.71	-14.02 -13.52	-18.00 -16.93	-52.32* -50.17
25.0	-68.48 -63.32	-80.58 -75.51	-123.04* -114.23	-36.62 -34.33	-54.41 -51.35	-63.60 -59.73	-29.47 -27.62	-40.46 -37.98	-66.76 -63.31
10.0	-121.76 -113.53	-122.55 -114.14	-123.36 -114.45	-72.45 -68.54	-90.88 -85.60	-60.16 -54.64	-73.68 -69.80	-90.66 -86.09	-83.59* -78.39

FIGURE 3-13 HEAT MAP OF ACCEPTABLE PARAMETER CONFIGURATIONS FOR PERC15 UNDER SAFIRE RECONSTRUCTION. VERY FEW CONFIGURATIONS RESULT IN ACCEPTABLE LEVELS OF CHANGE FROM THE REFERENCE VALUE, HOWEVER THE ONES THAT DO ARE THE SAME AS THOSE FOUND IN RA-950 UNDER SAFIRE (FIGURE 3-9). THIS IS SIMILAR TO WHAT OCCURRED IN WITH WFBP RECONSTRUCTIONS. (*) INDICATES ONE MISSING IMAGE DATASET (I.E. SCORES COMPUTED OVER 141 SUBJECTS RATHER THAN 142).

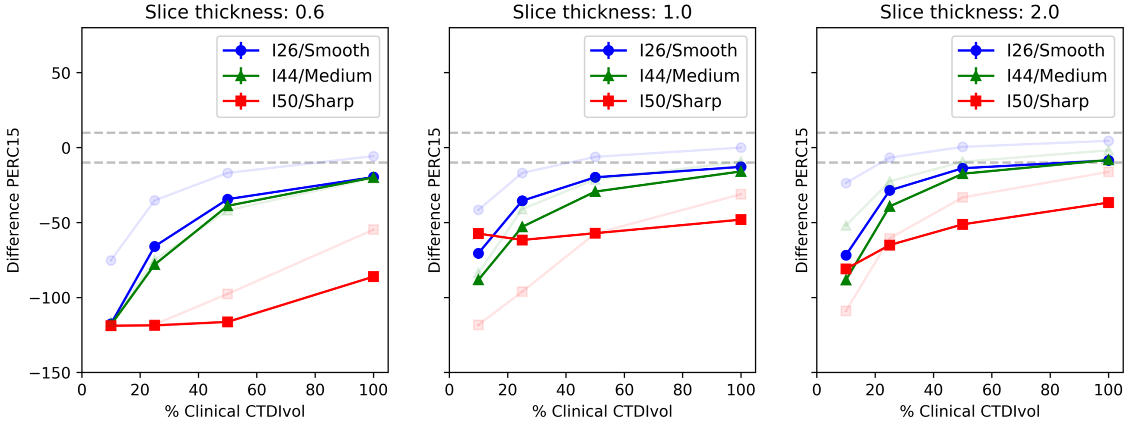


FIGURE 3-14 DIFFERENCE IN PERC15 UNDER SAFIRE RECONSTRUCTION USING DIFFERENT PARAMETER SETTINGS. REFERENCE CONDITION IS WFBP RECONSTRUCTION, 100% DOSE, 1.0 MM SLICE THICKNESS AND SMOOTH RECONSTRUCTION KERNEL. 5% THRESHOLD OF ACCEPTABILITY IS INDICATED WITH GRAY DASHED LINES. ORIGINAL WFBP PLOTS FOR PERC15 (FIGURE 3-7) ARE OVERLAID WITH REDUCED OPACITY FOR COMPARISON. STRIKINGLY DIFFERENT BEHAVIOR OCCURS FOR THE I50 SETTING COMPARED TO ALL OTHER WFBP KERNELS AND SAFIRE SETTINGS.

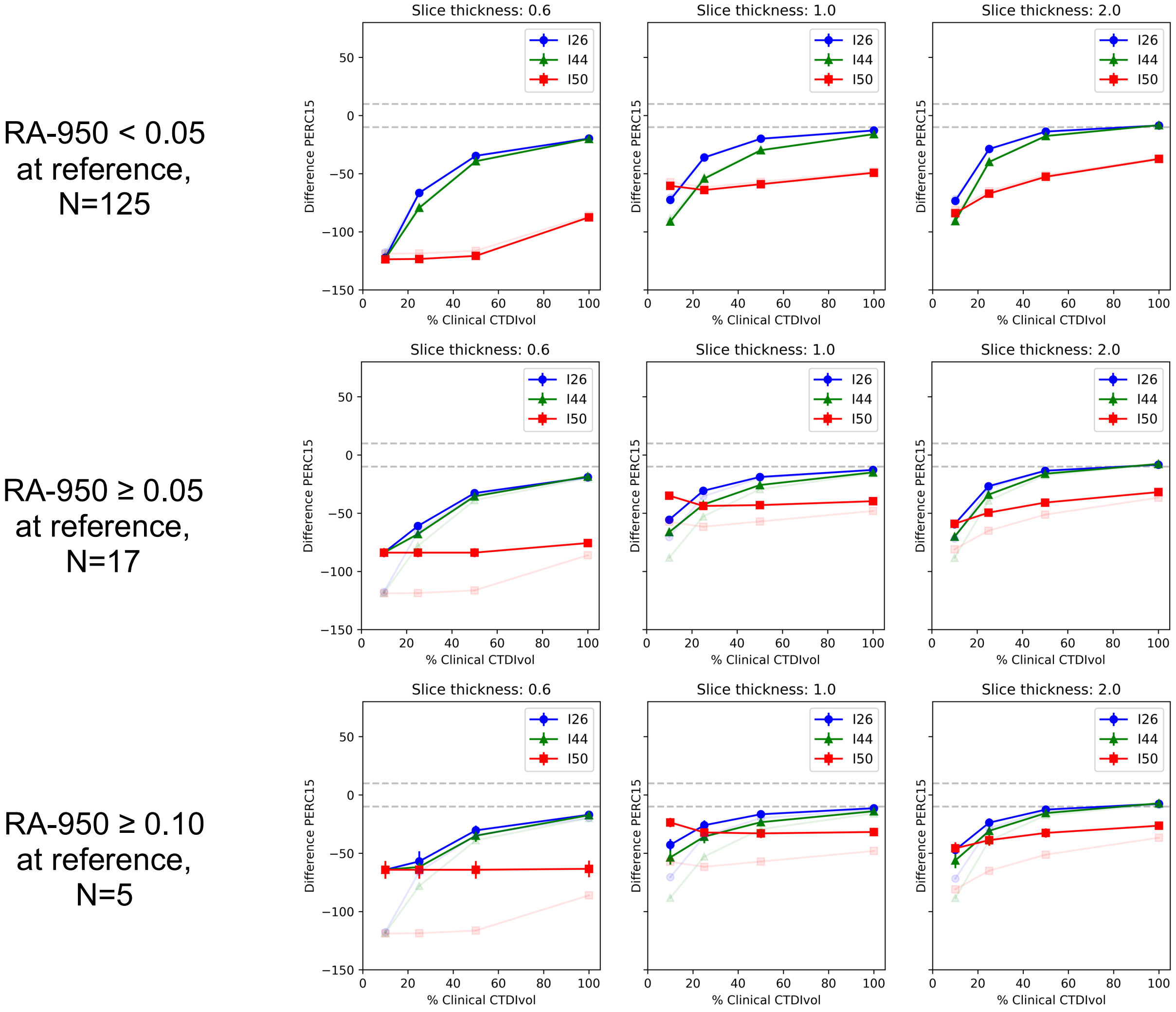


FIGURE 3-15: PERC15 RESPONSE IN SUBPOPULATIONS OF THE STUDY COHORT, GROUPED BY AMOUNT OF EMPHYSEMA AT REFERENCE (DETERMINED USING RA-950, SAME GROUPINGS AS FIGURE 3-5). ORIGINAL POOLED RESULTS, SHOWN IN FIGURE 3-14, ARE OVERLAID WITH REDUCED OPACITY.

Chapter 4 – Impacts of Adaptive Denoising Using Bilateral Filtering on Quantitative Emphysema Scoring Under a Broad Range of Reconstruction and Acquisition Conditions

4.1 Introduction

In Chapter 3 it was found that there were regions of CT parameter space, clustered around the reference reconstruction, that would result in limited amounts of change in RA-950 scoring and PERC15 scoring. Although the regions were relatively small, some clear guidance could be provided regarding the proper acquisition of a scan intended for quantitative emphysema scoring: wFBP with smooth or medium reconstruction kernels, with at least 1.0mm slice thickness should be utilized, assuming CTDIvols do not go below roughly 1.0mGy. It was generally found that parameter configurations that resulted in less image noise than the reference configuration were more likely to produce a score within the limits of "acceptability." This evaluation allows for a researcher or clinician to decide if the scan is acceptable for quantitative emphysema scoring based on the acquisition and reconstruction protocols, however does not provide many options if it is found that the scan and/or reconstruction is unacceptable for emphysema scoring.

If access to the raw data is still possible, the scan could be reconstructed using an acceptable protocol, however often this is typically not the case since raw data is usually cleared from the scanner after a short period of time (approximately one week at our institution). This leaves two options if quantitative imaging is required: redo the study, or perform image post-processing to "restore" the data to an acceptable state. Since performing CT studies multiple times is generally not possible (i.e. because of dose concerns), in the absence of further reconstructions directly from the raw data, post-processing of the image data may be the only viable pathway towards reliable quantitative emphysema scoring.

Surprisingly little work has been conducted investigating post-processing of the image data, however all have yielded promising results. Early studies investigated simple approaches, namely Gaussian filtration[3] [46] and slice averaging for the purposes of noise reduction [114]. Both approaches resulted in improved quantification of emphysema, with some limitations: the research end point for [114] was a plot of RA-950 score as a function of slice location, which is not generally utilized today making it difficult to compare (although promising results are presented showing clear differences in said plots between different GOLD status patients). Tylen et al. [46] improved separation of patients with and without emphysema, however still observed overlap between the two groups, suggesting some potential limitations of the proposed method. Further denoising approaches have been proposed such as non-linear "local noise-weighted" filtering which tunes the strength of the image filter depending on an estimate on local image variance for the voxel being denoised [115], and a very recently proposed machine-learning based "convolutional auto-encoder" for artifact and noise reduction [116]. Despite demonstrating more accurate, robust quantification of emphysema, a key limitation of these approaches is complexity, with each method requiring a number of different tuning parameters and/or training cases.

[3] The authors also propose several additional corrections to the image data, including a correction for "fluid pooling" and possible lung motion; the core of the denoising approach however lies in Gaussian blurring.

An extremely promising “image normalization” approach has been proposed by Gallardo-Estrella et al. [49]. This approach utilizes slice averaging to reach a 3.0mm, non-overlapping scan, filtering of the masked bullae (pockets of emphysema) to exclude any below 5mm^2 which are likely to be noise, and finally a Fourier domain-based kernel normalization scheme [48] to make images appear as if they were reconstructed using a Siemens B31f reconstruction kernel. As demonstrated on the NLST dataset, this approach dramatically improved the performance of emphysema scoring as a biomarker for mortality [49]. While the effect of the normalization process is similar to denoising, this approach is fundamentally limited. For instance, while this approach may work to contend with a scan incorrectly acquired using an over-enhancing/sharp reconstruction kernel or a very thin slice, once normalization is applied no further processing can be provided by the normalization scheme. Additionally, the kernel normalization process has been rigorously tested only in cohorts coming from well-controlled clinical trials (COPDGene and NLST) and only under three kernel configurations (Siemens B45f, GE “Standard” and “Bone”) and it remains to be seen if it will work effectively in a broader clinical cohort. Thus, denoising can potentially provide an alternative, perhaps complimentary approach to image post-processing.

In this chapter, we explore an existing image denoising algorithm, namely bilateral filtering, and its impact on the robustness of emphysema scoring using RA-950 and PERC15. Bilateral filtering is a non-linear, edge-preserving, approach to denoising [117] that has proven to be popular in many image-processing applications [118], however only recently has begun being explored in CT imaging [119]–[121]. One of the key benefits to the bilateral filter is its relative simplicity compared to the denoising algorithms described above [117].

The bilateral filter is formulated in the following manner:

$$I_{filtered}(\vec{x}) = \frac{1}{W_p}\sum_{\overrightarrow{x_i}\in\Omega} I(\vec{x}_i)\, f_r\big(I(\vec{x}_i) - I(\vec{x})\big) g_s(\|\vec{x}_i - \vec{x}\|)$$

where x is the pixel location being filtered, $I_{filtered}$ is the filtered pixel value, and Ω is a neighborhood of pixels, x_i, around the pixel being denoised. f_r and g_s are the “range” and “spatial” filter functions respectively. Finally, W_p is a normalization term:

$$W_p = \sum_{\overrightarrow{x_i}\in\Omega} f_r\big(I(\vec{x}_i) - I(\vec{x})\big) g_s(\|\vec{x}_i - \vec{x}\|)$$

While the range and spatial filter functions are arbitrary in the definition of the bilateral filter, a common choice is Gaussian functions for both, which is utilized in this work. A full treatment of the mathematical details of the bilateral filter is outside of the scope of this work, however the assumption and intuition underpinning this approach are that voxels that are close to one another in intensity value are likely to be from the same underlying material. Therefore, any slight differences observed are likely to be noise and should be smoothed (thus, small intensity differences result in a larger weight). Large differences are more likely to be different materials, such as the intensity difference between lung parenchyma and lung wall, and therefore should be preserved, i.e. lower weight in the bilateral filter. Bilateral filtering using Gaussian filter functions essentially accomplishes Gaussian filtering in relatively homogeneous regions, while avoiding Gaussian blurring across strong edges, giving the bilateral filter edge-preserving properties. The amount of edge-preserving versus Gaussian blurring can be modified via the tuning of several parameters, and by adjusting these parameters based on expected increases or decreases in noise, the bilateral filter can be made “adaptive.” These parameters as well as the settings utilized for them are discussed in the next section.

We theorize that the application of denoising using bilateral filtering prior to the scoring of emphysema will improve the robustness of the RA-950 and PERC15 metrics across the range of parameters investigated. Bilateral filtering is applied to the datasets developed in Chapter 3, and the same quantitative analysis is performed as described in Chapter 3, section 3.2 Methods, including an analysis of “acceptable” parameter configurations, as well as linear regression analysis to determine the impacts of individual parameters on the final quantitative emphysema result and any potential interactions between parameters.

4.2 Methods

The experiment performed in this chapter is methodologically identical the experiment performed in Chapter 3, as is the image data and study cohort, however with the application of bilateral filtering to denoise the image data prior to emphysema scoring and analysis. For information regarding the cohort, reconstructions, and quantitative analyses performed, readers are referred to section 3.2 Methods on page 96.

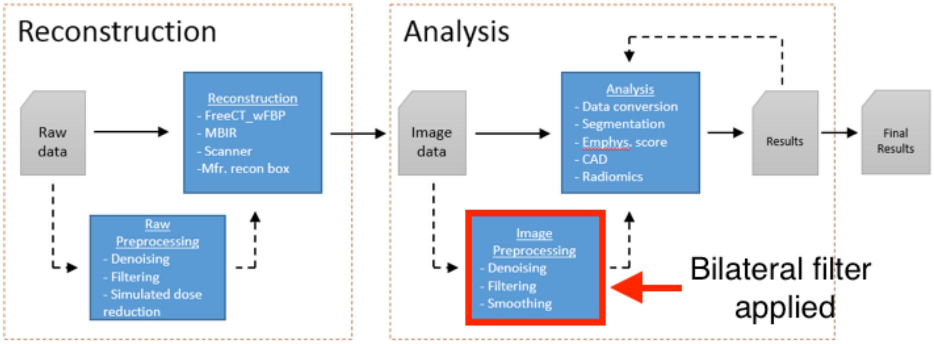


FIGURE 4-1 BLOCK DIAGRAM HIGHLIGHTING WHEN DENOISING WAS APPLIED IN THE PIPELINE WORKFLOW.

Bilateral filtering was applied after reconstruction, before analysis, as indicated in Figure 4-1. Filtering was conducted in three dimensions on the volumetric image datasets. The choice of Gaussian functions presents three tuning parameters: σ_r, the standard deviation of the range filter, σ_s, the standard deviation of the spatial filter, and w, the filter window width. For this work, the spatial filter standard deviation was fixed at $\sigma_s = 1$ for all parameter configurations. The window width was set to 5, with the window centered on the voxel being filtered. Finally, the standard deviation of the range filter, σ_r, was adjusted depending on dose and slice thickness according to the following heuristic formula:

$$\sigma_r(d_{test}, s_{test}) = \sqrt{2}^{\left(\frac{d_{ref}}{d_{test}}\right)*\left(\frac{s_{ref}}{s_{test}}\right)-1}$$

where d_{ref} and d_{test} are the reference dose and the test dose (reference is always 100%), and s_{ref} and s_{test} are the reference slice thickness (1.0mm) and the test slice thickness. Adjustments were not made for the reconstruction kernel/sharpness setting. Thus, with the reference condition, $\sigma_r = 1$. A full list of the different values employed for σ_r can be found in Table 4-1. This heuristic formula causes the bilateral filter to favor edge preserving in images where the noise is lower (i.e. higher dose, thick slices), but filter noisier parameter settings more aggressively at the expense of edge-preserving. The larger values of 22.627 and 228.070 cause the bilateral filter to behave essentially as a Gaussian blur. The bilateral filter was implemented in MATLAB (v2014a).

Table 4-1 Values for the standard deviation of the range filter (σ_r) as a function of dose and slice thickness.

		Dose (%)			
		100	50	25	10
Slice Thickness (mm)	2.0	0.841	1.000	1.414	4.000
	1.0	1.000	1.414	2.828	22.627
	0.6	1.260	2.245	7.127	228.070

4.3 Results

Example images of the denoising results can be found in Figure 4-2 (page 141) and Figure 4-9 (page 149). Enlarged version of some key examples are provided in Figure 4-3 and Figure 4-4. Qualitatively, bilateral filtering worked well to remove noise, in particular in the low dose cases, where the pocket of low-attenuation became extremely difficult to resolve without denoising (e.g. Figure 4-2, 10% dose, 0.6mm, sharp kernel). Visually, denoising seems slightly more effective at removing noise in the wFBP than the SAFIRE images. While it is not entirely clear the reason for this, one possible reason is the apparent increase in noise “structure” in the SAFIRE images (i.e. streaks, rather than simply image “graininess”); this will be discussed later in the discussion section of this chapter. Results for PERC15, in terms of trends and number of acceptable parameter configurations, was very similar to RA-950. For this reason, results for PERC15 are presented in their entirety in the appendix at the end of this chapter, and this chapter will primarily discuss RA-950.

4.3.1 WFBP

With regard to RA-950 scoring, denoising using bilateral filtering had a significant impact on the number of parameter configurations producing acceptable results. Indeed, as can be seen in Figure 4-5, for the pooled results, *every* parameter configuration evaluated resulted in an acceptable amount of change, in stark contrast to the unfiltered wFBP results where only a small subset resulted in acceptable levels of change. Line plots of the pooled results for the full study population are presented in Figure 4-6. For the subgroups of patients with emphysema and without emphysema, results are given in Figure 4-7. While results are still extremely promising, a more complex picture of denoising's impact on RA-950 scores emerges. It can be seen that patients with more emphysema typically undergo a systematic decrease in their RA-950 scores when bilateral filtering is applied, although in most cases this change still falls within the ±5% threshold of acceptability, and overall the robustness of RA-950 scoring is still dramatically improved with bilateral filtering. Possible reasons for the systematic decrease will be discussed in depth in the discussion section.

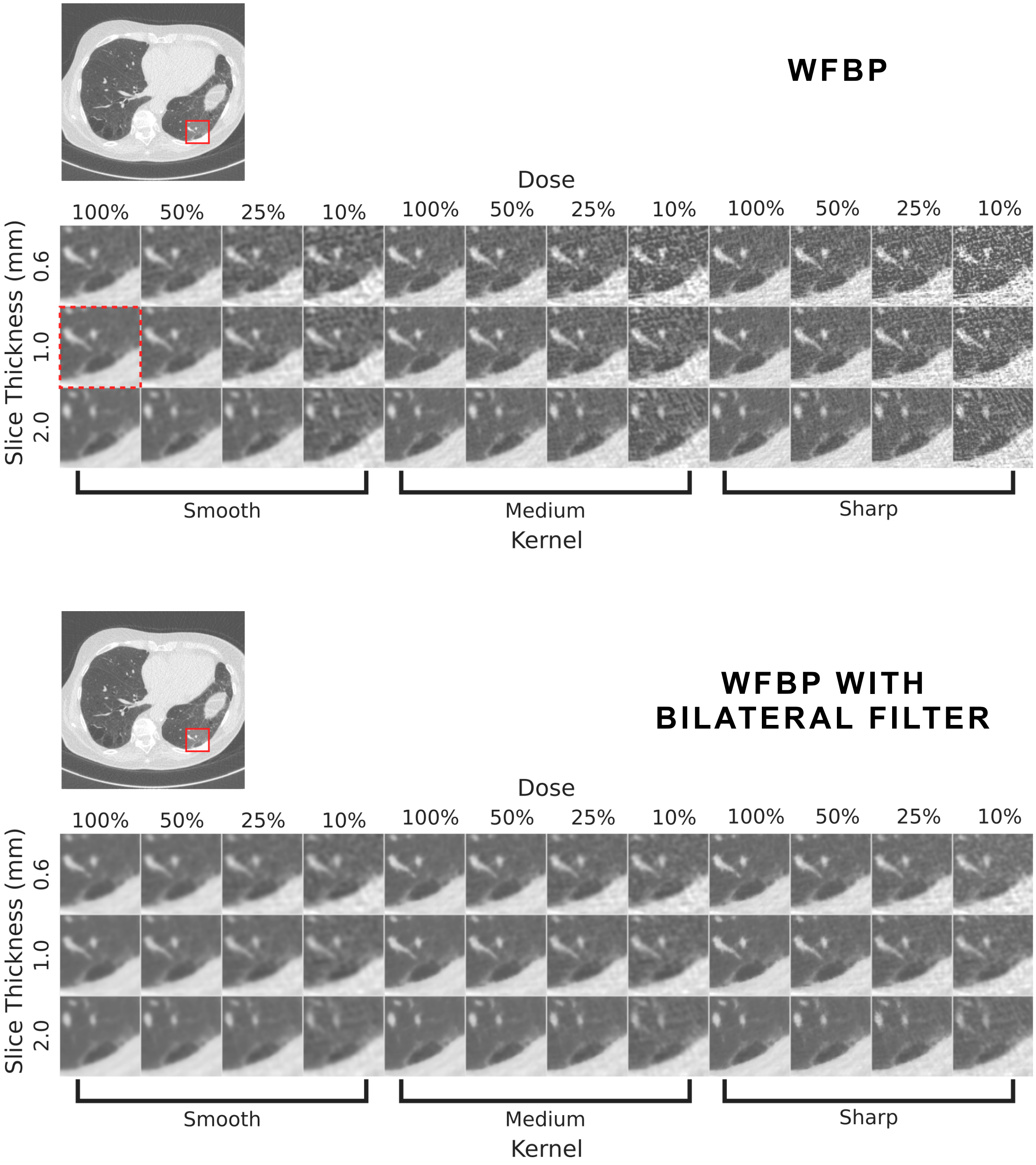


Figure 4-2 WFBP denoising results with bilateral filter (bottom row) presented with original wFBP images (top row). All images shown with window/level of 1600/-600. Reference reconstruction is outlined with a red, dashed line.

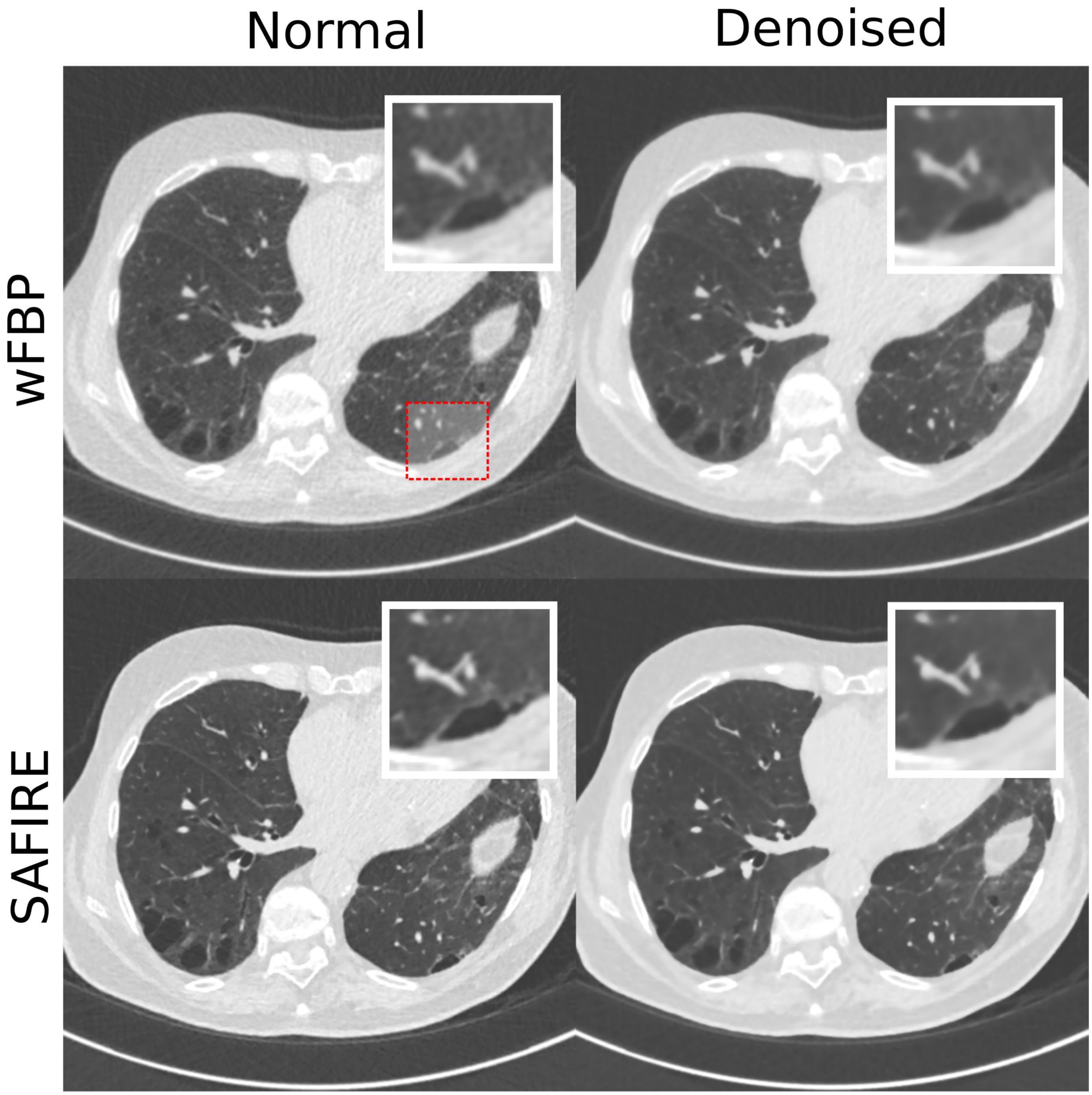


Figure 4-3 Enlarged ROIs illustrating bilateral filtering's effect on **100% dose** conditions. Top left: reference condition (wFBP, smooth kernel, 1.0mm slice thickness) *without* denoising. Top right: reference condition *with* bilateral filtering applied. Bottom left: SAFIRE, I26 setting, 1.0mm slice thickness, *without* denoising. Bottom right: SAFIRE, I26, 1.0mm *with* denoising applied

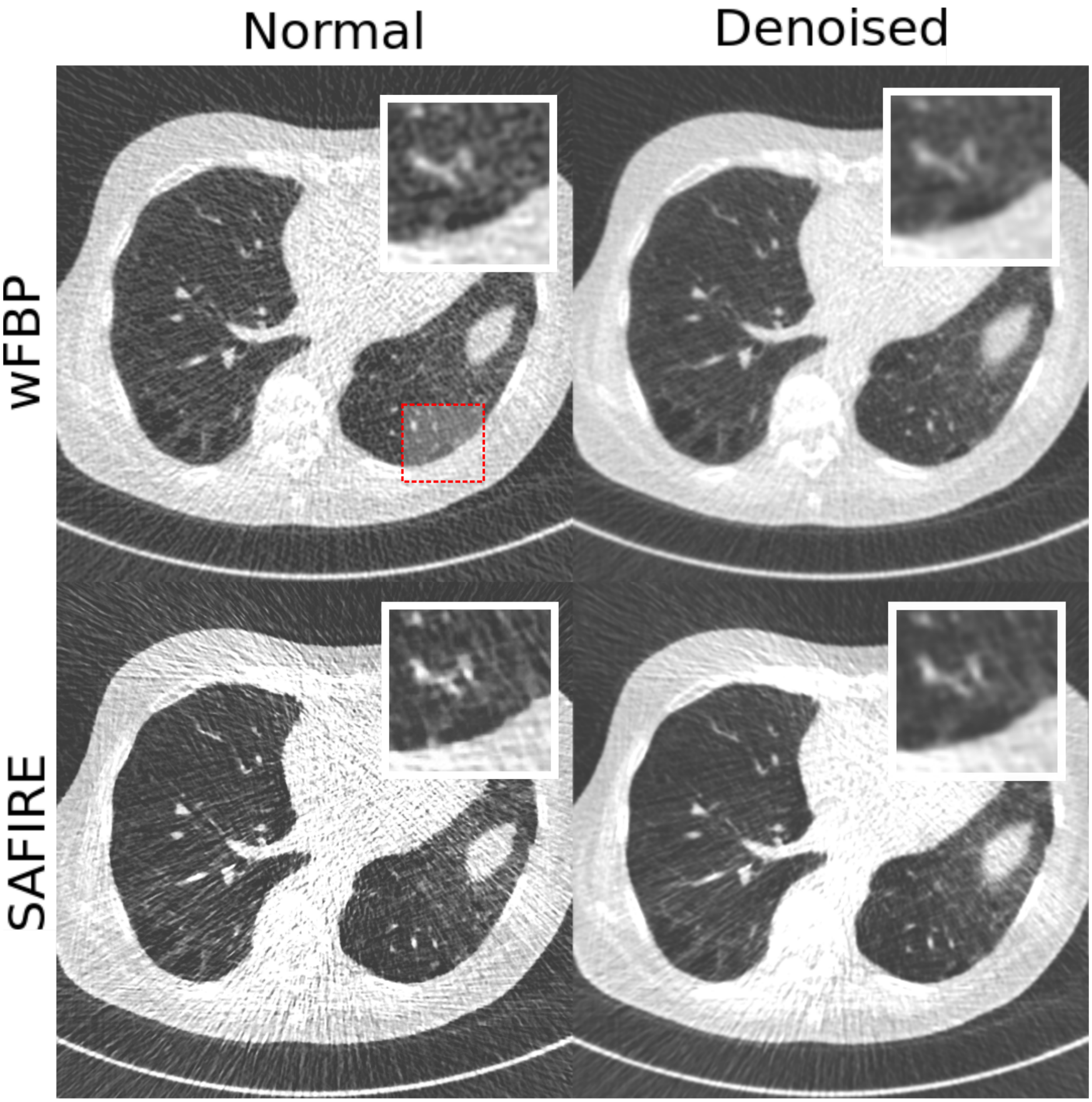
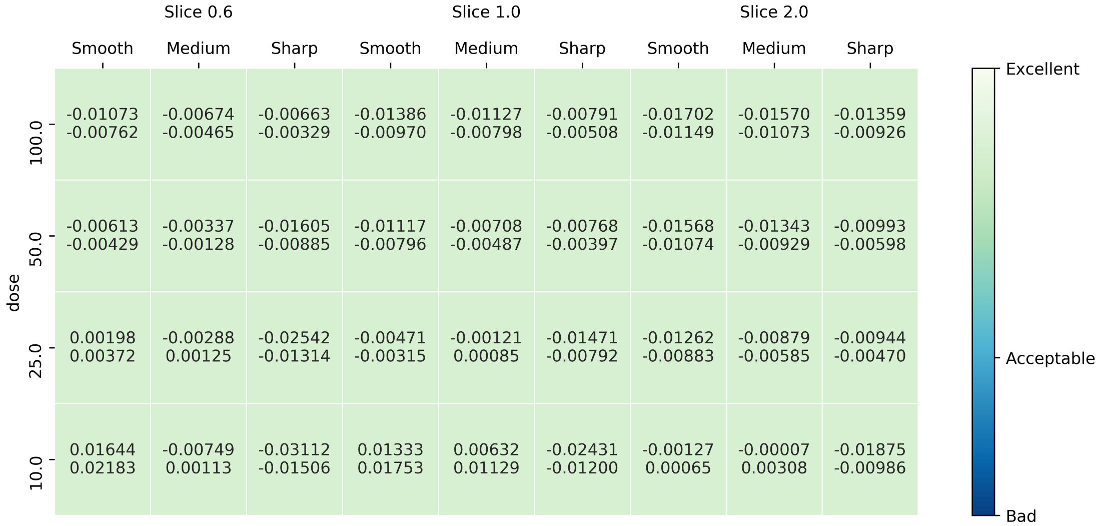


FIGURE 4-4 ENLARGED ROIS ILLUSTRATING BILATERAL FILTERING'S EFFECT ON **10% DOSE** CONDITIONS. TOP LEFT: REFERENCE CONDITION (WFBP, SMOOTH KERNEL, 1.0MM SLICE THICKNESS) *WITHOUT* DENOISING. TOP RIGHT: REFERENCE CONDITION *WITH* BILATERAL FILTERING APPLIED. BOTTOM LEFT: SAFIRE, I26 SETTING, 1.0MM SLICE THICKNESS, *WITHOUT* DENOISING. BOTTOM RIGHT: SAFIRE, I26, 1.0MM *WITH* DENOISING APPLIED



	Slice 0.6			Slice 1.0			Slice 2.0		
dose	Smooth	Medium	Sharp	Smooth	Medium	Sharp	Smooth	Medium	Sharp
100.0	-0.01073 -0.00762	-0.00674 -0.00465	-0.00663 -0.00329	-0.01386 -0.00970	-0.01127 -0.00798	-0.00791 -0.00508	-0.01702 -0.01149	-0.01570 -0.01073	-0.01359 -0.00926
50.0	-0.00613 -0.00429	-0.00337 -0.00128	-0.01605 -0.00885	-0.01117 -0.00796	-0.00708 -0.00487	-0.00768 -0.00397	-0.01568 -0.01074	-0.01343 -0.00929	-0.00993 -0.00598
25.0	0.00198 0.00372	-0.00288 0.00125	-0.02542 -0.01314	-0.00471 -0.00315	-0.00121 0.00085	-0.01471 -0.00792	-0.01262 -0.00883	-0.00879 -0.00585	-0.00944 -0.00470
10.0	0.01644 0.02183	-0.00749 0.00113	-0.03112 -0.01506	0.01333 0.01753	0.00632 0.01129	-0.02431 -0.01200	-0.00127 0.00065	-0.00007 0.00308	-0.01875 -0.00986

FIGURE 4-5 ACCEPTABLE PARAMETER CONFIGURATIONS FOR WFBP WITH BILATERAL FILTERING APPLIED. ALL 95% CONFIDENCE INTERVALS FALL WITHIN THE ESTABLISH ±5% THRESHOLD.

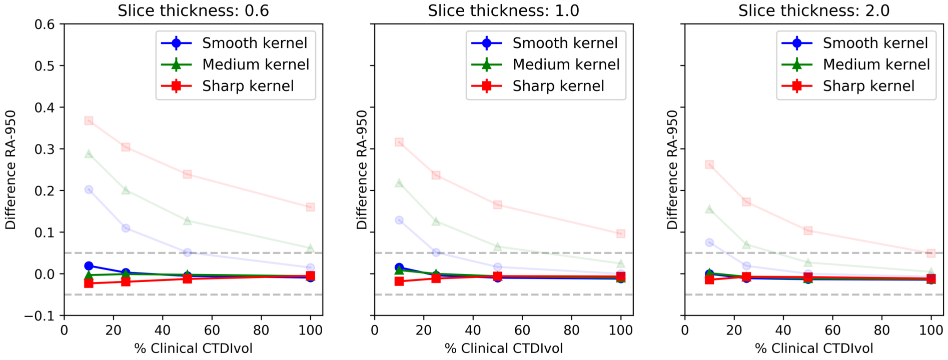


FIGURE 4-6 PLOT OF RA-950 CHANGE IN FULL POPULATION, RECONSTRUCTED WITH WFBP AND BILATERAL FILTERING, AS A FUNCTION OF ACQUISITION AND RECONSTRUCTION PARAMETER. ORIGINAL WFBP RESULTS (WITHOUT BILATERAL FILTERING) ARE OVERLAID WITH REDUCED OPACITY. GRAY, DASHED LINES SHOW ±5% THRESHOLD OF ACCEPTABLE CHANGE. WITH BILATERAL FILTERING, ROBUSTNESS OF RA-950 INCREASES DRAMATICALLY.

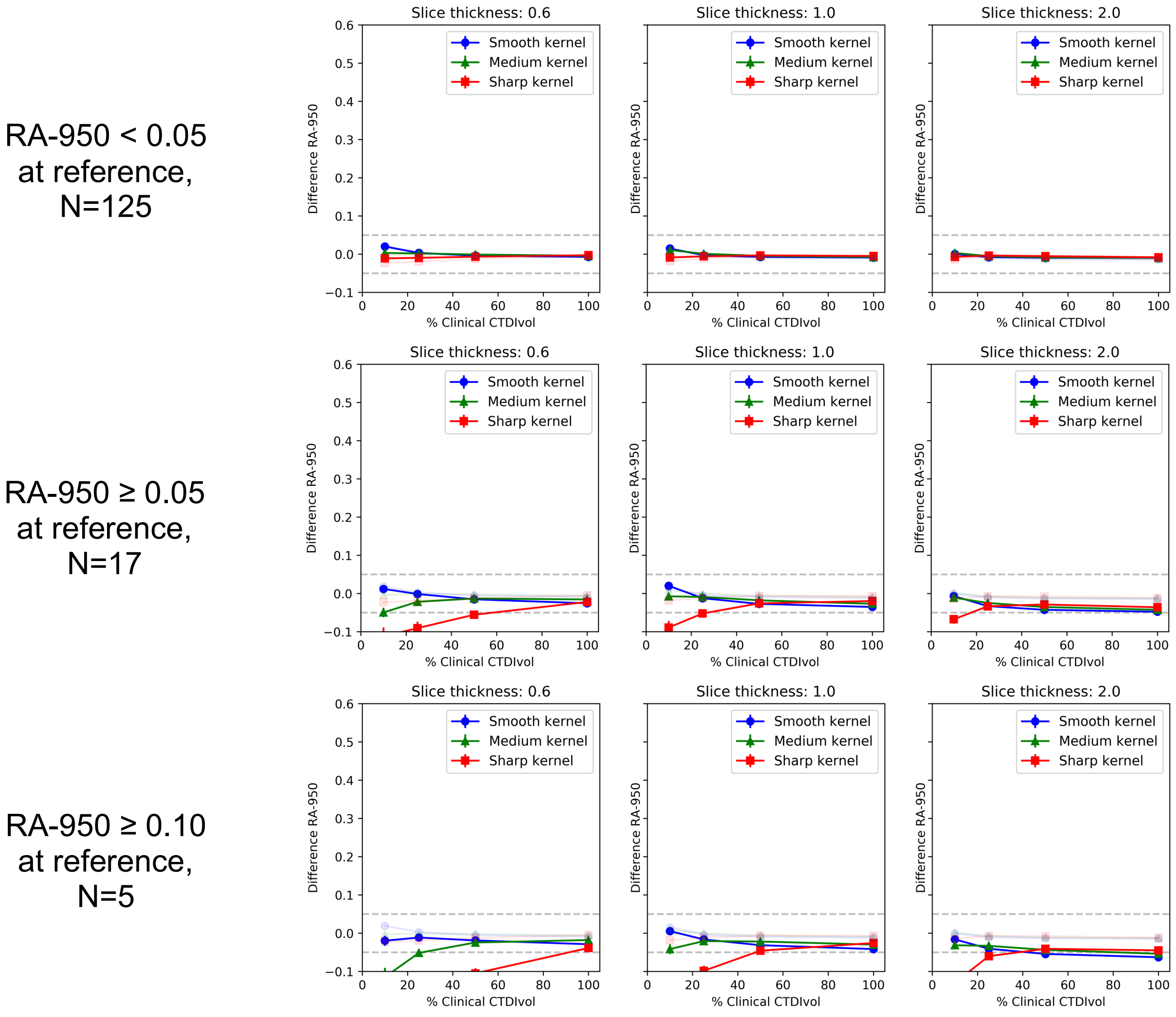


FIGURE 4-7 SUBGROUP PLOTS OF RA-950 CHANGE WITH WFBP+BILATERAL FILTERING. SUBGROUPS ARE FORMED BASED ON RA-950 SCORE AT THE REFERENCE CONDITION. ORIGINAL POOLED RESULTS (I.E. RESULTS PRESENTED IN FIGURE 4-6) ARE OVERLAID WITH REDUCED OPACITY.

Table 4-2 and Table 4-3 provide the results for the regression analysis (for RA-950 calculated on images reconstruction with wFBP and bilateral filter applied) with and without interaction terms. Both regressions resulted in an extremely poor fit (R^2 of 0.066 and 0.141 without interaction terms and with interaction terms, respectively). This poor fit, however, indicates that bilateral filtering was an effective means to reduce the impacts of kernel, slice thickness, and dose on RA-950 scoring. Sharp kernel and its interactions still had the strongest impact, however with a maximum regression coefficient absolute value of 0.036 (sharp kernel parameter, interaction model, Table 4-3), its effect relative to the ±0.05 threshold of acceptable change in RA-950 is minimal. Neither regression yielded a strong fit, however qualitatively in Figure 4-6 there does appear to be some interaction between kernel selection and slice thickness, so it may be reasonable to conclude that bilateral filtering induced some interactions not found with wFBP alone. However, given the similarity of R^2 values and coefficients (e.g. sharp kernel has largest coefficient in both models) and overall low coefficient values it is difficult to make any strong conclusions.

Table 4-2 Regression analysis results for RA-950 under wFBP with bilateral filtering, ***without*** interactions. (*) denotes values that are not statistically significant.

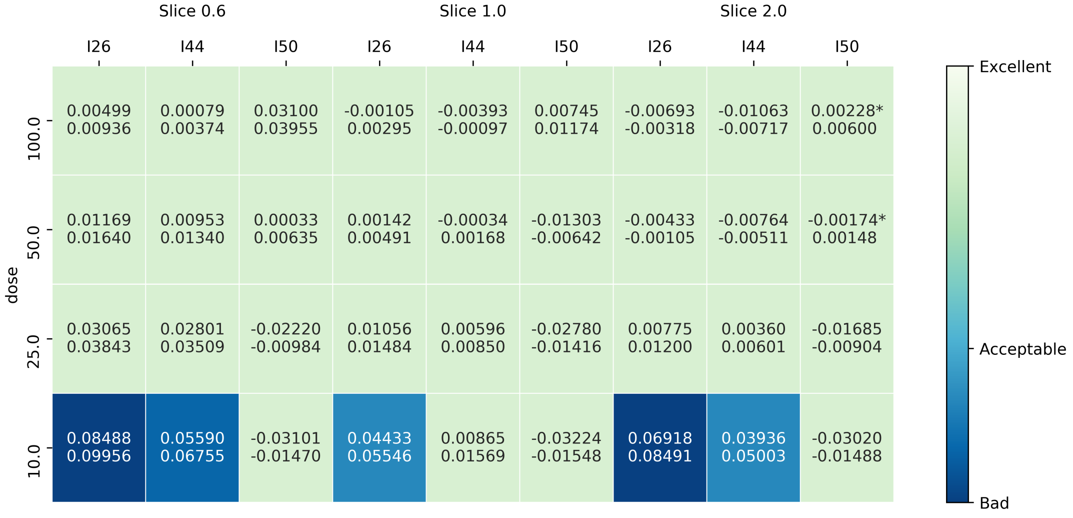
Intercept (Kernel: Smooth)	0.001 * (0.001)
Kernel: Medium	-0.001 * (0.001)
Kernel: Sharp	-0.009 (0.001)
Dose	0.000 (0.000)
Slice Thickness	-0.003 (0.000)
R-squared:	0.066
Adj. R-squared:	0.065
No. Observations:	5112

Table 4-3 Regression analysis results for RA-950 under wFBP with bilateral filtering, ***with*** interactions. (*) denotes values that were NOT statistically significant.

Intercept (Kernel: Smooth)	-0.001 * (0.001)
Kernel: Medium	-0.010 (0.002)
Kernel: Sharp	-0.036 (0.002)
Dose	0.000 (0.000)
Slice Thickness	-0.002 (0.001)
Dose x Slice Thickness	0.000 (0.000)
Dose x Kernel: Medium	0.000 (0.000)
Dose x Kernel: Sharp	0.000 (0.000)
Slice Thickness x Kernel: Medium	0.004 (0.001)
Slice Thickness x Kernel: Sharp	0.011 (0.001)
R-squared:	0.141
Adj. R-squared:	0.139
No. Observations:	5112

4.3.2 SAFIRE

SAFIRE with bilateral filtering (sample shown in Figure 4-9) resulted in fewer acceptable parameter configurations than wFBP with bilateral filtering, however still had a substantial improvement over wFBP without bilateral filtering and SAFIRE without bilateral filtering. An illustration of this is provided in Figure 4-8, along with the 95% confidence intervals of RA-950 change. When the dose was greater than 10% (~0.2 mGy CTDIvol), bilateral filtering greatly improved the robustness of emphysema scoring using SAFIRE. This is a promising result, since it was previously illustrated that the 10% dose is unusable for most applications (Figure 3-12, page 128). Interestingly, at 10% dose, the I50 sharpness setting resulted in the most robust performance; this result however is strongly influenced by the cohort largely not having emphysema, and does not hold for subjects with >0.05 emphysema at reference (shown in Figure 4-11).



dose	Slice 0.6 I26	Slice 0.6 I44	Slice 0.6 I50	Slice 1.0 I26	Slice 1.0 I44	Slice 1.0 I50	Slice 2.0 I26	Slice 2.0 I44	Slice 2.0 I50
100.0	0.00499 0.00936	0.00079 0.00374	0.03100 0.03955	-0.00105 0.00295	-0.00393 -0.00097	0.00745 0.01174	-0.00693 -0.00318	-0.01063 -0.00717	0.00228* 0.00600
50.0	0.01169 0.01640	0.00953 0.01340	0.00033 0.00635	0.00142 0.00491	-0.00034 0.00168	-0.01303 -0.00642	-0.00433 -0.00105	-0.00764 -0.00511	-0.00174* 0.00148
25.0	0.03065 0.03843	0.02801 0.03509	-0.02220 -0.00984	0.01056 0.01484	0.00596 0.00850	-0.02780 -0.01416	0.00775 0.01200	0.00360 0.00601	-0.01685 -0.00904
10.0	0.08488 0.09956	0.05590 0.06755	-0.03101 -0.01470	0.04433 0.05546	0.00865 0.01569	-0.03224 -0.01548	0.06918 0.08491	0.03936 0.05003	-0.03020 -0.01488

FIGURE 4-8 ACCEPTABLE PARAMETER CONFIGURATIONS FOR SAFIRE WITH BILATERAL FILTERING APPLIED. MOST 95% CONFIDENCE INTERVALS FALL WITHIN THE ESTABLISH ±5% THRESHOLD, HOWEVER I26 AND THE 10% DOSE SETTING SEEMED TO CONSISTENTLY RESULT IN SCORES THAT WERE TOO HIGH. (*) INDICATES ONE MISSING IMAGE DATASET (I.E. SCORES COMPUTED OVER 141 SUBJECTS RATHER THAN 142).

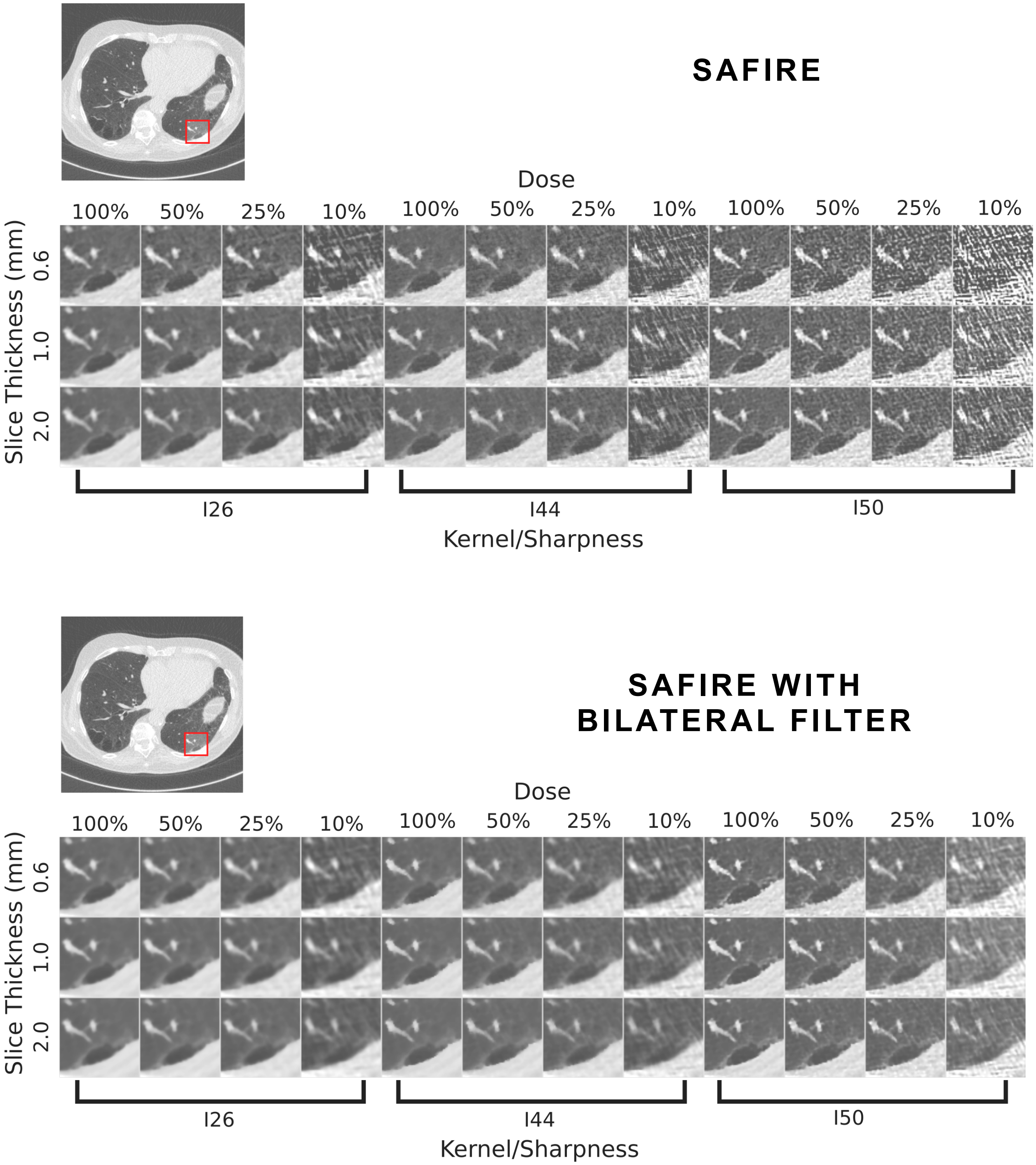


Figure 4-9 SAFIRE denoising results with bilateral filter (bottom row) presented with original SAFIRE images (top row). All images shown with window/level of 1600/-600. Reference acquisition/reconstruction can be found in Figure 4-2.

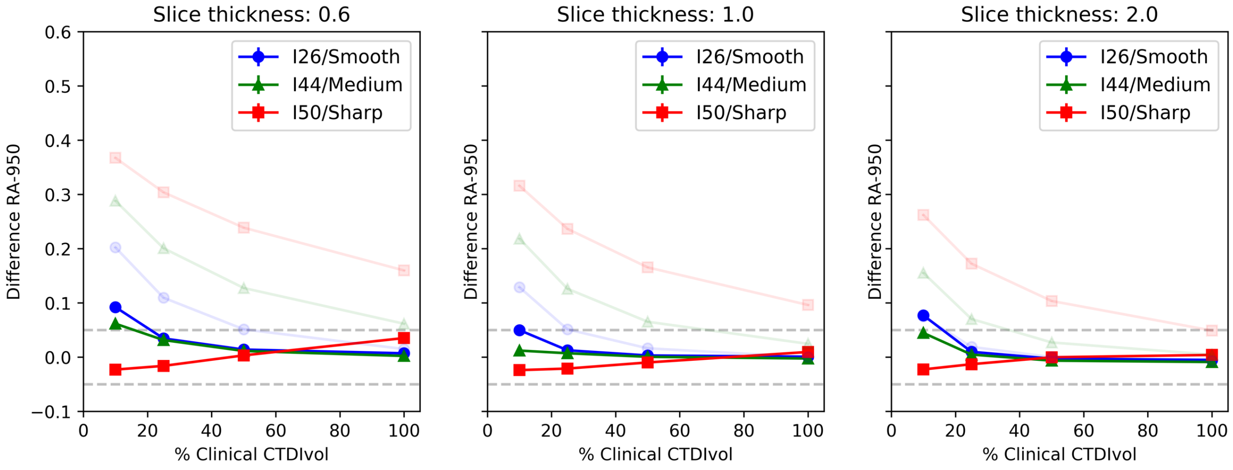


FIGURE 4-10 LINE PLOTS OF RA-950 CHANGE UNDER SAFIRE RECONSTRUCTION WITH BILATERAL FILTERING AS FUNCTION OF PARAMETER CONFIGURATION. ORIGINAL WFBP RESULTS WITHOUT BILATERAL FILTERING ARE OVERLAID WITH REDUCED OPACITY FOR COMPARISON.

As can be seen in Figure 4-10, the RA-950 scored under the I50 setting continues to display behavior different than I26, I44, and all wFBP kernel settings. This effect is more pronounced when considering emphysematous patients only, which can be found in Figure 4-11. As with wFBP with bilateral filtering, emphysematous patients undergo a decrease in RA-950 scores at high-dose, smooth-setting reconstructions (e.g. 100% dose, 2.0mm slice thickness, I26 reconstruction), and with sharp kernel, low-dose settings, there is a strong decrease in RA-950 scores, falling well outside of the ±5% limits of acceptable score change.

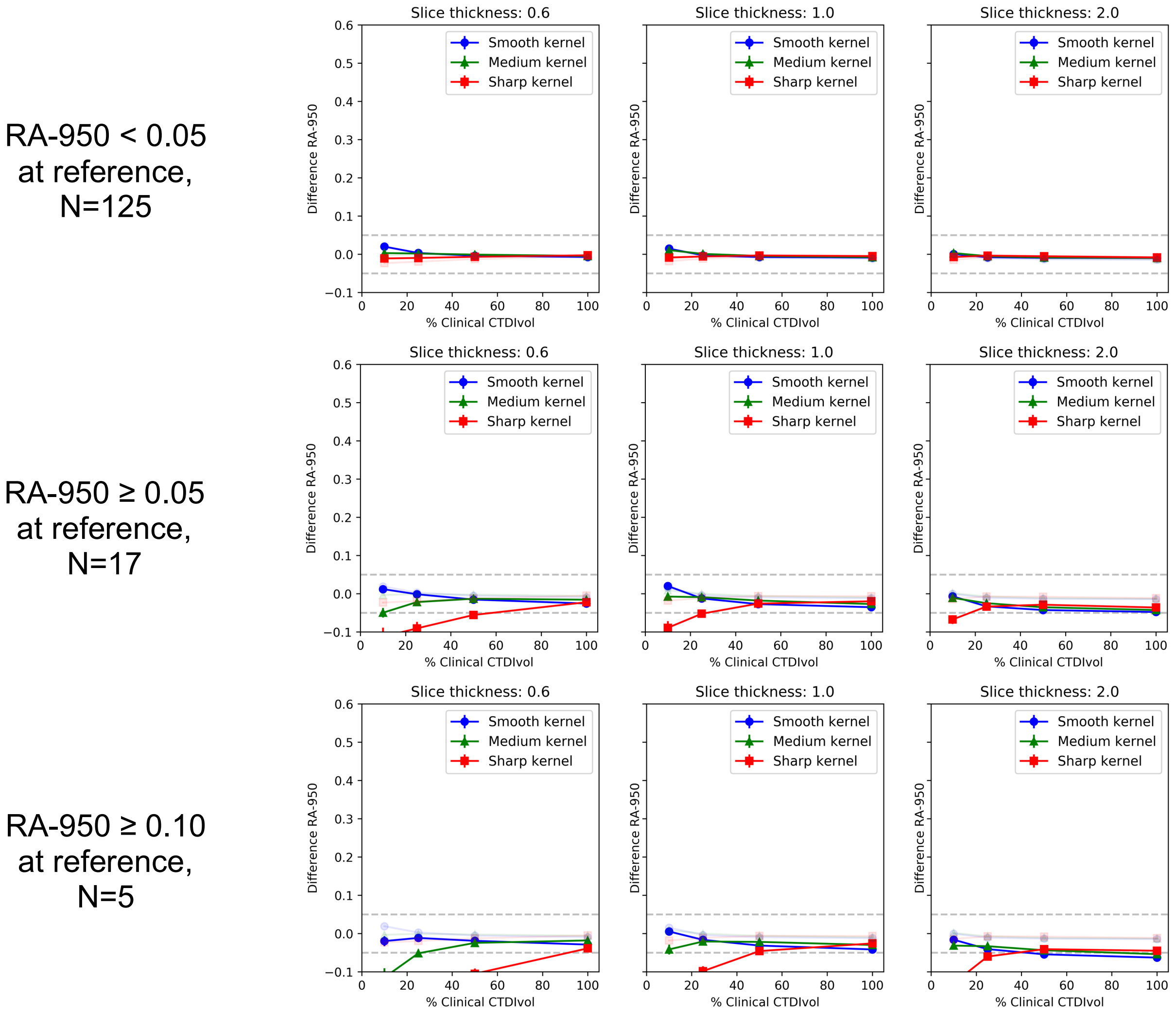


FIGURE 4-11 SUBGROUP PLOTS OF RA-950 CHANGE WITH SAFIRE+BILATERAL FILTERING. ORIGINAL POOLED RESULTS (FOR SAFIRE WITH BILATERAL FILTERING, FIGURE 4-10) ARE OVERLAID WITH REDUCED OPACITY.

Table 4-4 and Table 4-5 present the results for the regression analysis of RA-950 change both with interaction terms and without. R^2 values for both were low (0.169 and 0.357 for the non-interaction and interaction models, respectively), indicating a poor fit to the data with the variables utilized, although a somewhat improved fit than in the models of wFBP with bilateral filtering. This however, as with wFBP with bilateral filtering above, is a positive result, supporting the use of bilateral filtering for removing the effects of parameter configuration, since none of the parameters was especially predictive of the change in RA-950. The regression analysis quantitatively demonstrates that bilateral filtering was an effective means of removing most of the effects of kernel, dose, and slice thickness on RA-950 scoring in this cohort. While including interactions in the model resulted in an improved fit, the overall poor fit of the interaction model (R^2 of 0.357) and small coefficients relative to ±5% threshold of acceptability make it difficult to conclude that the interaction model is substantially better than the non-interaction model.

It is worth noting however, that wFBP with bilateral filtering resulted in smaller R^2 values relative to the SAFIRE results, as well as smaller coefficients. This implies that wFBP with bilateral filtering depends less on parameter selection than SAFIRE, and as a result is likely to be better choice for robust quantitative emphysema scoring with RA-950.

Table 4-4 Regression analysis results for RA-950 under SAFIRE with bilateral filtering, without interactions.

Intercept (Kernel: Smooth)	0.029 (0.001)
Kernel: Medium	-0.011 (0.001)
Kernel: Sharp	-0.031 (0.001)
Dose	0.000 (0.000)
Slice Thickness	-0.008 (0.001)
R-squared:	0.169
Adj. R-squared:	0.168
No. Observations:	5103

Table 4-5 Regression analysis results for RA-950 under SAFIRE with bilateral filtering, with interactions. (*) denotes values that were NOT statistically significant.

Intercept (Kernel: Smooth)	0.024 (0.002)
Kernel: Medium	-0.022 (0.003)
Kernel: Sharp	-0.087 (0.003)
Dose	0.000 (0.000)
Slice Thickness	-0.004 (0.001)
Dose x Slice Thickness	0.000 (0.000)
Dose x Kernel: Medium	0.000 (0.000)
Dose x Kernel: Sharp	0.001 (0.000)
Slice thickness x Kernel: Medium	0.000 * (0.002)
Slice thickness x Kernel: Sharp	0.005 (0.002)
R-squared:	0.357
Adj. R-squared:	0.355
No. Observations:	5103

4.4 Discussion

Denoising using bilateral filtering, when adapted based on dose and slice thickness setting, greatly improved the robustness of RA-950 scoring under both wFBP and SAFIRE. wFBP with bilateral filtering still demonstrated better performance (i.e. more acceptable parameter configurations) as well as fewer apparent interactions than SAFIRE with bilateral filtering and thus, we argue that it is the better choice for robust, quantitative emphysema scoring. While other iterative algorithms might yield highly robust noise reduction and accurate quantitative imaging, it is not clear what the SAFIRE algorithm at the settings utilized is optimizing, yielding unpredictable quantitative results under some conditions.

One possible explanation of why bilateral filtering was more effective in wFBP than SAFIRE is the increase in noise structure that seems to occur with SAFIRE. In the 25% and 10% dose cases show in Figure 4-9, there appears to be a dramatic increase in “streaking” across the image, while the equivalent conditions in Figure 4-2 seem to only increase in “graininess.” It is possible that because bilateral filtering is an edge-preserving denoising approach that these streaks were left untouched since they seem to indicate a true image feature rather than just noise. Although we would still recommend wFBP be used for RA-950 scoring and not SAFIRE, a more tailored denoising algorithm could improve the removal of such structured noise.

The appendix at the end of the chapter illustrates that highly similar results are observed for PERC15. As in Chapter 3, there were slightly more parameter configurations that resulted in unacceptable levels of change for both SAFIRE and wFBP with bilateral filtering. While this does not directly contradict studies that have found that PERC15 may be a slightly more repeatable measure of emphysema for the purposes of longitudinal patient evaluation [23], [24], it does suggest that PERC15 may more susceptible to reconstruction and acquisition parameters, at least in a lung screening population. In particular, SAFIRE reconstructions appear to have unclear effects on the image data, which causes PERC15 scoring changes that appear largely unpredictable. As stated in

Chapter 3 however, these results are dependent on what is defined as “acceptable” levels of change.

Use of the sharp kernel or sharp iterative settings should still likely be excluded for the purposes of quantitative emphysema scoring. While bilateral filtering appears to dramatically improve the robustness of scores in the sharp reconstructions, in all patients, especially those with high-emphysema, this still appears to be the parameter and setting that results in the most complex and difficult to predict behavior of RA-950 and PERC15 scoring. This is supported by the regression model analysis, in which the sharp kernel, in all cases, had the largest coefficient. Thus, it plays the largest “role” in causing change in RA-950 score and as a result, should be avoided if clinical emphysema scoring were to be performed.

One interesting effect observed in the subpopulation analysis was a systematic decrease in RA-950 scores (and corresponding increase in PERC15) among subjects with high baseline RA-950 scores (i.e. higher emphysema), which could reflect a very real concern about any denoising approach: the over-smoothing of actual regions of emphysema, which could result in under-scoring or under-diagnosis. While this experiment was not designed to determine exactly why this was the case, qualitative review of several high-emphysema cases, as illustrated in Figure 4-12 indicate that there does appear to be some loss at the borders of emphysematous regions, however it also appears that bilateral filtering helps smooth out some extraneous voxels from the reference condition mask (i.e. noise “speckle”), which would result in the observed shift down in RA-950 scores (i.e. a net improvement in the accuracy of scores). In either case, improvements to the denoising method could reduce this effect by considering image-based features for tuning of the filter. For example, instead of a heuristic tuning of bilateral filter based on knowledge of dose reduction and slice thickness change, one could utilize the standard deviation of an ROI placed in the trachea or lung parenchyma of a subject. This would prevent over-smoothing of a case that was already “acceptable” for emphysema scoring. Possible future work discussed in the next chapter explores several pathways forward to both determine the impact of the denoising approach on density mask features as well as possible means for improving the patient-specific and parameter-specific tuning of denoising strength.

A potential concern with the utilization of denoising is the destruction of underlying disease information. For example, in the case of emphysema scoring this would potentially reflect the loss of emphysematous voxels at the boundary of emphysema pockets, or the elimination of smaller pockets of emphysema due to the blurring or averaging operations typically incorporated into the filter. As Figure 4-12 illustrates, this loss of mask borders and elimination of small emphysema pockets can occur with the bilateral filter, however it only appears to result in unacceptable amounts of change in emphysema score for challenging conditions (e.g., high noise), such as <0.2mGy CTDIvol with a sharp kernel.

Were the loss of the density mask at borders and small regions a substantial issue, it would be reflected in our results due to the experimental design utilized in this work. By comparing scores computed on an unfiltered wFBP reconstruction with scores computed on the filtered reconstructions, substantial loss of the emphysema mask due to the bilateral filter would be reflected as an "unacceptable" decrease in score (i.e. a decrease of greater than 5%). While this was observed to a slightly greater extent in subjects with >0.05 RA-950 score at reference (Figure 4-7, in particular the low-dose, sharp kernel reconstructions), it can be seen that the bilateral filter still improves the robustness of the emphysema scoring relative to the scoring performed on the unfiltered image datasets.

The extent to which the bilateral filter removes emphysema borders and small pockets could be rigorously evaluated in a future experiment by conducting a simulation study utilizing a phantom such as the XCAT phantom [122] and simulating a series of emphysema pockets of varying sizes. The phantom could then be reconstructed utilizing the same parameters in this experiment and scored. Since the amount of emphysema is simulated and known exactly, the loss due to the filtering process could be determined exactly. Results could then be utilized to further improve the bilateral filter tuning scheme or set limits on when a scan is too compromised with noise for denoising to be effective.

Finally, it is worth noting that bilateral filtering, while effective for emphysema scoring, may not be the ideal filter selection for other quantitative tasks. Bilateral filtering is fundamentally a blurring/averaging operation (with some limited measure of edge-preservation), and it is unlikely that bilateral filtering will preserve more complex information such as texture or fine structures. Since emphysema tends to occur in pockets, and emphysema scoring is based on density measurements, this averaging process is effective and reasonable for the task evaluated here, however other denoising schemes (such as that utilized in [123] to classify voxels into lung parenchymal abnormalities such as ground glass, lung fibrosis, etc.) would likely be more effective for different clinical tasks. Future work should compare the application and effects of different denoising approaches for a given diagnostic task, and across different diagnostic tasks.

A larger discussion of limitations of this work and future work will be discussed in detail in Chapter 5, since much of it applies to both Chapter 3 and Chapter 4.

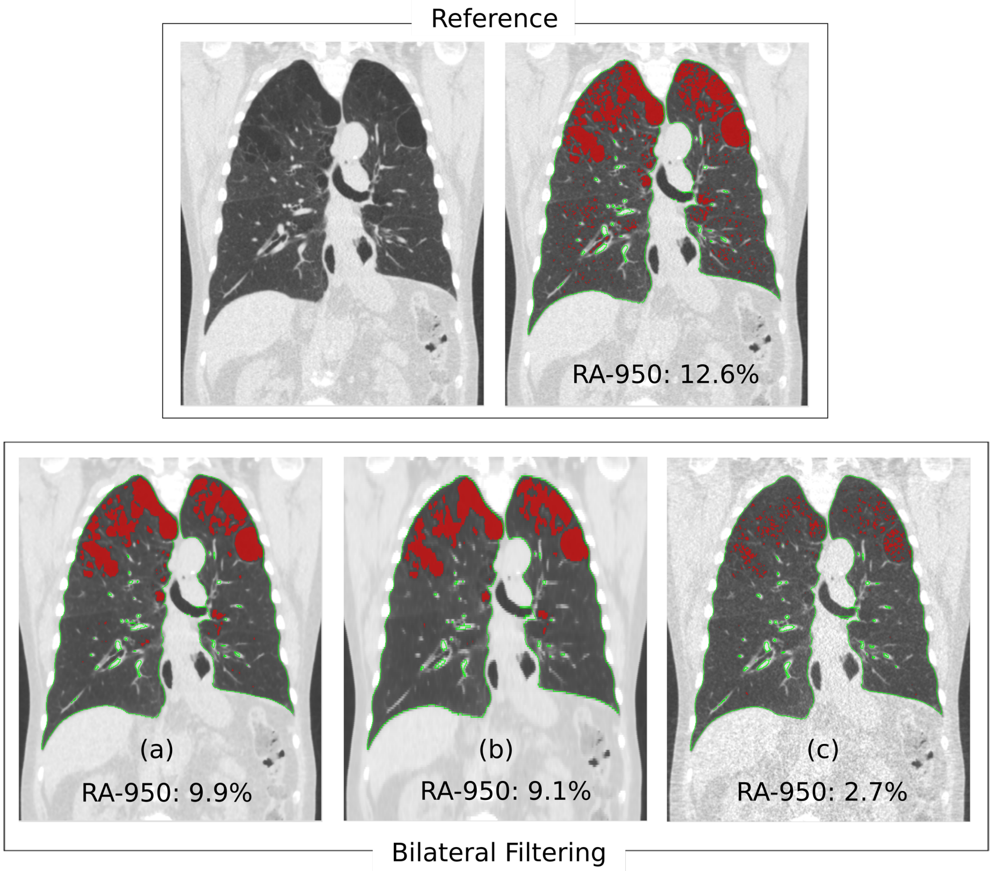


Figure 4-12 Sample coronal images of RA-950 mask in a high-baseline emphysema patient. 100% dose is shown. Top row shows the reference reconstruction and corresponding RA-950 mask. Bottom row shows three examples with bilateral filtering: (a) reference reconstruction (100% dose, 1.0mm slice thickness, smooth reconstruction kernel) (b) 100% dose, 2.0mm slice thickness, smooth kernel, and (c) 25% dose, 0.6mm slice thickness, sharp kernel (not usable).

4.5 Conclusions

Denoising, in particular with bilateral filtering, presents an extremely viable pathway forward for the clinical use of quantitative emphysema scoring. Specifically, it appears to dramatically increase the robustness of RA-950 and PERC15 scores to commonly varied imaging parameters, namely acquisition dose, slice thickness, and sharpness setting (either kernel or iterative setting). While the results here demonstrated improvements for both wFBP and SAFIRE, results support the argument that wFBP is the better choice for clinical, robust emphysema scoring since more acceptable parameter configurations were found with wFBP. wFBP is also a more "accessible" reconstruction configuration since, at present, all scanners still provide a version of filtered-backprojection reconstruction. This does not guarantee that there are not proprietary, non-linear, "black-box" pre- or post-processing methods applied to either the projection or raw data with wFBP, however wFBP is a fundamentally more well-understood and widely available algorithm than the currently heterogeneous field of iterative reconstruction algorithms, with the additional benefit that the underlying reconstruction method is linear and preserves photon counting physics and statistics (e.g. Poisson process, mean values, etc.). In the long run, this may not prove to be necessary for reliable quantitative imaging, however it at least gives at present some intuition around how noise magnitude, noise power spectrum shape, and spatial resolution depend on acquisition and reconstruction parameters.

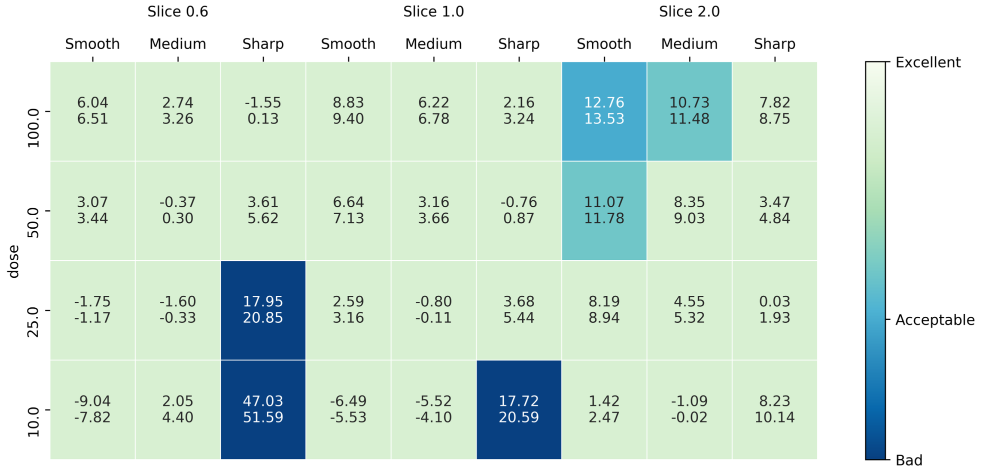
All of the results presented strongly suggest that some amount of denoising can readily be applied to image data to improve emphysema score robustness, however our results also suggest that some care must be taken with applying denoising to patients with higher emphysema, if an accurate score is required. PERC15 and RA-950 displayed similar behavior within the parameter configurations explored for wFBP with bilateral filtering. However, RA-950 proved to be slightly more robust under the tests utilized here, and provides the added benefit of localizing pockets of emphysema for visual review or further quantitative analysis (such as number of bullae/pockets, evaluations of the size of pockets, filtering of pockets by size, etc.). Since both are straightforward to compute and have similar requirements (i.e. a lung segmentation), and are equally robust under wFBP with bilateral filtering[4], we would recommend that both be used in a complimentary manner.

Finally, improvements to the underlying denoising approach should be investigated. In this study, knowledge of the reference versus test parameter configuration was utilized to adapt the strength of the denoising. This information is not typically available, and furthermore does not account for factors such as patient size or reconstruction kernel/sharpness. A reasonable next step in the investigation of denoising with bilateral filtering might be to adjust the strength of denoising based on an ROI measurement inside either the lung parenchyma or trachea, making the approach more patient- and scan-specific and could make the denoising more effective.

[4] With the exception of the 10% dose configurations. This however, is far below what would be considered “acceptable” for radiologist review and would likely be sent back for a rescan.

Appendix: PERC15 Results

WFBP with bilateral filtering



dose	Slice 0.6 Smooth	Slice 0.6 Medium	Slice 0.6 Sharp	Slice 1.0 Smooth	Slice 1.0 Medium	Slice 1.0 Sharp	Slice 2.0 Smooth	Slice 2.0 Medium	Slice 2.0 Sharp
100.0	6.04 6.51	2.74 3.26	-1.55 0.13	8.83 9.40	6.22 6.78	2.16 3.24	12.76 13.53	10.73 11.48	7.82 8.75
50.0	3.07 3.44	-0.37 0.30	3.61 5.62	6.64 7.13	3.16 3.66	-0.76 0.87	11.07 11.78	8.35 9.03	3.47 4.84
25.0	-1.75 -1.17	-1.60 -0.33	17.95 20.85	2.59 3.16	-0.80 -0.11	3.68 5.44	8.19 8.94	4.55 5.32	0.03 1.93
10.0	-9.04 -7.82	2.05 4.40	47.03 51.59	-6.49 -5.53	-5.52 -4.10	17.72 20.59	1.42 2.47	-1.09 -0.02	8.23 10.14

FIGURE 4-13 ACCEPTABLE PARAMETER CONFIGURATIONS FOR PERC15, SCORED ON WFBP WITH BILATERAL FILTERING RECONSTRUCTIONS. NEARLY ALL 95% CONFIDENCE INTERVALS FALL WITHIN THE ESTABLISH ±10HU THRESHOLD. AS SEEN WITH RA-950, THERE IS SOME DEVIATION IN THE "SMOOTHEST" CASES LIKELY CAUSED BY THE DENOISING ALGORITHM.

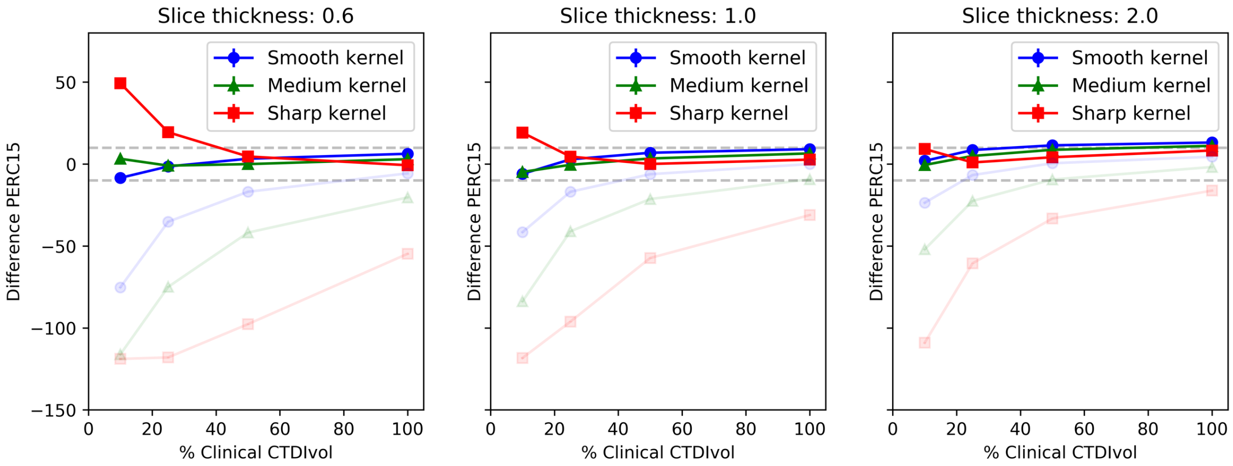


FIGURE 4-14 PLOT OF PERC15 CHANGE IN FULL POPULATION, RECONSTRUCTED WITH WFBP AND BILATERAL FILTERING, AS A FUNCTION OF ACQUISITION AND RECONSTRUCTION PARAMETER. ORIGINAL WFBP RESULTS (WITHOUT BILATERAL FILTERING) ARE OVERLAID WITH REDUCED OPACITY. GRAY, DASHED LINES SHOW ±10HU THRESHOLD OF ACCEPTABLE CHANGE. WITH BILATERAL FILTERING, ROBUSTNESS OF PERC15 INCREASES DRAMATICALLY.

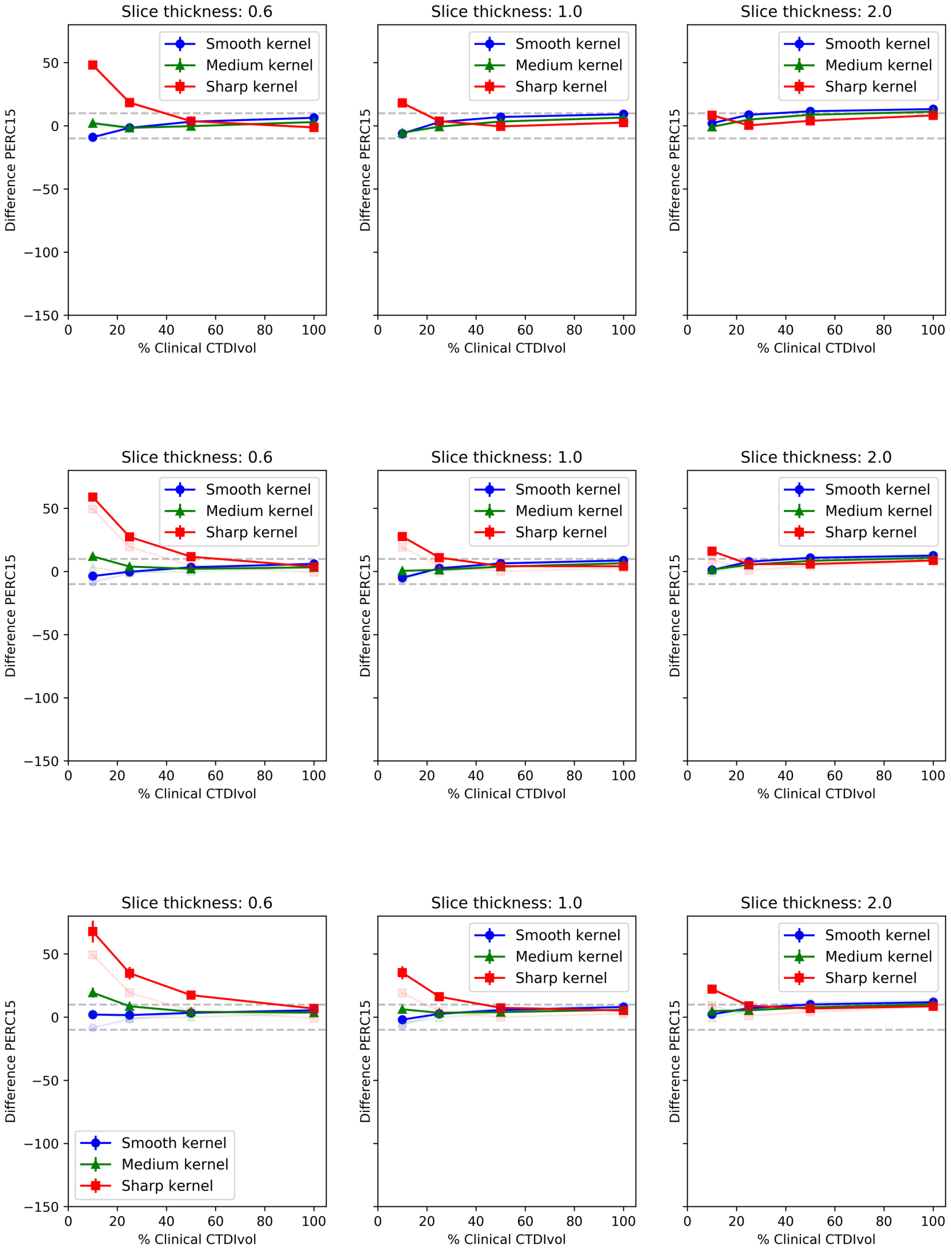
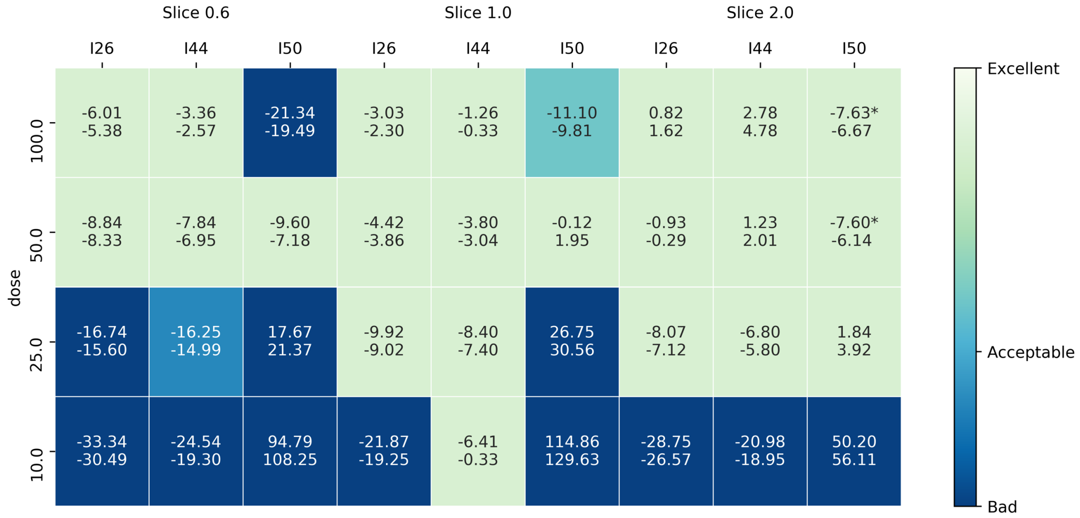


FIGURE 4-15 SUBGROUP PLOTS OF PERC15 CHANGE WITH WFBP+BILATERAL FILTERING. ORIGINAL POOLED RESULTS ARE OVERLAID WITH REDUCED OPACITY.

SAFIRE with bilateral filtering



dose	Slice 0.6			Slice 1.0			Slice 2.0		
	I26	I44	I50	I26	I44	I50	I26	I44	I50
100.0	-6.01 -5.38	-3.36 -2.57	-21.34 -19.49	-3.03 -2.30	-1.26 -0.33	-11.10 -9.81	0.82 1.62	2.78 4.78	-7.63* -6.67
50.0	-8.84 -8.33	-7.84 -6.95	-9.60 -7.18	-4.42 -3.86	-3.80 -3.04	-0.12 1.95	-0.93 -0.29	1.23 2.01	-7.60* -6.14
25.0	-16.74 -15.60	-16.25 -14.99	17.67 21.37	-9.92 -9.02	-8.40 -7.40	26.75 30.56	-8.07 -7.12	-6.80 -5.80	1.84 3.92
10.0	-33.34 -30.49	-24.54 -19.30	94.79 108.25	-21.87 -19.25	-6.41 -0.33	114.86 129.63	-28.75 -26.57	-20.98 -18.95	50.20 56.11

FIGURE 4-16 ACCEPTABLE PARAMETER CONFIGURATIONS FOR PERC15, SCORED ON SAFIRE WITH BILATERAL FILTERING RECONSTRUCTIONS. THE 10% DOSE CONFIGURATIONS TYPICALLY RESULTED IN UNACCEPTABLE SCORES. (*) INDICATES ONE MISSING IMAGE DATASET (I.E. SCORES COMPUTED OVER 141 SUBJECTS RATHER THAN 142).

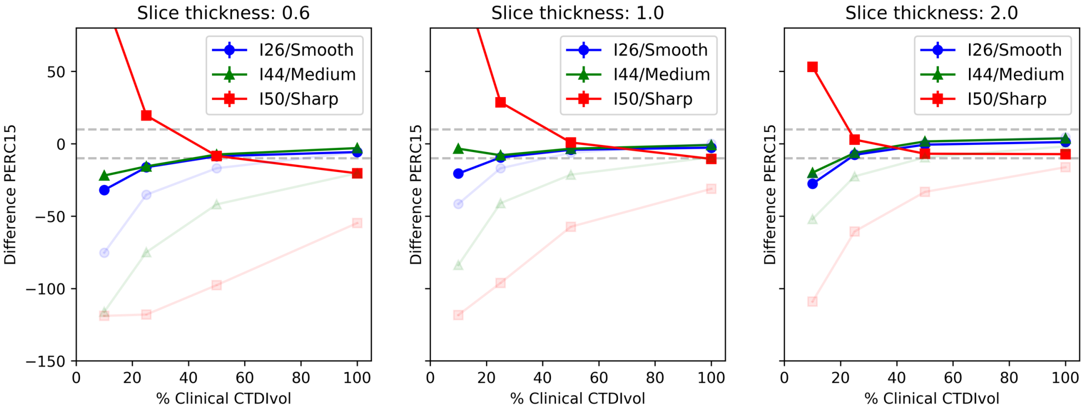


FIGURE 4-17 PLOT OF PERC15 CHANGE IN FULL POPULATION, RECONSTRUCTED WITH SAFIRE AND BILATERAL FILTERING, AS A FUNCTION OF ACQUISITION AND RECONSTRUCTION PARAMETER. ORIGINAL WFBP RESULTS (WITHOUT BILATERAL FILTERING) ARE OVERLAID WITH REDUCED OPACITY. GRAY, DASHED LINES SHOW ±10HU THRESHOLD OF ACCEPTABLE CHANGE. WITH BILATERAL FILTERING, ROBUSTNESS OF PERC15 INCREASES DRAMATICALLY, HOWEVER IS SUBSTANTIALLY WORSE THAN RA-950 AS WELL AS ALL WFBP WITH BILATERAL FILTERING RESULTS.

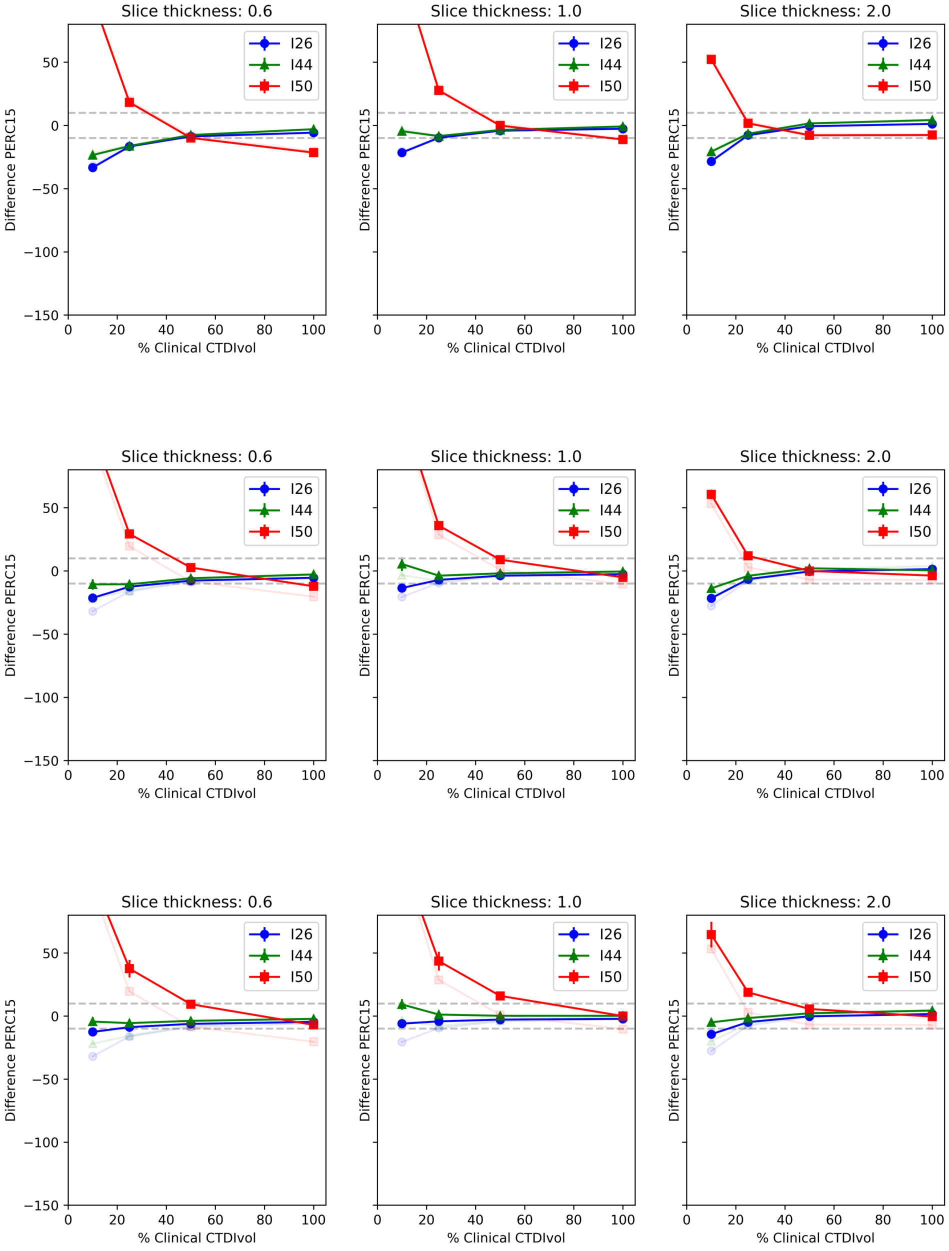


Figure 4-18 Subgroup plots of PERC15 change with SAFIRE+bilateral filtering. Original pooled results are overlaid with reduced opacity.

Chapter 5 – Conclusions

5.1 Review

At the beginning of this dissertation, we discussed the concept that an ideal quantitative imaging test would only reflect a patient's underlying disease state or anatomy, and would not be impacted by factors such as scanner manufacturer, acquisition protocol, or reconstruction. In practice however, each of these has some effect on the final quantitative score, which creates significant concerns when trying to implement clinical quantitative imaging tests that would be used to make diagnoses or guide treatment for a patient.

Previous studies have explored the fact that parameter changes cause variation in quantitative imaging, although most have only explored variation in one parameter, or in limited cohorts of patients. To overcome this, much of the initial efforts of this dissertation were infrastructure development to accelerate quantitative imaging studies, namely automating the dataset creation process and quantitative imaging tests. Through the development of free, open-source, CT reconstruction software, and a customized GPU pipeline framework, quantitative imaging studies that previously required upwards of six months to complete can now be carried out in less than one week. This enables much more thorough explorations in terms of the number of parameters, parameter configurations tested, and cohort size. Furthermore, because of simplicity of generating these datasets and the automated processing of quantitative imaging data, more quantitative imaging applications can be tested, such as computer automated detection/diagnosis (CAD), automated segmentation systems, and more recently developed deep learning tests.

Using the developed infrastructure, in Chapter 3 it was shown that while emphysema scoring with both RA-950 and PERC15 are susceptible to changes in dose, slice thickness, and kernel selection, there do exist small “regions” of parameter space that can give reasonably similar results to a reference reconstruction. This is shown for wFBP in Figure 3-3, Figure 3-4, Figure 3-6, and Figure 3-7; SAFIRE results are given in Figure 3-9 and Figure 3-10, as well as the appendix of Chapter 3. In practice, this could allow a researcher or clinician to evaluate (using a look-up table of acceptable parameter configurations) whether or not a scan can be utilized for quantitative emphysema scoring, and the likely difference in score they would expect to see as a result. For this to be utilized clinically however, maps such as Figure 3-3 (i.e. the heat map of acceptable and unacceptable parameter configurations) would need be generated for different scanners, manufacturers, reconstruction algorithms, and subpopulations (e.g. differing amounts of emphysema, which is difficult to know in advance in many cases) etc. As was shown in Figure 3-9 and Figure 3-10, while similar emphysema scores can be found using the SAFIRE iterative reconstruction algorithm, there were certain configurations for which no clear pattern emerged regarding which configurations would yield reliable scores, namely the I50 sharpness setting. Although I26 and I44 SAFIRE settings behaved more similarly to wFBP than I50, emphysema score changes at 25% and 10% still were worse than the smooth and medium settings for wFBP. Overall, while there is reasonable robustness of emphysema scoring to parameter configuration when wFBP is utilized, identifying these regions of robustness and translating this approach directly into the clinic would prove to be challenging for emphysema scoring across all scanners, manufacturers, and reconstruction algorithms.

Since an exhaustive search of all possible parameters, hardware, and software is unlikely, post-processing of the image data may be the most viable pathway forward. In Chapter 4 it was shown that denoising via the use of a bilateral filter removed nearly all of the effects of slice thickness, reconstruction kernel, and acquisition dose on quantitative emphysema scores, making RA-950 and PERC15 score values across a wide range of parameter configurations almost equal to their values at reference. Figure 4-5 and Figure 4-6 illustrate acceptable parameter configurations and trends for wFBP with denoising, and Table 4-2 and Table 4-3 present the regression results establishing that nearly all of the impact of parameter selection has on quantitative RA-950 scoring has been removed. While denoising with bilateral filtering did work better in wFBP (e.g. comparing and contrasting Figure 4-5 with Figure 4-8), it had a similar effect with SAFIRE iterative reconstructions lending some optimism that the approach could work for other reconstruction algorithms. We did find that some care must be taken in patients with greater amounts of emphysema, however even in these patients, more parameter configurations proved to be acceptable with denoising than without. Based on these results, as well as other recent explorations into "image normalization" schemes for emphysema scoring [49], denoising represents one of the most promising pathways to achieving reliable, robust, and widespread clinical quantitative imaging.

5.2 Limitations and Future Work

Cohort and lung screening protocol

The cohort for this study was 142 subjects scanned at UCLA using the CT lung screening protocol. A lung screening population, smokers with a smoking history of 30+ pack years, is likely to have emphysema, however most subjects in this study had minimal amounts of emphysema (< 0.05 RA-950 score at reference, N=125). Thus, all of the results and conclusions presented reflect a non-emphysematous population, however the subgroup analyses presented show that the underlying amount of emphysema does impact these results. To better understand the effects of amount of emphysema, other experiments should be conducted that capture a larger subject population with mild, moderate, and severe levels of emphysema and conduct similar analyses as those presented above. Furthermore, different emphysema subtypes (e.g. centrilobular versus paraseptal emphysema) may be susceptible in different ways to parameter configurations and denoising and should potentially be considered in future work.

While the lung screening protocol agrees well with the guidelines established by the NLST [40], [41] it is markedly lower dose than a diagnostic or quantitative CT [42]. Work has been done establishing that these low-dose protocols typically produce scores similar to their full-dose counterparts [33] suggesting that as long as the reference condition is similar to what would typically be utilized for quantitative imaging (in terms reconstruction algorithm, kernel, and slice thickness) this represents scores that would be similar to a true, full-dose quantitative scan (such as that recommended for SPIROMICS). Our selection of wFBP, smooth kernel, and 1.0mm slice thickness is similar to the quantitative protocols established at our institution (Siemens wFBP, B31f reconstruction kernel, 1.0mm slice thickness). Future work would ideally begin with a clinical quantitative CT protocol as reference and investigate the same sorts of parameter configurations and changes in quantitative emphysema scoring.

Selection of the reference protocol

Since no specific “truth” value is known for RA-950 and PERC15 scores for the subjects evaluated, the figure of merit was *change* in score relative to a reference parameter configuration. Thus, the selection of the reference protocol is critical to the overall work. As seen in Figure 4-12, while the reference protocol is a reasonable match for the clinical quantitative imaging protocol, it will occasionally contain voxels typically regarded as extraneous (i.e. “noise speckle”) which bilateral filtering removed and thus “lowered” the score. However, the bilateral filter may have simply improved the quality of the reference result. It is possible that there are other justifiable, reasonable selections for the reference configuration. As a result, researchers should be conscious of the impacts that the reference selection can have on the results, clearly state the selected reference, and explain why it was selected.

Scanner and reconstruction algorithms

All of the data in this work originated from the Siemens Definition AS 64 located in Medical Plaza 200 at UCLA. All of this data was reconstructed utilizing either FreeCT_wFBP (not on the scanner) or Siemens SAFIRE iterative reconstruction (on the scanner). The resulting dataset is extensive, however this only captures a small portion of all possible reconstructions. Ideally, future work would capture more scanner models and more reconstruction algorithms,

Exploring other scanners and reconstruction algorithms is challenging for two reasons: raw data access and a lack of on-board (i.e. on the scanner) automation. A key development enabling this work was the use of the pipeline and the pipeline relies on access to clinical raw projection data, as well as full automation of the reconstruction process. Work with image data from non-pipeline reconstructions is possible (as was done with the SAFIRE iterative reconstructions for this project), however substantially slows down the investigation process, and limits the number of configurations that can be explored.

Raw data access is often restricted, and raw data cannot generally be removed from most scanners. If it can be removed, there is the further challenge of extracting projections and necessary metadata from the file, which is often encoded in various schemes to reduce file sizes. Wider investigations covering multiple scanners or reconstruction schemes would be possible if manufacturers were to: (1) allow users to export raw data and (2) adopt the recently proposed, DICOM-based projection data format for more universal access [57]. It is unlikely that manufacturers will pursue this route in the near future, so in addition to FreeCT_wFBP, FreeCT_ICD, and the pipeline, we hope that future development from the research community will help enable broader access to both the raw projection data as well as clinically-similar reconstruction algorithms for use in research.

Denoising adjustment

Some measure of “adaptation” in the bilateral filter was achieved through the use of a heuristic rule to tune the range filter standard deviation. This rule however was developed based on knowledge of the relative parameter changes between configurations, which is not always known, and additionally fails to account for other factors that affect image noise magnitude and structure, such as kernel sharpness, or patient size. An improved scheme would consider image-based features to decide the proper denoising strength, such as standard deviation of a region of interest placed in the trachea. Future work should test several such schemes and see if it improves results. This will additionally make the denoising scheme more readily applicable to other datasets, since no knowledge beyond what is available in the image would be required for tuning.

Finally, while one of the strengths of the bilateral filter is its simplicity, it is also possible that different denoising schemes may be more effective, in particular with patients who have substantial emphysema. Future work should investigate alternative denoising schemes alongside, or coupled with bilateral filtering. Some good candidate approaches for alternative denoising algorithms could be BM3D denoising [124] or non-local means denoising [125], which have proven immensely effective in digital photography applications. We would also recommend that in addition to advanced denoising methods, more simplistic denoising approaches also be tested, such as basic median filtering or Gaussian filtering; these simple filtering approaches may help researchers understand what filtering approaches and filter features are most important to reliable quantitative results (e.g. preservation of mean values, filter linearity, and/or edge-preserving characteristics, etc.). In particular, denoising with bilateral filtering might make an excellent complement to the image normalization approach proposed in [49].

Defining "acceptable" variation

This work centers heavily on the concept of an "acceptable" amount of change that can be incurred in a quantitative imaging metric. Thus, the definition of what is or is not acceptable is central to any of the results obtained. For this dissertation, “acceptable” amount of change in emphysema scores were determined by considering the limitations of the emphysema scoring approach as well as the clinical implications of the potential score changes. In this work, we identified a change of ≤0.05 in RA-950 and a change of ≤10HU in PERC15 as acceptable.

Whenever defining what an acceptable tolerance for error would be, it is important to consider the diagnostic task. For emphysema scoring, there are two reasonably obvious tasks: prediction of the “correct” amount of emphysema, and evaluation of change over time. Accurate estimation of disease amount and tracking of disease change over time would likely require different levels of accuracy to be effective. In the context of lung cancer screening where scanning is conducted yearly or more frequently, it would be ideal if both tracking and accurate estimation could be performed and highly valuable to patient care since it could improve early diagnosis of emphysema in a population in which emphysema is likely to occur. However, while lung cancer screening presents an excellent opportunity for the tracking of emphysema score, it is unlikely that it is accurate enough to perform tracking beyond relatively crude designations (e.g. similar to the GOLD designations of mild, moderate, severe).

Gietema et al. [126] determined in a cohort of 157 lung screening subjects that RA-950 scoring could potentially assess a 1.1% change in patient score with 95% accuracy when perfect protocol implementation is utilized: patients are scanned on the same scanner using the exact same parameter configurations. While this is an extremely promising result, it is unlikely that such a high sensitivity is required for effective clinical quantitative imaging; it is also unlikely that a subject would be scanned often enough to observe such small changes (in [126], patients were scanned within 3 months). A larger threshold may be acceptable for establishing a diagnosis, i.e. baseline amount of emphysema for a patient, and a potential prognosis. The patient could then be called back to the clinic for a full-dose quantitative CT scan if detailed tracking would be needed. Thus, for this study in which the cohort was derived from a population scanned with the lung cancer screening protocol, a wider threshold of acceptability was utilized.

As far as establishing a diagnosis is concerned, current methods for diagnosis and prognosis of COPD typically group patients into categories such as mild, moderate, severe, and very severe. While these groupings arise primarily from evaluation methods such as FEV_1, the BODE index, or SGRQ (Saint George's Respiratory Questionnaire), there is evidence that patients in the "very severe" category typically display quantitative RA-950 scores >=30% [127], with other grouping falling roughly linearly below that (i.e. mild: 0-10%, moderate: 10-20%, and severe: 20-30%). Thus, we theorize that a change of >5% score would be likely to affect the prognosis and treatment course of a patient, and therefore we set the second threshold of "acceptable" change in score at 5%.

Finally, no such correlations between prognostic functional tests and the other quantitative metrics (i.e. PERC15) have been established in the same manner as for RA-950 described above. The PERC15 threshold of acceptability was selected to be ±10HU based on the work of [22]. In this work, a COPD subgroup (subjects with GOLD stage III and IV) had a mean PERC15 score of -985HU, while the control subgroup (GOLD stage 0 and I) had a mean PERC15 score of -963. Stratifying between the two values in increments of 10HU would result in a reasonable, albeit crude, correspondence between PERC15 score and GOLD status.

With denoising however, substantially more accurate results may be possible. Coupled with an image normalization procedure (such as that discussed in the Chapter 4 introduction), tighter tolerances may be possible, and longitudinal tracking may be substantially more viable with both clinical quantitative imaging or lung screening scans.

5.3 Implications for Clinical Quantitative Emphysema Scoring

WFBP versus SAFIRE

One of the primary goals of this dissertation is to determine implementable, achievable pathways for clinical usage of quantitative emphysema scoring. From the results described above, wFBP should be utilized for RA-950 and PERC15 scoring instead of SAFIRE. This is due to the fact that parameter configurations using wFBP as the reconstruction method were more likely than SAFIRE reconstructions to produce RA-950 results close to the reference score. Furthermore, SAFIRE presented non-trivial interactions between dose, slice thickness, and sharpness setting (i.e. the iterative equivalent of kernel) that make it difficult to predict whether or not a given parameter configurations would predictably produce RA-950 scores close to reference. These SAFIRE behaviors are further complicated when considering different underlying levels of patient emphysema, which appear to cause conflicting, unpredictable behavior across the different parameter configurations. wFBP on the other hand demonstrated predictable behavior with noisier reconstruction resulting in higher RA-950 scores. While interactions were statistically significant, none were clinically significant, meaning that parameters can be adjusted independently without the risk of complex, unexpected behaviors in RA-905 scoring.

It is important to emphasize that these results do not mean that iterative reconstruction approaches cannot be used for quantitative emphysema scoring, but rather highlight that with the current state of iterative algorithms, extreme caution should be exercised when attempting quantitative imaging on iterative reconstructions. The term “iterative reconstruction” is presently being applied to a wide variety of different algorithms that all approach the problem of denoising, artifact reduction, and edge-preservation differently. SAFIRE is a hybrid algorithm, beginning with an FBP reconstruction, followed by an iterative denoising step and some level of raw-data domain verification of the denoising and artifact reduction. Other approaches often also referred to as “iterative reconstruction”, such as GE’s ASiR, and Toshiba’s AIDR 3D, are at their core an FBP reconstruction that has been denoised in the image domain using an iterative post-processing technique. Finally, a third class of “iterative” algorithms implement a fully “model-based” approach that is highly computationally intensive however attempts a detailed modeling of the CT system properties and iterative updates the reconstructed volume through careful comparison with the raw data. Examples of this approach include Toshiba’s FIRST algorithm and GE’s Veo algorithm, as well as FreeCT_ICD discussed in Chapter 2.

In the current paradigm of iterative reconstruction methods, it will be difficult or impossible to ever reach a strong conclusion regarding acceptable and unacceptable configurations, or whether or not iterative reconstruction as a whole is beneficial to or hinders quantitative imaging. With some clarification from the manufacturers as to the underlying approach, it is conceivable that a class of iterative algorithms (e.g. model-based iterative algorithm with an edge-preserving penalty function) could be found that works exceedingly well for quantitative imaging, achieving denoising and artifact reduction while preserving underlying physical properties of the data thought to be important for accurate quantitative imaging. As iterative reconstruction matures, more research will hopefully emerge helping to characterize each algorithm and its impacts on quantitative imaging. Prior to that point however, wFBP algorithms represents a more reliable, more predictable, algorithm for use with quantitative imaging.

Acquisition

The doses explored represent lung screening doses and below. Sample reconstructions from all doses investigated are shown in Figure 5-1 with wFBP reconstruction and Figure 5-2 with SAFIRE reconstruction. Lung screening is already a fairly low-dose protocol (~2mGy CTDIvol) when compared to routine diagnostic exams (~10-15mGy CTDIvol). The 10% dose configuration was investigated in an attempt to find a lower bound of possible acceptable configurations. While our results suggest that this dose level may not be entirely out of the question with denoising, the overall image quality would not be acceptable for radiologist review. 25% dose levels and above however could be viable in patients for the purposes of quantitative imaging, assuming that some type of denoising is applied, and kernel/sharpness setting selection is not sharp or over-enhancing.

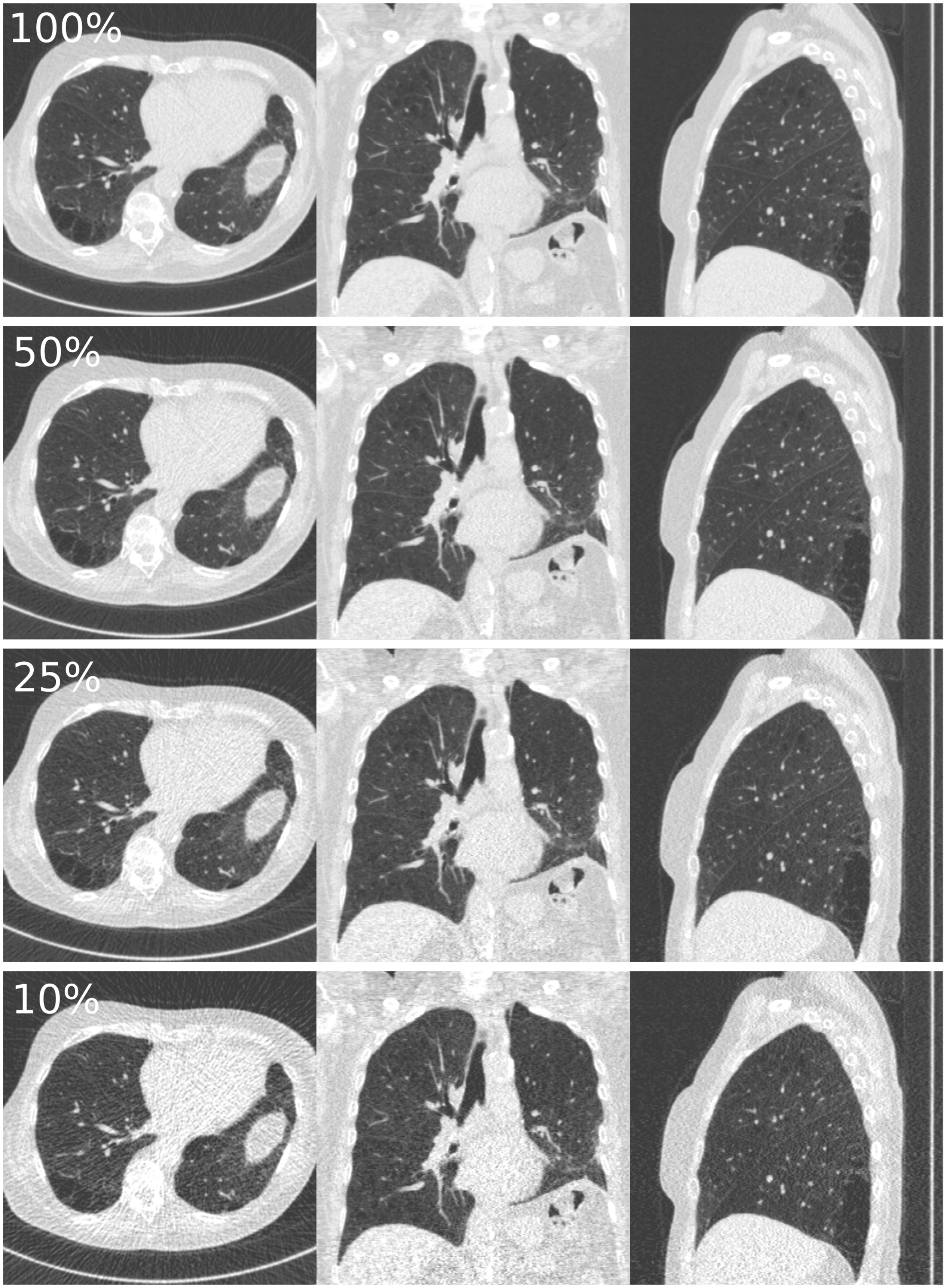


FIGURE 5-1 EXAMPLES OF DIFFERENT DOSE LEVELS INVESTIGATED WITH **wFBP**. IMAGES ARE RECONSTRUCTED WITH SMOOTH KERNEL AND 1.0MM SLICE THICKNESS. NO DENOISING IS APPLIED.

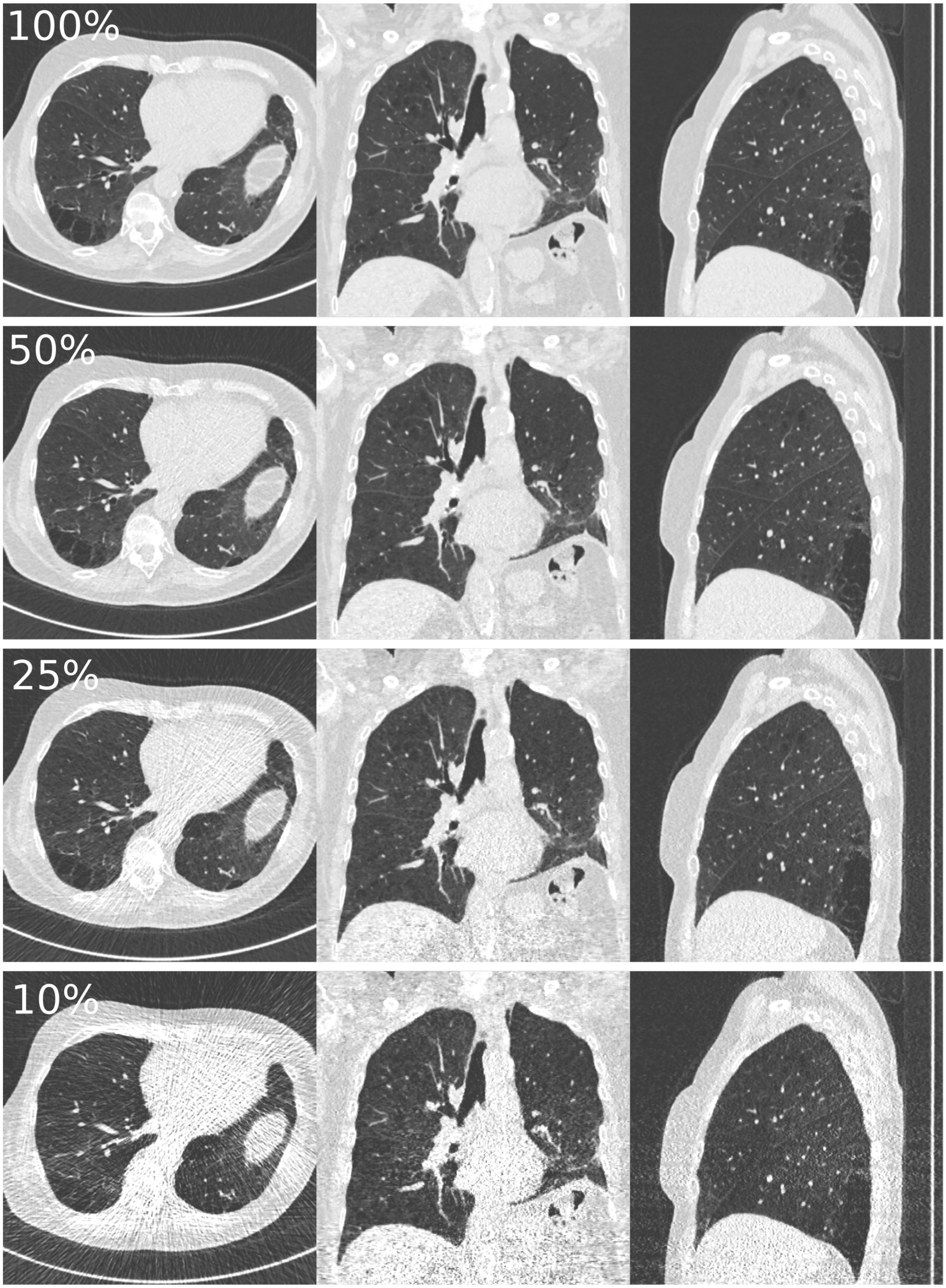


Figure 5-2 Examples of different dose levels investigated with **SAFIRE**. Images are reconstructed with smooth kernel and 1.0mm slice thickness. No denoising is applied.

Reconstruction Parameters

As previous studies have found, and this work supports: sharp or over-enhancing kernels should not be used when performing quantitative emphysema scoring. In all cases, including with both SAFIRE and wFBP, these kernels produced results that fell outside of what would be considered "acceptable" in any routine clinical practice. When coupled with denoising, these results improved, however it is easy to identify patients and settings for whom a sharp kernel would produce aberrant results. Anyone attempting to perform quantitative imaging using a sharp kernel reconstruction should use extreme caution when interpreting the results.

Based on our results, thicker slices resulted in more parameter configurations that yielded acceptable results. As a result, we would recommend reconstructing 1.0mm slices and above for emphysema scoring. We did not directly investigate slice thicknesses above 2.0mm, however based on results of other studies [49] 3.0mm slice thicknesses should be perfectly acceptable. We would not recommend scoring emphysema directly on 0.6mm slices, due to the increased noise observed. With thin slice reconstructions, one can often simply average slices together to achieve a pseudo-thicker-slice reconstruction and it is on this reconstruction that we would attempt emphysema scoring. While we did not investigate this approach directly, other studies have investigated it and found that it behaves as expected [49].

Denoising

In our study, denoising dramatically improved the robustness of quantitative emphysema scoring and removed much of the effects of the parameter tested. While quantifying acceptable parameters is important for the understanding of how parameter selection impacts quantitative emphysema scoring, it appears that denoising is the best pathway forward to bring quantitative imaging to the clinic. While further testing needs to be done to ensure performance across different levels and types of emphysema, the results of this study show that the potential application of denoising to any lung screening exam could make it usable for quantitative emphysema scoring. With slightly more validation, and an improved tuning-scheme, we would recommend that any study utilizing quantitative emphysema scoring as an end-point employ some form of denoising.

5.4 Final Thoughts

In this work, it has been shown that denoising has the potential to turn a non-ideal quantitative imaging test into something much closer to the ideal test presented in Chapter 1 (Figure 1-4). While more work is needed, these results highlight that there is a very clear potential path to extend quantitative emphysema scoring into routine clinical practice. Perhaps more importantly, the work presented in this dissertation lays the groundwork for many possible further studies that will hopefully result in more effective and accurate quantitative imaging. In particular, the types of tests and analyses presented here can be conducted much faster with the infrastructure developments of Chapter 2. Additionally, improved tuning schemes for the bilateral filter as well as the exploration of other denoising approaches would be straightforward and valuable extensions of this work. Finally, we hope that the free, open-source software tools provided will help provide a solid foundation for the future development of new reconstruction algorithms, possibly even those built with quantitative imaging in mind.

This work has concentrated on two attenuation-based quantitative measures of emphysema (RA-950 and PERC15), however other quantitative approaches are currently under investigation such as parametric response mapping (PRM) of the lung tissues [27], texture features [128], or deep learning-based schemes [129]. Broadly, we see two clear extensions of this work for other quantitative imaging tests. First is the potential for denoising to improve the robustness of quantitative imaging and quantitative imaging tasks. We believe that tasks that rely mostly on attenuation-based measurements will benefit from denoising; this could be densitometry (nodules, bone, etc.), the attenuation-based version of PRM, or automated segmentation algorithms. Tasks that may benefit from denoising are possibly volumetry or measures of "shape," however some additional care must be taken regarding the borders of the object being measured, since many denoising approaches fundamentally rely on blurring/averaging operations. Finally, we do not expect most "texture" image features (such as gray level co-occurrence matrix, Haralick features [130], etc.) to benefit from denoising, since denoising often fundamentally changes the expression of many features that impact image texture: noise magnitude, noise structure, edges, etc. Thus the denoising process could fundamentally damage potentially valuable textures, or create new textures that reflect the denoising process and not the underlying subject biology.

The second broad application of this work, is the methodology applied to test robustness of a given quantitative imaging approach. The tools and techniques developed here are readily transferred to other tasks through the development of custom analysis modules, and generation of custom, suitable datasets for the given task. If any quantitative task is to gain widespread clinical adoption, we foresee an evaluation similar to that provided here for RA-950 and PERC15, being necessary to build confidence in the approach. Because we have provided nearly all of our tools as free and open-source software, we hope that others begin testing their quantitative imaging techniques in a similar manner.

Bibliography

[1] N. Lung and S. Trial, “The National Lung Screening Trial: Overview and Study Design,” *Radiology*, vol. 258, no. 1, pp. 243–253, 2011.

[2] I. Isherwood, R. Rutherford, B. Pullan, and P. Adams, “Bone-Mineral Estimation By Computer-Assisted Transverse Axial Tomography,” *Lancet*, 1976.

[3] C. E. Cann, H. K. Genant, F. O. Kolb, and B. Ettinger, “Quantitative computed tomography for prediction of vertebral fracture risk,” *Bone*, vol. 6, no. 1, pp. 1–7, 1985.

[4] H. K. Genant, C. E. Cann, B. Ettinger, and G. S. Gordan, “Quantitative computed tomography of vertebral spongiosa: A sensitive method for detecting early bone loss after oophorectomy,” *Ann. Intern. Med.*, vol. 97, no. 5, pp. 699–705, 1982.

[5] W. A. Kalender, E. Klotz, and C. Suess, “Vertebral bone mineral analysis: an integrated approach with CT.,” *Radiology*, vol. 164, no. 2, pp. 419–423, 1987.

[6] N. L. Muller, C. A. Staples, R. R. Miller, and R. T. Abboud, “‘Density mask’. An objective method to quantitate emphysema using computed tomography,” *Chest*, vol. 94, no. 4, pp. 782–787, 1988.

[7] P. Genevois and Y. J.-C. De Vuyst P, de Maertelaer V, zanen J, Jacobovitz D, Cosio MG, “Comparison of Computed Density and Microscopic Morphometry in Pulmonary Emphysema,” *Am. J. Respir. Crit. Care Med.*, vol. 154, pp. 187–192, 1996.

[8] E. A. Zerhouni, J. F. Spivey, R. H. Morgan, F. P. Leo, F. P. Stitik, and S. S. Siegelman, “Factors Influencing Quantitative CT Measurements of Solitary Pulmonary Nodules,” *J. Comput. Assist. Tomogr.*, vol. 6, no. 6, 1982.

[9] E. A. Zerhouni *et al.*, “A Standard Phantom for Quantitative CT Analysis of Pulmonary Nodules,” *Radiology*, vol. 149, pp. 767–773, 1983.

[10] E. C. McCullough and R. L. Morin, “CT-number variability in thoracic geometry,” *Am. J. Roentgenol.*, vol. 141, no. 1, pp. 135–140, 1983.

[11] A. J. Buckler and R. Boellaard, “Standardization of quantitative imaging: the time is right, and 18F-FDG PET/CT is a good place to start.,” *J. Nucl. Med.*, vol. 52, no. 2, pp. 171–2, 2011.

[12] D. Couper *et al.*, “Design of the subpopulations and intermediate outcomes in COPD study (SPIROMICS),” *Thorax*, vol. 69, no. 5, pp. 491–494, 2014.

[13] E. A. Regan *et al.*, “Genetic epidemiology of COPD (COPDGene) study design,” *COPD J. Chronic Obstr. Pulm. Dis.*, vol. 7, no. 1, pp. 32–43, 2010.

[14] N. N. Jarjour *et al.*, “Severe asthma: Lessons learned from the National Heart, Lung, and Blood Institute Severe Asthma Research Program,” *American Journal of Respiratory and Critical Care Medicine*, vol. 185, no. 4. pp. 356–362, 2012.

[15] R. G. Barr *et al.*, “Subclinical atherosclerosis, airflow obstruction and emphysema: The MESA Lung Study,” *Eur. Respir. J.*, vol. 39, no. 4, pp. 846–854, 2012.

[16] B. R. Celli *et al.*, “Standards for the diagnosis and treatment of patients with COPD: A summary of the ATS/ERS position paper,” *Eur. Respir. J.*, vol. 23, no. 6, pp. 932–946, 2004.

[17] M. Heron, “National Vital Statistics Reports Deaths : Leading Causes for 2012,” 2015.

[18] P. R. Goddard, E. M. Nicholson, G. Laszlo, and I. Watt, “Computed tomography in pulmonary emphysema.,” *Clin. Radiol.*, vol. 33, pp. 379–387, 1982.

[19] A. Madani, A. Van Muylem, and P. A. Gevenois, “Pulmonary emphysema: effect of lung volume on objective quantification at thin-section CT.,” *Radiology*, vol. 257, no. 1, pp. 260–8, 2010.

[20] D. A. Lynch and M. A. Al-Qaisi, “Quantitative Computed Tomography in Chronic Obstructive Pulmonary Disease,” *J. Thorac. Imaging*, vol. 28, no. 5, pp. 284–290, 2013.

[21] A. Madani, J. Zanen, V. de Maertelaer, and P. A. Gevenois, “Pulmonary emphysema: objective quantification at multi-detector row CT--comparison with macroscopic and microscopic morphometry.,” *Radiology*, vol. 238, no. 3, pp. 1036–1043, 2006.

[22] C. P. Heussel *et al.*, “Fully automatic quantitative assessment of emphysema in computed tomography: Comparison with pulmonary function testing and normal values,” *Eur. Radiol.*, vol. 19, no. 10, pp. 2391–2402, 2009.

[23] A. Dirksen, “Monitoring the Progress of Emphysema by Repeat Computed Tomography Scans with Focus on Noise Reduction,” *Proc. Am. Thorac. Soc.*, vol. 5, no. 9, pp. 925–928, 2008.

[24] J. Stolk *et al.*, “Repeatability of lung density measurements with low-dose computed tomography in subjects with alpha-1-antitrypsin deficiency-associated emphysema.,” *Invest. Radiol.*, vol. 36, no. 11, pp. 648–651, 2001.

[25] M. Mishima *et al.*, “Complexity of terminal airspace geometry assessed by lung computed tomography in normal subjects and patients with chronic obstructive pulmonary disease.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 96, no. 16, pp. 8829–34, 1999.

[26] A. Madani, A. Van Muylem, V. de Maertelaer, J. Zanen, and P. A. Gevenois,

“Pulmonary Emphysema: Size Distribution of Emphysematous Spaces on Multidetector CT Images—Comparison with Macroscopic and Microscopic Morphometry,” *Radiology*, vol. 248, no. 3, pp. 1036–1041, 2008.

[27] C. J. Galbán *et al.*, “Computed tomography–based biomarker provides unique signature for diagnosis of COPD phenotypes and disease progression,” *Nat. Med.*, vol. 18, no. 11, pp. 1711–1715, 2012.

[28] J. L. Boes *et al.*, “Parametric Response Mapping Monitors Temporal Changes on Lung CT Scans in the Subpopulations and Intermediate Outcome Measures in COPD Study (SPIROMICS),” *Acad. Radiol.*, vol. 22, no. 2, pp. 186–194, 2015.

[29] J. JL Boes, M Bule, BA Hoff, R Chamberlain, DA Lynch, Stojanovska, F. J. Martinez, M. K. Han, E. A. Kazerooni, and D. Brian, “The Impact of Sources of Variability on Parametric Response Mapping of Lung CT Scans,” *Tomography*, vol. 1, no. 1, pp. 69–77, 2015.

[30] K. L. Boedeker *et al.*, “Emphysema: Effect of Reconstruction Algorithm on CT Imaging Measures,” *Radiology*, vol. 232, pp. 295–301, 2004.

[31] G. J. Kemerink, R. J. S. Lamers, G. R. P. Thelissen, and J. M. A. Van Engelshoven, “Scanner Conformity in CT Densitometry of the Lungs,” *J. Comput. Assist. Tomogr.*, vol. 20, no. 1, pp. 24–33, 1996.

[32] G. J. Kemerink, H. H. Kruize, R. J. S. Lamers, and J. M. A. Van Engelshoven, “CT lung densitometry: Dependence of CT number histograms on sample volume and consequences for scan protocol comparability,” *J. Comput. Assist. Tomogr.*, vol. 21, no. 6, pp. 948–954, 1997.

[33] D. S. Gierada *et al.*, “Comparison of standard- and low-radiation-dose CT for quantification of emphysema,” *Am. J. Roentgenol.*, vol. 188, no. 1, pp. 42–47, 2005.

[34] B. M. Trotta, A. V. Stolin, M. B. Williams, S. B. Gay, A. S. Brody, and T. A. Altes, “Characterization of the relation between CT technical parameters and accuracy of quantification of lung attenuation on quantitative chest CT,” *Am. J. Roentgenol.*, vol. 188, no. 6, pp. 1683–1690, 2007.

[35] D. S. Gierada *et al.*, “Effects of CT Section Thickness and Reconstruction Kernel on Emphysema Quantification. Relationship to the Magnitude of the CT Emphysema Index,” *Acad. Radiol.*, vol. 17, no. 2, pp. 146–156, 2010.

[36] M. Nishio *et al.*, “Emphysema quantification by low-dose CT: Potential impact of adaptive iterative dose reduction using 3D processing,” *Am. J. Roentgenol.*, vol. 199, no. 3, pp. 595–601, 2012.

[37] M. Nishio *et al.*, “Emphysema Quantification Using Ultralow-Dose CT With Iterative Reconstruction and Filtered Back Projection,” *AJR. Am. J. Roentgenol.*,

vol. 206, no. June, pp. 1–9, 2016.

[38] O. M. Mets *et al.*, "The effect of iterative reconstruction on computed tomography assessment of emphysema, air trapping and airway dimensions," *Eur. Radiol.*, vol. 22, no. 10, pp. 2103–2109, 2012.

[39] J. Y. Choo, J. M. Goo, C. H. Lee, C. M. Park, S. J. Park, and M. S. Shim, "Quantitative analysis of emphysema and airway measurements according to iterative reconstruction algorithms: Comparison of filtered back projection, adaptive statistical iterative reconstruction and model-based iterative reconstruction," *Eur. Radiol.*, vol. 24, no. 4, pp. 799–806, 2014.

[40] D. R. Aberle *et al.*, "Reduced Lung Cancer Mortality with Low-Dose Computed Tomographic Screening.," *N. Engl. J. Med.*, vol. 365, no. 5, pp. 395–409, 2011.

[41] C. H. Cagnon, D. D. Cody, M. F. McNitt-Gray, J. A. Seibert, P. F. Judy, and D. R. Aberle, "Description and implementation of a quality control program in an imaging-based clinical trial.," *Acad. Radiol.*, vol. 13, no. 11, pp. 1431–1441, 2006.

[42] J. P. Sieren *et al.*, "SPIROMICS protocol for multicenter quantitative computed tomography to phenotype the lungs," *Am. J. Respir. Crit. Care Med.*, vol. 194, no. 7, pp. 794–806, 2016.

[43] ACR, "CT Accreditation Phantom Instructions," pp. 1–14, 2013.

[44] T. A. A. of P. in Medicine, "THE ALLIANCE FOR QUALITY COMPUTED TOMOGRAPHY." [Online]. Available: https://www.aapm.org/pubs/CTProtocols/. [Accessed: 25-Jan-2018].

[45] S. Healthineers, "The Stellar Detector." [Online]. Available: https://www.healthcare.siemens.com/computed-tomography/technologies-innovations/stellar-detector. [Accessed: 25-Jan-2018].

[46] U. Tylen, O. Friman, M. Borga, and J.-E. Angelhed, "An improved algorithm for computerized detection and quantification of pulmonary emphysema at high resolution computed tomography (HRCT)," *Proc. SPIE*, vol. 4321, no. May 2001, pp. 254–262, 2001.

[47] R. A. Blechschmidt, R. Werthschützky, and U. Lörcher, "Automated CT Image Evaluation of the Lung : A Morphology-Based Concept," *IEEE Trans. Med. Imaging*, vol. 20, no. 5, pp. 434–442, 2001.

[48] L. Gallardo-Estrella *et al.*, "Normalizing computed tomography data reconstructed with different fi1 L. Gallardo-Estrella, D.A. Lynch, M. Prokop, et al., Normalizing computed tomography data reconstructed with different filter kernels: effect on emphysema quantification, Eur. Radiol. 2," *Eur. Radiol.*, vol. 26, no. 2, pp. 478–486, 2016.

[49] L. Gallardo-Estrella *et al.*, “Normalized emphysema scores on low dose CT: Validation as an imaging biomarker for mortality,” *PLoS One*, vol. 12, no. 12, pp. 1–12, 2017.

[50] F. J. Martinez *et al.*, “Predictors of mortality in patients with emphysema and severe airflow obstruction,” *Am. J. Respir. Crit. Care Med.*, vol. 173, no. 12, pp. 1326–1334, 2006.

[51] D. S. Gierada *et al.*, “Quantitative CT Assessment of Emphysema and Airways in Relation to Lung Cancer Risk,” *Radiology*, vol. 261, no. 3, pp. 950–959, 2011.

[52] S. G. Armato *et al.*, “The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans,” *Med. Phys.*, vol. 38, no. 2, pp. 915–931, 2011.

[53] H. J. Aerts *et al.*, “Data From NSCLC-Radiomics-Genomics,” *The Cancer Imaging Archive*, 2017. [Online]. Available: http://doi.org/10.7937/K9/TCIA.2015.L4FRET6Z. [Accessed: 13-Nov-2017].

[54] S. Young, H. J. G. Kim, M. M. Ko, W. W. Ko, C. Flores, and M. F. McNitt-Gray, “Variability in CT lung-nodule volumetry: Effects of dose reduction and reconstruction methods,” *Med. Phys.*, vol. 42, no. 5, pp. 2679–2689, 2015.

[55] S. Young *et al.*, “The effect of radiation dose reduction on computer-aided detection (CAD) performance in a low-dose lung cancer screening population,” *Med. Phys.*, vol. 44, no. 4, pp. 1337–1346, 2017.

[56] P. Lo, S. Young, H. J. Kim, M. S. Brown, and McNitt, “Variability in CT lung-nodule quantification : Effects of dose reduction and reconstruction methods on density and texture based features,” *Med. Phys.*, vol. 4854, no. 43, 2016.

[57] B. Chen, X. Duan, Z. Yu, S. Leng, L. Yu, and C. McCollough, “Technical Note: Development and validation of an open data format for CT projection data,” *Med. Phys.*, vol. 42, no. 12, pp. 6964–6972, 2015.

[58] S. Zabić, Q. Wang, T. Morton, and K. M. Brown, “A low dose simulation tool for CT systems with energy integrating detectors.,” *Med. Phys.*, vol. 40, no. 3, p. 31102, Mar. 2013.

[59] M. Kachelriess, O. Watzke, and W. a Kalender, “Generalized multi-dimensional adaptive filtering for conventional and spiral single-slice, multi-slice, and cone-beam CT,” *Med. Phys.*, vol. 28, no. 4, p. 475, 2001.

[60] J. Hsieh, “Adaptive streak artifact reduction in computed tomography resulting from excessive x-ray photon noise.,” *Med. Phys.*, vol. 25, no. 11, pp. 2139–2147, 1998.

[61] J. Hoffman, S. Young, F. Noo, and M. McNitt-Gray, “Technical Note : FreeCT _

wFBP : A robust , efficient , open-source implementation of weighted filtered backprojection for helical , fan-beam CT," *Med. Phys.*, vol. 43, no. 3, p. 10 pp., 2016.

[62] H. J. Aerts *et al.*, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nat Commun*, vol. 5, p. 4006, 2014.

[63] T. P. Coroller *et al.*, "CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma.," *Radiother. Oncol.*, vol. 114, no. 3, pp. 345–350, 2015.

[64] B. Zhao *et al.*, "Exploring intra- and inter-reader variability in uni-dimensional, bi-dimensional, and volumetric measurements of solid tumors on CT scans reconstructed at different slice intervals," *Eur. J. Radiol.*, vol. 82, no. 6, pp. 959–968, 2013.

[65] B. Zhao, Y. Tan, W. Y. Tsai, L. H. Schwartz, and L. Lu, "Exploring Variability in CT Characterization of Tumors: A Preliminary Phantom Study," *Transl. Oncol.*, vol. 7, no. 1, pp. 88–93, 2014.

[66] J. Solomon and E. Samei, "A generic framework to simulate realistic lung, liver and renal pathologies in CT imaging," *Phys. Med. Biol.*, vol. 59, no. 21, pp. 6637–6657, 2014.

[67] M. S. Brown *et al.*, "Toward clinically usable CAD for lung cancer screening with computed tomography," *Eur. Radiol.*, vol. 24, no. 11, pp. 2719–2728, 2014.

[68] M. F. McNitt-Gray *et al.*, "Determining the Variability of Lesion Size Measurements from CT Patient Data Sets Acquired under 'No Change' Conditions," *Transl. Oncol.*, vol. 8, no. 1, pp. 55–64, 2015.

[69] N. Petrick *et al.*, "Comparison of 1D, 2D, and 3D Nodule Sizing Methods by Radiologists for Spherical and Complex Nodules on Thoracic CT Phantom Images," *Acad. Radiol.*, vol. 21, no. 1, pp. 30–40, 2014.

[70] P. Lo, S. Young, H. J. G. Kim, J. Hoffman, M. Brown, and M. F. McNitt-Gray, "The Effects of CT Acquisition and Reconstruction Conditions On Computed Texture Feature Values of Lung Lesions," in *AAPM*, 2015.

[71] S. Rit, M. Vila Oliva, S. Brousmiche, R. Labarbe, D. Sarrut, and G. C. Sharp, "The Reconstruction Toolkit (RTK), an open-source cone-beam CT reconstruction toolkit based on the Insight Toolkit (ITK)," *J. Phys. Conf. Ser.*, vol. 489, no. 1, p. 12079, 2014.

[72] A. Maier *et al.*, "CONRAD--a software framework for cone-beam imaging in radiology.," *Med. Phys.*, vol. 40, no. 11, p. 111914, 2013.

[73] J. M. Boone *et al.*, "OSCaR : Open Source Cone-beam Reconstructor," no. 1, pp. 1–25.

[74] X. Tang, J. Hsieh, R. a Nilsen, S. Dutta, D. Samsonov, and A. Hagiwara, “A three-dimensional-weighted cone beam filtered backprojection (CB-FBP) algorithm for image reconstruction in volumetric CT-helical scanning.,” *Phys. Med. Biol.*, vol. 51, no. 4, pp. 855–874, 2006.

[75] K. Stierstorfer, A. Rauscher, J. Boese, H. Bruder, S. Schaller, and T. Flohr, “Weighted FBP - a simple approximate 3D FBP algorithm for multislice spiral CT with good dose usage for arbitrary pitch,” *Phys. Med. Biol.*, vol. 49, no. 11, pp. 2209–2218, Jun. 2004.

[76] H. Kudo, T. Rodet, F. Noo, and M. Defrise, “Exact and approximate algorithms for helical cone-beam CT,” *Phys. Med. Biol.*, vol. 49, no. 13, pp. 2913–2931, 2004.

[77] D. Heuscher, K. Brown, and F. Noo, “Redundant data and exact helical cone-beam reconstruction.,” *Phys. Med. Biol.*, vol. 49, no. 11, pp. 2219–2238, 2004.

[78] UCLA CT Physics and Recon Group, “FreeCT Home Page.” [Online]. Available: http://cvib.ucla.edu/freect. [Accessed: 23-Jan-2018].

[79] J. Hsieh and X. Tang, “Tilted cone-beam reconstruction with row-wise fan-to-parallel rebinning.,” *Phys. Med. Biol.*, vol. 51, no. 20, pp. 5259–76, Oct. 2006.

[80] T. G. Flohr, K. Stierstorfer, S. Ulzheimer, H. Bruder, A. N. Primak, and C. H. McCollough, “Image reconstruction and image quality evaluation for a 64-slice CT scanner with z-flying focal spot,” *Med Phys*, vol. 32, no. 8, pp. 2536–2547, 2005.

[81] M. Kachelrieß, M. Knaup, C. Penßel, and W. A. Kalender, “Flying Focal Spot (FFS) in Cone – Beam CT,” vol. 0, no. C, pp. 3759–3763, 2004.

[82] P. J. La Rivière and X. Pan, “Sampling and aliasing consequences of quarter-detector offset use in helical CT,” *IEEE Trans. Med. Imaging*, vol. 23, no. 6, pp. 738–749, 2004.

[83] K. Sourbelle, “FORBILD Thorax Phantom,” 2014. [Online]. Available: http://www.imp.uni-erlangen.de/phantoms/thorax/thorax.htm.

[84] J. Hoffman, F. Noo, K. Mcmillan, S. Young, and M. McNitt-Gray, “Assessing nodule detection on lung cancer screening CT: the effects of tube current modulation and model observer selection on detectability maps,” in *Proc. SPIE Medical Imaging*, 2016.

[85] B. Zhao *et al.*, “Evaluating Variability in Tumor Measurements from Same-day Repeat CT Scans of Patients with Non–Small Cell Lung Cancer,” *Radiology*, vol. 252, no. 1, pp. 263–272, Jul. 2009.

[86] E. Zheng, Yuese; Solomon, Justin; Choudhury, Kingshuk; Marin, Daniele ; Samei, “Accuracy and variability of texture-based radiomics features of lung lesions across CT imaging conditions,” in *SPIE Medical Imaging*, 2017, p. 101325F.

[87] P. J. Pickhardt *et al.*, “Abdominal CT With Model-Based Iterative Reconstruction (MBIR): Initial Results of a Prospective Trial Comparing Ultralow-Dose With Standard-Dose Imaging,” *AJR. Am. J. Roentgenol.*, vol. 199, no. December, pp. 1266–1274, 2012.

[88] J. Xu and B. M. W. Tsui, “Iterative image reconstruction in helical cone-beam x-ray CT using a stored system matrix approach.,” *Phys. Med. Biol.*, vol. 57, no. 11, pp. 3477–97, 2012.

[89] P. M. Joseph, “An Improved Algorithm for Reprojecting Rays Through Pixel Images,” *IEEE Trans. Med. Imag.*, vol. 1, no. 2, pp. 192–196, Nov. 1982.

[90] K. Hahn, H. Schöndube, K. Stierstorfer, J. Hornegger, and F. Noo, “A comparison of linear interpolation models for iterative CT reconstruction,” *Med. Phys.*, vol. 43, no. 12, pp. 6455–6473, 2016.

[91] M. Guo and H. Gao, “Memory-Efficient Algorithm for Stored Projection and Backprojection Matrix in Helical CT,” *Med. Phys.*, vol. 44, no. 4, pp. 1287–1300, 2017.

[92] “Open MP Libraries.” [Online]. Available: http://www.openmp.org/. [Accessed: 23-Jan-2018].

[93] J. Hoffman, S. Young, F. Noo, and M. McNitt-Gray, “Technical Note: FreeCT_wFBP: A robust, efficient, open-source implementation of weighted filtered backprojection for helical, fan-beam CT,” *Med. Phys.*, vol. 43, no. 3, pp. 1411–1420, 2016.

[94] American College of Radiology, “CT Accreditation Phantom Instructions,” 2013. .

[95] D. Matenine, G. Côté, J. Mascolo-Fortin, Y. Goussard, and P. Després, “System matrix computation vs storage on GPU: a comparative study in cone beam CT,” *Med. Phys.*, 2017.

[96] J. Hoffman, “CTBB Pipeline Github,” 2017. [Online]. Available: https://github.com/captnjohnny1618/CTBB_Pipeline_Package. [Accessed: 13-Nov-2017].

[97] S. Young, J. M. Hoffman, F. Noo, and M. F. McNitt-Gray, “Vendor-Independent, Model-Based Iterative Reconstruction On a Rotating Grid with Coordinate-Descent Optimization for CT Imaging Investigations,” in *Annual Meeting of the AAPM*, 2016.

[98] G. Lauritsch and H. Bruder, “FORBILD Head Phantom.” .

[99] D. Heuscher, K. Brown, and F. Noo, “Redundant data and exact helical cone-beam reconstruction,” *Phys. Med. Biol.*, vol. 49, no. 11, pp. 2219–2238, 2004.

[100] K. Hahn, H. Schondube, K. Stierstorfer, and F. Noo, “Impact of statistical weights and edge preserving regularization on image quality in iterative CT reconstruction,” *Fully 3D*, no. lcd, pp. 51–54, 2015.

[101] K. Stierstorfer, T. Flohr, and H. Bruder, “Segmented multiple plane reconstruction: a novel approximate reconstruction scheme for multi-slice spiral CT.,” *Phys. Med. Biol.*, vol. 47, no. 15, pp. 2571–81, Aug. 2002.

[102] T. Martin, J. Hoffman, J. R. Alger, M. Mcnitt-Gray, and D. J. Wang, “Low-dose CT perfusion with projection view sharing,” *Med. Phys.*, pp. 101–113, 2017.

[103] Z. Yu, F. Noo, F. Dennerlein, A. Wunderlich, G. Lauritsch, and J. Hornegger, “Simulation tools for two-dimensional experiments in x-ray computed tomography using the FORBILD head phantom,” *Phys. Med. Biol.*, vol. 57, no. 13, pp. N237–N252.

[104] F. Noo, J. Pack, and D. Heuscher, “Exact helical reconstruction using native cone-beam geometries.,” *Phys. Med. Biol.*, vol. 48, no. 23, pp. 3787–3818, 2003.

[105] M. Kachelrieß, S. Schaller, and W. a. Kalender, “Advanced single-slice rebinning in cone-beam spiral CT,” *Med. Phys.*, vol. 27, no. 4, p. 754, 2000.

[106] J. Hoffman, G. Kim, J. Goldin, M. Brown, and M. Mcnitt-Gray, “A Pilot Study Evaluating the Robustness of Density Mask Scoring (RA-950), a Quantitative Measure of Chronic Obstructive Pulmonary Disease, to CT Parameter Selection Using a High-Throughput, Automated, Computational Research Pipeline,” in *Annual Meeting of the AAPM*, 2017.

[107] J. Hoffman, G. Kim, J. Goldin, M. Brown, and M. McNitt-Gray, “Robustness Evaluation of RA-950 Scoring in a Cohort of CT Lung Screening Patients Across a Large Range of CT Acquisition and Reconstruction Conditions,” in *Radiological Society of North America 2017 Scientific Assembly and Annual Meeting*, 2017.

[108] D. Blackwell, J. Lucas, and T. Clarke, “Summary health statistics for U.S. adults: National Health Interview Survey, 2012.,” 2014.

[109] A. S. Buist *et al.*, “International variation in the prevalence of COPD (the BOLD Study): a population - based prevalence study . PubMed Commons,” *Lancet (London, England)*, vol. 370, no. 9589, pp. 741–50, 2015.

[110] L. B. Gerald and W. C. Bailey, “Global initiative for chronic obstructive lung disease,” *J. Cardiopulm. Rehabil.*, vol. 22, no. 4, pp. 234–244, 2002.

[111] M. Kirby *et al.*, “Management of COPD: Is there a role for quantitative imaging?,” *Eur. J. Radiol.*, vol. 86, pp. 335–342, 2016.

[112] N. Emaminejad *et al.*, “WE-G-201-4: Evaluation of CAD Nodule Detection Performance in Low Dose CT Lung Cancer Screening Across a Range of Dose

Levels, Slice Thicknesses and Reconstruction Kernels," in *58th Annual Meeting and Exhibition of the AAPM*, 2017.

[113] K. Grant and R. Raupach, "SAFIRE : Sinogram Affirmed Iterative Reconstruction," *(WHITE Pap. Somat. Sess.*, pp. 1–8, 2012.

[114] W. J. Kostis, S. C. Fluture, D. F. Yankelevitz, and C. I. Henschke, "Method for analysis and display of distribution of emphysema in CT scans," vol. 5032, no. May 2003, pp. 199–206, 2003.

[115] A. M. R. Schilham, B. Van Ginneken, H. Gietema, and M. Prokop, "Local noise weighted filtering for emphysema scoring of low-dose CT images," *IEEE Trans. Med. Imaging*, vol. 25, no. 4, pp. 451–463, 2006.

[116] M. Nishio *et al.*, "Convolutional auto-encoders for image denoising of ultra-low-dose CT," *Heliyon*, vol. 3, no. 8, p. e00393, 2017.

[117] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," *Sixth Int. Conf. Comput. Vis. (IEEE Cat. No.98CH36271)*, pp. 839–846, 1998.

[118] F. Banterle, M. Corsini, P. Cignoni, and R. Scopigno, "A low-memory, straightforward and fast bilateral filter through subsampling in spatial domain," *Comput. Graph. Forum*, vol. 31, no. 1, pp. 19–32, 2012.

[119] Q. Yang, A. Maier, N. Maass, and J. Hornegger, "Edge-preserving bilateral filtering for images containing dense objects in CT," *IEEE Nucl. Sci. Symp. Conf. Rec.*, pp. 0–4, 2013.

[120] M. D. Lekan, "Impact of bilateral filter parameters on medical image noise reduction and edge preservation," University of Toledo, 2009.

[121] A. Manduca *et al.*, "Projection space denoising with bilateral filtering and CT noise modeling for dose reduction in CT," *Med Phys*, vol. 36, no. 11, pp. 4911–4919, 2009.

[122] W. P. Segars, G. Sturgeon, S. Mendonca, J. Grimes, and B. M. W. Tsui, "4D XCAT phantom for multimodality imaging research.," *Med. Phys.*, vol. 37, no. 9, pp. 4902–4915, 2010.

[123] H. J. G. Kim *et al.*, "Classification of parenchymal abnormality in scleroderma lung using a novel approach to denoise images collected via a multicenter study," vol. 18, no. 9, pp. 1199–1216, 2013.

[124] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising with block-matching and 3D filtering," vol. 6064, p. 606414, 2006.

[125] A. Buades, B. Coll, and J. M. Morel, "A Review of Image Denoising Algorithms, with a New One," *Multiscale Model. Simul.*, vol. 4, no. 2, pp. 490–530, 2005.

[126] H. a Gietema, A. M. Schilham, B. van Ginneken, R. J. van Klaveren, J. W. J. Lammers, and M. Prokop, “Monitoring of smoking-induced emphysema with CT in a lung cancer screening setting: detection of real increase in extent of emphysema.,” *Radiology*, vol. 244, no. 3, pp. 890–897, 2007.

[127] C. H. Martinez *et al.*, “Relationship between quantitative CT metrics and health status and BODE in chronic obstructive pulmonary disease,” *Thorax*, vol. 67, no. 5, pp. 399–406, 2012.

[128] D. V. Fried *et al.*, “Prognostic value and reproducibility of pretreatment ct texture features in stage III non-small cell lung cancer,” *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 90, no. 4, pp. 834–842, 2014.

[129] M. Gao, Z. Xu, L. Lu, A. Harrison, R. Summers, and D. Mollura, “Holistic Interstitial Lung Disease Detection using Deep Convolutional Neural Networks: Multi-label Learning and Unordered Pooling,” 2017.

[130] R. M. Haralick, “Statistical and structural approaches to texture,” *Proc. IEEE*, vol. 67, no. 5, pp. 786–804, 1979.