

5-2019

QUANTITATIVE IMAGING FOR PRECISION MEDICINE IN HEAD AND NECK CANCER PATIENTS

Rachel Ger

Follow this and additional works at: https://digitalcommons.library.tmc.edu/utgsbs_dissertations



Part of the [Medicine and Health Sciences Commons](#)

Recommended Citation

Ger, Rachel, "QUANTITATIVE IMAGING FOR PRECISION MEDICINE IN HEAD AND NECK CANCER PATIENTS" (2019). *The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences Dissertations and Theses (Open Access)*. 923.
https://digitalcommons.library.tmc.edu/utgsbs_dissertations/923

This Dissertation (PhD) is brought to you for free and open access by the The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences at DigitalCommons@TMC. It has been accepted for inclusion in The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences Dissertations and Theses (Open Access) by an authorized administrator of DigitalCommons@TMC. For more information, please contact digitalcommons@library.tmc.edu.

QUANTITATIVE IMAGING FOR PRECISION MEDICINE IN HEAD AND NECK CANCER PATIENTS

by

Rachel Beth Ger, B.S.

APPROVED:

Laurence E. Court, Ph.D.
Advisory Professor

Clifton D. Fuller, M.D., Ph.D.

Rebecca M. Howell, Ph.D.

Rick R. Layman, Ph.D.

Heng Li, Ph.D.

R. Jason Stafford, Ph.D.

Shouhao Zhou, Ph.D.

APPROVED:

Dean, The University of Texas

MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences

QUANTITATIVE IMAGING FOR PRECISION MEDICINE IN HEAD AND NECK CANCER PATIENTS

A

DISSERTATION

Presented to the Faculty of

The University of Texas

MD Anderson Cancer Center UTHealth

Graduate School of Biomedical Sciences

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

by

Rachel Beth Ger, B.S.
Houston, Texas

May, 2019

Dedication

Dedicated to my parents. You have been my role models from day one and have shown me to never give up. I owe all of my success to you. Thank you for always supporting me and giving me all of the skills to succeed.

Acknowledgments

I would like to start by thanking my advisor, Dr. Laurence Court for his constant support and perpetual optimism. For all the times that I came into his office with my head down after experiments produced negative results, he always managed to see the positive that was drawn from it. I would not be the researcher or person I am today without the knowledge you have given me over the past five years. I am so thankful that I was able to stumble across your lab group right before I made my decision on which group to join.

Thank you to my committee Drs. Clifton Fuller, Rebecca Howell, Rick Layman, Heng Li, Jason Stafford, and Shouhao Zhou. Your input at every meeting has been invaluable and has helped shape my project and me.

Thank you Dr. Rebecca Howell for being my constant point of reference and female role model. You have always made yourself available whenever I have needed advice, whether it was project related, career related, or just about general life. Your perpetual encouragement of me throughout all of the stages of my graduate career have helped me become the strong, confident researcher that I am.

Thank you to my parents who listened to me talk non-stop about my project despite me never really explaining it well enough to actually understand what my issue of the day was. I never could have made it this far without your constant support of me in all of my endeavors.

Thank you to every member of the Courtyard, past and present, who have all contributed to my success. The team science and complete environment of support have made the past five years beyond enjoyable. I will miss all of our Science Clubs, paper-in-a-days, ideas-fests, happy-hours, and every other little lab activity.

QUANTITATIVE IMAGING FOR PRECISION MEDICINE IN HEAD AND NECK CANCER PATIENTS

Rachel Beth Ger, B.S.

Advisory Professor: Laurence Court, Ph.D.

The purpose of this work was to determine if prediction models using quantitative imaging measures in head and neck squamous cell carcinoma (HNSCC) patients could be improved when noise due to imaging was reduced. This was investigated separately for salivary gland function using dynamic contrast enhanced magnetic resonance imaging (DCE-MRI), overall survival using computed tomography (CT)-based radiomics, and overall survival using positron emission tomography (PET)-based radiomics. From DCE-MRI, where T1-weighted images are serially acquired after injection of contrast, quantitative measures of diffusion can be obtained from the series of images. Radiomics is the study of the relationship of voxels to one another providing measures of texture from the area of interest. Quantitative information obtained from imaging could help in radiation treatment planning by providing quantifiable spatial information with computational models for assigning dose to regions to improve patient outcome, both survival and quality of life. By reducing the noise within the quantitative data, the prediction accuracy could improve to move this type of work closer to clinical practice.

For each imaging modality sources of noise that could impact the patient analysis were identified, quantified, and if possible minimized during the patient analysis. In MRI, a large potential source of uncertainty was the image registration. To evaluate this, both physical and synthetic phantoms were used, which showed that registration of MR images was high, with all root mean square errors below 3 mm. Then, 15 HNSCC patients with pre-, mid-, and post-treatment DCE-MRI scans were evaluated. However, differences in algorithm output were found to be a large source of noise as different algorithms could not consistently rank patients as above or below the median for quantitative metrics from DCE-MRI. Therefore, further analysis using this modality was not pursued.

In CT, a large potential source of noise that could impact patient analysis was the inter-scanner variability. To investigate this a controlled protocol was designed and used to image, along with the local head and chest protocols, a radiomics phantom on 100 CT scanners. This demonstrated that the inter-scanner variability could be reduced by over 50% using a controlled protocol compared to local protocols. Additionally, it was shown that the reconstruction parameters impact feature values while most acquisition parameters do not, therefore, most of this benefit can be achieved using a radiomics reconstruction with no additional dose to the patient. Then to evaluate this impact in patient studies, 726 HNSCC patients with CT images were used to create and test a Cox proportional hazards model for overall survival. Those patients with the same imaging protocol were subset and a new Cox proportional hazards model was created and tested in order to determine if the reduction in noise due to controlling the imaging protocol translated into improved prediction. However, noise between patient populations from different institutions was shown to be larger than the reduction in noise due to a controlled imaging protocol.

In PET, a large potential source of noise that could impact patient analysis was the imaging protocol. A phantom scanned on three different scanners and vendors demonstrated that on a single vendor, imaging parameter choices did not affect radiomics feature values, but inter-scanner variances could be large. Then, 686 HNSCC patients with PET images were used to create and test a Cox proportional hazards model for overall survival. Those patients with the same imaging protocol were subset and a new Cox proportional hazards model was created and tested in order to determine if the reduction in noise due to controlling the imaging protocol on a vendor translated into improved prediction. However, no predictive radiomics signature could be determined for any subset of the patient cohort that resulted in significant stratification of patients into high and low risk.

This study demonstrated that the imaging variability could be quantified and controlled for in each modality. However, for each modality there were larger sources of noise identified that did not

allow for improvement in prediction modeling of salivary gland function or overall survival using quantitative imaging metrics for MRI, CT, or PET.

Table of Contents

Dedication	iii
Acknowledgments	iv
List of Illustrations	xii
List of Tables	xiv
Chapter 1 : Introduction	1
Chapter 2 : Purpose and Central Hypothesis	5
Chapter 3 : Magnetic Resonance Imaging Registration Uncertainty	6
3.1 Introduction	6
3.2 Methods and Materials	7
3.2.1 Porcine Phantom	7
3.2.2 Synthetic Images	10
3.2.3 Imaging	14
3.2.4 Registration Techniques	15
3.2.5 Statistical Methods	18
3.3 Results	18
3.3.1 Porcine Phantom	18
3.3.2 Synthetic Images	21
3.4 Discussion	26
3.5 Conclusions	28
Chapter 4 : Dynamic Contrast-Enhanced Magnetic Resonance Imaging Feature Stability	29
4.1 Introduction	29
4.2 Methods and Materials	30
4.2.1 Algorithms	30
4.2.2 DROs	33
4.2.3 Patients	33
4.2.4 Statistical Methods	37
4.3 Results	39
4.3.1 DROs	39
4.3.2 Patients	44
4.4 Discussion	48
Chapter 5 : Computed Tomography Radiomics Feature Dependence on Tube Voltage	54
5.1 Introduction	54
5.2 Methods	55

5.2.1 <i>Effective Atomic Number of Phantom Materials</i>	55
5.2.2 <i>Phantom Scans</i>	56
5.2.3 <i>Patient Scans</i>	57
5.2.4 <i>Radiomics Feature Analysis</i>	58
5.2.5 <i>Statistical Methods</i>	60
5.3 Results	61
5.3.1 <i>Effective Atomic Number</i>	61
5.3.2 <i>Spearman Correlation of Features with Tube Voltage</i>	63
5.3.3 <i>Patient-Normalized Phantom Range</i>	67
5.4 Discussion.....	70
5.5 Conclusions	74
Chapter 6 : Impact of Head and Neck Artifacts on Computed Tomography Radiomics Features	76
6.1 Introduction	76
6.2 Methods and Materials	77
6.2.1 <i>Streak Artifact</i>	77
6.2.2 <i>Bone Artifact</i>	81
6.3 Results	84
6.3.1 <i>Streak Artifacts</i>	84
6.3.2 <i>Bone Artifact</i>	88
6.4 Discussion.....	90
6.5 Conclusion	91
Chapter 7 : Inter-Scanner Variability of Radiomics Features on Computed Tomography Scanners	92
7.1 Introduction	92
7.2 Methods	93
7.2.1 <i>Methods and Materials</i>	93
7.2.2 <i>CT Scans</i>	95
7.2.3 <i>Patient Scans</i>	95
7.2.4 <i>Radiomics Feature Extraction</i>	96
7.2.5 <i>Statistical Methods</i>	99
7.3 Results	102
7.3.1 <i>Scanners</i>	102
7.3.2 <i>Feature Stability</i>	105
7.3.3 <i>Resample the z Dimension</i>	105
7.3.4 <i>Imaging Variability</i>	107

7.3.5 Quality Assurance Using a Radiomics Phantom	115
7.4 Discussion	118
7.5 Conclusion	123
Chapter 8 : PET Imaging Protocol Effect on Radiomics Feature Values	124
8.1 Introduction	124
8.2 Methods	124
8.2.1 Phantom Scans	124
8.2.2 Patients	127
8.2.3 Feature Extraction	128
8.2.4 Statistical Analysis	131
8.3 Results	131
8.4 Discussion	135
8.5 Conclusion	137
Chapter 9 : CT- and PET-Based Radiomics Survival Modeling of HNSCC Patients	138
9.1 Introduction	138
9.2 Materials and Methods	138
9.2.1 CT Patients	138
9.2.2 PET Patients	139
9.2.3 Feature Extraction	143
9.2.4 Model Building	143
9.3 Results	145
9.3.1 CT Patients	145
9.3.2 PET Patients	148
9.4 Discussion	150
9.5 Conclusion	152
Chapter 10 : Discussion	153
Future Applications	156
Conclusion	159
Appendix A: Supplemental Material for Chapter 4	160
Appendix B: Supplemental Material for Chapter 5	168
Appendix C: Supplemental Material for Chapter 6	237
Appendix D: Supplemental Material Chapter 7	241
Appendix E: Supplemental Material for Chapter 8	259
Appendix F: Supplemental Material for Chapter 9	271

Bibliography:.....	276
Vita	288

List of Illustrations

Figure 3-1: Porcine Phantom Representative Deformation.	9
Figure 3-2: Workflow of Synthetic Image Generation.	12
Figure 3-3: Parotid Deformation in Synthetic Image Generation.	13
Figure 3-4: Porcine Phantom Registration Error.	20
Figure 3-5: Root Mean Square (RMS) Applied Deformation and Registration Error in T1-Weighted Synthetic Images.	22
Figure 3-6: Root Mean Square (RMS) Applied Deformation and Registration Error in T2-Weighted Synthetic Images.	24
Figure 4-1: No Noise DRO Performance.	40
Figure 4-2: Heat Maps of DRO Error.	42
Figure 4-3: Percent of Values Removed.	45
Figure 4-4: Differences in Ktrans from Algorithms.	47
Figure 5-1: Acrylic and Cork Changes with Tube Voltage.	66
Figure 5-2: Heat map of Patient-Normalized Phantom Range Values.	68
Figure 6-1: Streak Artifact.	78
Figure 6-2: CT Images of Bone Phantom.	82
Figure 7-1: Radiomics Phantom used for Inter-Scanner Analysis.	94
Figure 7-2: Image Thickness Histograms.	104
Figure 7-3: Pearson Correlation of Feature Value with Image Thickness.	106
Figure 7-4: Bar Plots of Variation from Head and Controlled Protocols.	110
Figure 7-5: Histograms of Number of Scanners that have a Percentage of features outside of patient bounds.	117
Figure 8-1: Slices of Hoffman Phantom.	129
Figure 8-2: Bar Plots of Features by Reliability Level.	134

Figure 9-1: Patient survival curves for CT radiomics models.	146
Figure 9-2: Patient survival curves for PET radiomics models.	149

List of Tables

Table 3-1. Parameter Settings for the Dual-force Demons Deformable Image Registration	17
Table 3-2. Maximum Registration Error from Synthetic Images and the Porcine Phantom.....	25
Table 4-1. Description of Algorithms	32
Table 4-2. Study Patient Demographics.....	35
Table 5-1. Radiomics Features Analyzed.....	59
Table 5-2. Estimated Atomic Number of Phantom Cartridges and Human Tissues.....	62
Table 5-3. Percentage of Features with Significant Spearman Correlation.....	64
Table 6-1. Percentage of features with significantly different values with streak artifacts and with artifact slices removed	85
Table 6-2. Number of features that were not robust across the volume range for each slice removal and preprocessing technique	87
Table 6-3. Number of features with significantly different values when measured in a phantom with a PLA or PVC rod	89
Table 7-1. Radiomics Features Analyzed.....	98
Table 7-2. Number of features for each protocol and preprocessing type that have imaging variability compared to inter-patient variability from linear mixed-effects models above the cutoff	112
Table 8-1. Parameters Changed to Investigate Impact on Radiomics Features	126
Table 8-2. Radiomics Features Used in PET Analysis.....	130
Table 9-1. Patient Demographics	141

Chapter 1 : Introduction

Head and neck squamous cell carcinoma (HNSCC) is the sixth most common cancer worldwide with more than 500,000 new cases annually [1]. The standard of care for HNSCC patients includes radiation therapy. Over a 40-year span with advances in radiation therapy and other cancer treatment modalities, the 5-year survival has improved from 53% in 1975 to 69% in 2010 [2]. However, radiation therapy also impacts normal tissues. Xerostomia, salivary gland dysfunction which adversely impacts swallowing and taste (and thus eating) as well as speech, is common due to the proximity of the normal tissue structures to the tumor: over 90% of patients suffer from xerostomia during treatment and 68% still suffer two years after treatment even with modern treatments like intensity modulated radiation therapy [3]. While there have been improvements in patient survival and localization of radiation to reduce xerostomia, there is evidence that quantitative imaging could further improve these. Dynamic-contrast enhanced magnetic resonance imaging (DCE-MRI) and radiomics have shown potential in being able to do this.

The current radiation treatment plan standard of care for HNSCC uses one assigned dose to the tumor and one dose to stay below for each normal tissue structure to avoid toxicity, ignoring the known heterogeneity in these tissues. Quantitative information obtained from imaging could help in radiation treatment planning by providing quantifiable spatial information with computational models for assigning dose to regions to improve patient outcome.

In this dissertation we examined three different approaches to understanding and predicting patient outcomes from radiotherapy: DCE-MRI for salivary gland function, CT-based radiomics for overall survival, and PET-based radiomics for overall survival. In each case we first examined the expected major sources of uncertainty, before examining whether these approaches could provide additional information to inform clinical decision making by improving prediction models that included quantitative imaging metrics.

DCE-MRI

DCE-MRI is a noninvasive tool for examination of the microvasculature of tumors and normal tissue which uses T1-weighted MRI scans acquired serially after injection of a contrast agent. The perfusion and permeability metrics estimated from pharmacokinetic modeling of DCE-MRI data may provide an indirect measure of tumor hypoxia, a condition associated with poor prognosis in HNSCC [4, 5]. Therefore, it may be possible to build prognostic models to help tailor HNSCC treatments to individual patients based on that patient's DCE-MRI signature. Investigators have used DCE-MRI to assess therapeutic response of HNSCC and have shown associations between DCE-MRI metrics and changes in salivary glands and mandible [6-11].

Additionally, a sub-region within the parotid glands was identified which was directly associated with salivary gland function after one year [12]. This region is believed to be the stem cell region of the parotid gland, thus the spatial relationship of dose within the salivary glands is important. As patient quality of life is one of the primary concerns after radiation therapy, identifying an individual patient's risk for developing normal tissue complications is vital to creating and adapting individualized radiation therapy plans and determining interventions to mitigate negative effects. Previous studies have only assessed DCE-MRI of parotid glands post-treatment using global metrics, e.g. mean. With the knowledge of this stem cell region within parotid glands and the ability of DCE-MRI to quantitatively describe the microvasculature, this could allow voxel-based tracking of these DCE-MRI parameters in the salivary glands to further define the parotid stem cell region for patients. This would allow improved guidelines for normal tissue doses in salivary glands which would result in improved patient quality of life.

Radiomics

Radiomics involves evaluating images on a voxel level to extract quantitative image features (i.e. texture). This process relies on the assumption that there is more information contained within the

images than the human eye can extract and these textures and patterns are related to the gene microenvironment within that tumor or tissue [13]. Interest in radiomics has grown as radiomics features have been shown to improve survival models when combined with conventional prognostic factors (e.g., age) [14-21].

Radiomics studies have primarily been conducted on images from non-small cell lung cancer (NSCLC) patients. Using computed tomography (CT) images, Fried et al. were able to identify radiomics features that significantly improved risk stratification compared to conventional prognostic factors alone for overall survival, locoregional control, and freedom from distant metastases [22]. Similarly, Fave et al. were able to identify radiomics and delta-radiomics features that significantly improved risk stratification [14]. Thawani et al. summarized 11 CT radiomics studies using NSCLC patients where each study chose a different assortment of radiomics features to study, all found at least one radiomics feature that created a significant model related to survival [23]. Radiomics studies in NSCLC have found similar results using positron emission tomography (PET) images: Fried et al. were able to identify radiomics feature correlated with survival and additionally use it to identify subgroups of patients who did or did not receive a benefit from dose escalation [15, 24]; and Cook et al. summarized four other studies that also found radiomics features associated with patient outcomes [25].

This analysis has more recently transferred into head and neck patients. Studies using CT images of head and neck patients have found radiomics features significantly associated with local control, tumor failure, overall survival, and human papillomavirus (HPV) status [17, 18, 20, 26-29]. Similar findings using PET images of head and neck patients have been found where radiomics features were significantly associated with local control, tumor failure, overall survival, and freedom from distant metastases [18, 29-31].

However, there are known imaging protocol variabilities that can add noise to patient cohorts in studies. For CT images, the impacts of differences in kernel, pixel size, image thickness, and tube current have been studied [32-38]. For parameters such as pixel size, it has been shown that resampling

can reduce imaging differences [32, 37], while for parameters such as the reconstruction kernel, it has been shown that combining patient data that includes both sharp and smooth kernels can lead to large discrepancies [34]. For PET images, acquisition and reconstruction parameters have been shown to impact radiomics features; particularly, the number of iterations, matrix size, and smoothing filter have demonstrated variability, mostly in lung patient cohorts [39-49]. Reuzé et al. even demonstrated that models using radiomics features developed on one scanner may not be applicable to images from a different scanner [50].

Radiomics studies require large patient cohorts to power the modeling process. Due to this, images acquired under different imaging protocols from the same institution or different institutions are included in a patient cohort. If a protocol could be designed to minimize the known imaging protocol variabilities, this would reduce the noise within a patient cohort. This would likely result in improved patient outcome modeling performance.

Study Goal

The goals of this study were to determine if DCE-MRI parameters could be correlated with salivary gland dose response on a voxel basis, and to determine if minimizing CT and PET imaging protocol variabilities improves patient outcome modeling when using radiomics features. The DCE-MRI parameters would allow to better define salivary gland dose limits when creating patient treatment plans. If radiomics models could be improved by reducing the noise in the patient cohorts through imaging protocols, this would encourage others to focus on only including patients whose imaging protocols matched, thus producing better studies and allowing radiomics to approach use within clinics. These two quantitative measures, DCE-MRI and radiomics, were used to investigate improving both salivary gland function models and overall survival models.

Chapter 2 : Purpose and Central Hypothesis

Central Hypothesis:

DCE-MRI parameter changes during treatment are associated with salivary gland toxicity and pre-treatment CT and PET-based radiomics features are predictive of patient outcome in HNSCC.

Specific Aim 1: Determine DCE-MRI parameters associated with salivary gland dose response.

Specific Aim 1 Hypothesis: DCE-MRI parameters at pre- and mid-treatment time points are associated with normal tissue outcomes.

Project 1.1: Determine the uncertainty in MRI-to-MRI deformable image registration.

Project 1.2: Identify DCE-MRI parameters associated with dose response in salivary glands.

Specific Aim 2: Identify imaging protocols that minimize variability in CT-based radiomics features to improve patient outcome models.

Specific Aim 2 Hypothesis: Reducing the variability in CT-based radiomics features due to imaging protocols improves prediction accuracy when these features are used in HNSCC patient outcome models (overall survival, local-regional control, and freedom from distant metastases).

Project 2.1: Evaluate the variability in CT-based radiomics features due to tube voltage, artifacts in the head and neck region, and inter-scanner variability.

Project 2.2: Determine CT-based radiomics features predictive of patient outcome.

Specific Aim 3: Determine imaging protocols that minimize variability in PET-based radiomics features to improve patient outcome models.

Specific Aim 3 Hypothesis: Reducing the variability in PET-based radiomics features due to imaging protocols improves prediction accuracy when these features are used in HNSCC patient outcome models.

Project 3.1: Evaluate inter-scanner variability of PET-based radiomics features.

Project 3.2: Identify PET-based radiomics features predictive of patient outcome.

Chapter 3 : Magnetic Resonance Imaging Registration Uncertainty

This chapter is based upon:

Ger RB, Yang Y, Ding Y, Jacobsen MC, Fuller CD, Howell RM, Li H, Stafford RJ, Zhou S, Court LE. Accuracy of Deformable Image Registration on Magnetic Resonance Images in Digital and Physical Phantoms. *Medical Physics* doi: 10.1002/mp.12406. Volume 44, Issue 10, Pages 5153-5161. ©Wiley.

The permissions for reuse of these materials were obtained from Wiley.

3.1 Introduction

The use of MRI has increased because it allows for non-invasive evaluation of patients without ionizing radiation and provides superior soft tissue contrast in comparison with CT. With these attributes, MRI has great potential for use in longitudinal studies [9, 10, 51-53]. A longitudinal study is when data is acquired for the same subjects over a period of time and may have different designs based on the type of study, such as prospective or retrospective [54]. An example of a prospective longitudinal study is the recent DCE-MRI study evaluating mandible changes in patients that were scanned before chemoradiotherapy treatment started, 3-4 weeks after initiation of treatment, and 6-8 weeks after treatment concluded [55]. An example of a retrospective longitudinal study is the recent delta radiomics study on non-small cell lung cancer patient outcome predictions [14]. However, for MRI to be useful in the setting of longitudinal studies, accurate deformable image registration is needed.

Many commercial and in-house image registration systems have been benchmarked using CT [56-61]. For MRI, various in-house algorithms have been validated for different anatomic sites, including the liver [62, 63], prostate [64-66], and breast [67, 68]. However, image registration error on commercial software using MRI has not been widely reported. A b-spline-based commercial software, Velocity (Velocity AI version 3.0.1, Varian Medical Systems, Palo Alto, CA), has been evaluated for CT-based image registration and shown to have an average registration error below 5 mm [57, 69-73]. In

addition, our in-house demons-based algorithm has been validated for CT with registration error below 2 mm [58, 59, 74]. Both systems are used for MRI applications, but have not been validated for this use.

The current study aimed to evaluate two registration systems, Velocity commercial deformable image registration software and an in-house demons-based algorithm, for MRI. This was accomplished using synthetic images derived from patient longitudinal deformations and a phantom with implanted markers.

3.2 Methods and Materials

The deformable image registration uncertainty of the two registration systems was evaluated using two methods: synthetic images and a porcine phantom.

3.2.1 Porcine Phantom

Porcine meat was implanted with ten 0.35mm gold markers. These markers are currently the smallest commercially available gold markers and therefore do not appear in the MR images for the imaging protocol used. This allowed for accurate assessment of the imaging registration error. If the markers appeared on the image they could bias the registration at those points leading to inaccurate registration error estimates.

The porcine tissue was placed in a plastic container with movable dividers (United States Plastic Company, Lima, OH) to secure it in place. The porcine tissue was imaged using T1-weighted and T2-weighted MRI sequences where the markers were not visible and then imaged using CT where the markers were identifiable. The porcine tissue was then deformed by changing the placement of the dividers (as shown in Figure 3-1) and re-imaged. This process was repeated three times for a total of four sets of T1-weighted, T2-weighted, and CT images. The container is 27.6 cm x 21.0 cm x 12.7 cm with 5 notches in the short direction and 7 notches in the long direction. The notches are spaced 3.25 cm apart, and the first and last notches are 1.6 cm from the edge in the short direction and 1.8 cm in

the long direction. Long dividers are placed using the notches in the short direction and short dividers are placed using the notches in the long direction. The four different positions were as follows: (1) no dividers in the short direction, 1 divider in the first notch in the long direction; (2) 1 divider in the first notch in the short direction, 1 divider in the first notch in the long direction, (3) 1 divider in the first notch and 1 divider in the last notch in the short direction, 1 divider in the first notch in the long direction; and (4) 1 divider in the first notch and 1 divider in the last notch in the short direction, 1 divider in the first notch and 1 divider in the last notch in the long direction.

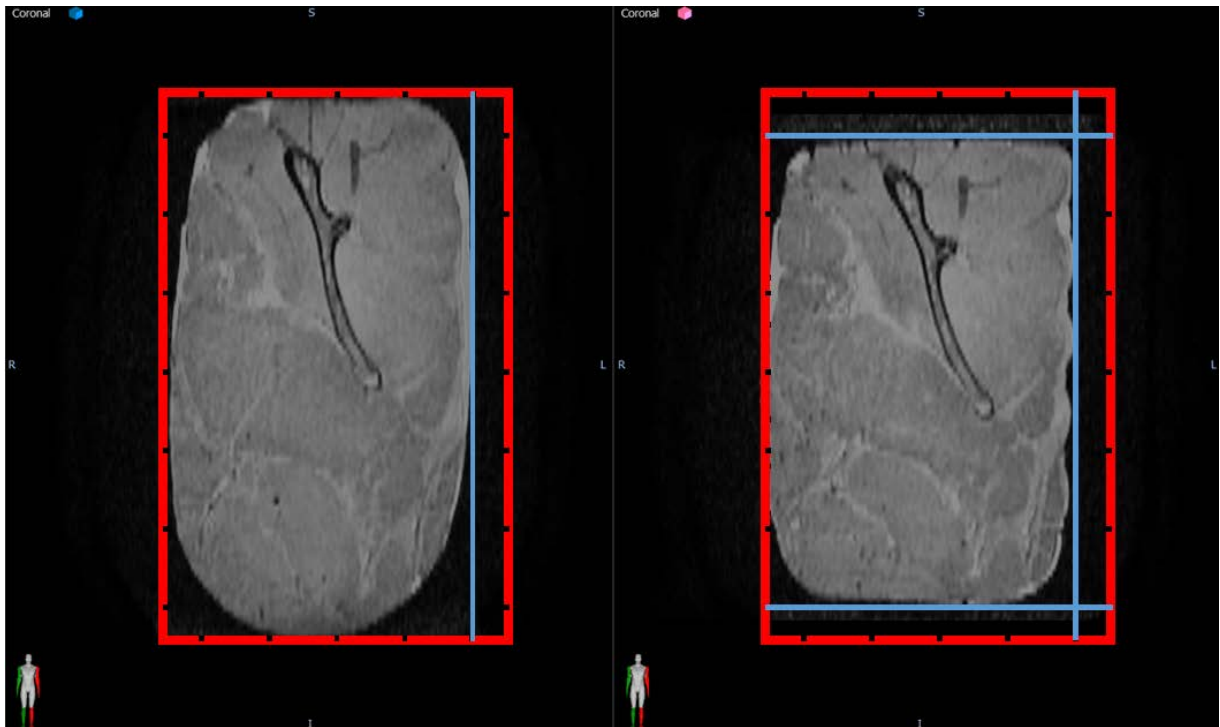


Figure 3-1: Porcine Phantom Representative Deformation.

A deformation was applied to the porcine phantom by moving the dividers. The red box represents the container with the grooves for the movable dividers and the position of the dividers is shown in blue. The original position is shown on the left and a deformed position using more movable dividers to secure the phantom in place is shown on the right.

The images were imported into the registration system where the MR images were rigidly registered to the CT image for each divider position and then deformably registered to the MR images at different divider positions. The gold markers were contoured for each divider position on the CT images and their center location was extracted. The markers were propagated through the registrations to determine the virtual location, and the error was measured by the distance between the virtual location and the known location. For example, the gold markers were transferred from CT-1 to MR-1 through rigid registration, then from MR-1 to MR-2 through deformable registration, then from MR-2 to CT-2 through rigid registration, and the error was measured as the distance between the propagated marker location on CT-2 and the actual marker location in CT-2. This method was applied to both registration systems.

3.2.2 Synthetic Images

The image registration methodology from Yu, *et al.* [75] was followed for 28 patients' with human papillomavirus-positive oropharyngeal squamous cell carcinoma who were treated with definitive chemoradiotherapy. The patients were selected from a prospective trial under a protocol approved by the institutional review board at MD Anderson Cancer Center with study-specific informed consent. The first 28 patients to complete the three MRI scans were included. Patients underwent MRI scans from December 2013 to October 2015. Patient median age was 57 (range 46-70), with 26 men and 2 women. The median left parotid volume was 30.9 cm³ (range 22.4-47.1 cm³), median right parotid volume was 33.5 cm³ (range 18.0-46.7 cm³), median left submandibular volume was 9.6 cm³ (range 5.7-19.7 cm³), median right submandibular volume was 9.4 cm³ (range 4.3-17.5 cm³), and median sublingual volume was 4.8 cm³ (range 1.6-9.4 cm³).

Patients underwent T1-weighted and T2-weighted MRI scans before treatment (within 1 week prior to treatment), during treatment (3-4 weeks after the start of treatment), and after treatment (6-8 weeks after completion of treatment). This methodology has been described previously [75]; briefly,

models were trained on the patient images using an in-house demons-based algorithm. An intra-patient variation model was created by deforming each patient's mid-treatment and post-treatment images to the pre-treatment image. One patient was selected to be the template for the synthetic images based on being the median age and having salivary glands with volumes near the median volume for all glands. An inter-patient variation model was created by deforming each patient's pre-treatment image to the selected patient's pre-treatment image. 95% of the variation was included in the intra- and inter-patient variation models. This was done separately for T1-weighted and T2-weighted images.

Synthetic pre-treatment images were created by deforming the selected patient's pre-treatment image using the inter-patient variation model. Synthetic post-treatment images were created by deforming the synthetic pre-treatment image using the intra-patient variation model. For each synthetic pre-treatment image, four synthetic post-treatment images were created. Four synthetic pre-treatment images were created, resulting in 16 image registrations for both T1- and T2-weighted images. This process is demonstrated in Figure 3-2. An example of the applied deformation is shown in Figure 3-3.

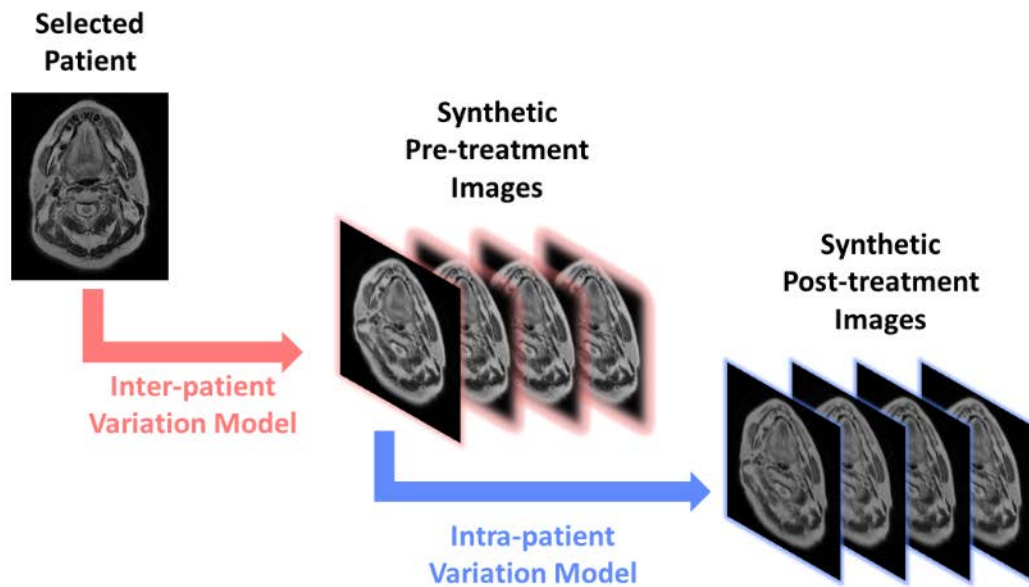


Figure 3-2: Workflow of Synthetic Image Generation.

The generation of synthetic images is represented visually. The selected patient's pre-treatment image is deformed by the inter-patient variation model to generate the synthetic pre-treatment images (highlighted in pink). For each of the synthetic pre-treatment images, four synthetic post-treatment images (highlighted in blue) are generated by deforming the synthetic pre-treatment image using the intra-patient variation model.

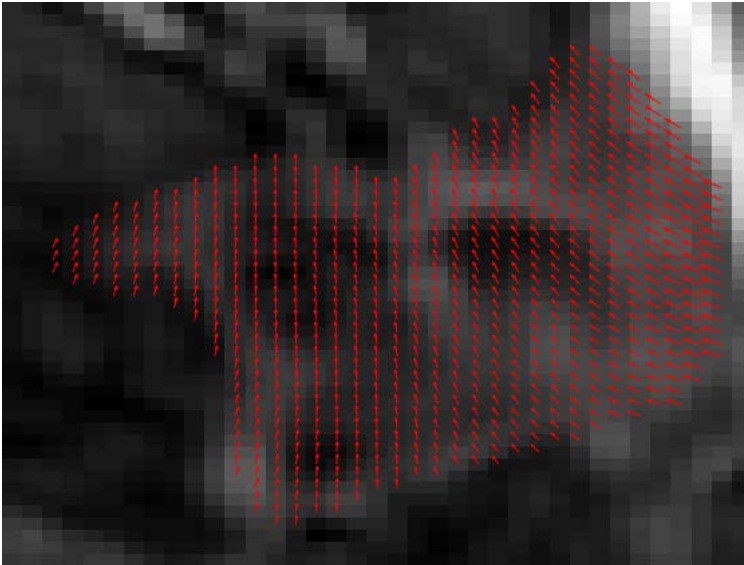


Figure 3-3: Parotid Deformation in Synthetic Image Generation.

An example of one of the deformations applied to the T1-weighted synthetic pre-treatment image left parotid to generate a synthetic post-treatment image. The arrows show the direction of the left-right and anteroposterior deformation and the size shows the magnitude of the deformation for that voxel.

The synthetic post-treatment images were registered to the synthetic pre-treatment images in the image registration system and the deformation vector field for each registration was exported. The difference between the registration system's deformation vector field and the applied deformation from the intra-patient variation model was evaluated within the salivary glands (left/right parotid, left/right submandibular, and sublingual). This method was not applied to the in-house demons-based algorithm because the algorithm was used to generate the inter- and intra-patient variation models used to generate the synthetic images for this method, and this would cause bias.

3.2.3 Imaging

3.2.3.1 Porcine MRI protocol

The T1-weighted MRI sequence was a three-dimensional spoiled gradient recalled echo sequence with flip angle 12° , repetition time 4.96 ms, echo time 2.1 ms, effective number of excitations 2, pixel bandwidth 325 Hz, field of view 25.6 cm, slice thickness 1 mm, and pixel size 1 mm \times 1 mm. The T2-weighted MRI sequence was a two-dimensional fast spin echo sequence with flip angle 90° , repetition time 5884 ms, echo time 98 ms, effective number of excitations 2, pixel bandwidth 325 Hz, field of view 25.6 cm, slice thickness 2.5 mm, gap 1.5 mm, and acquisition pixel size 1 mm \times 1 mm, and zero filling interpolation \times 2. The MRI sequences were executed on the same 3.0T GE machine as the patient scans using an 8US TORSOPA coil (GE Healthcare, Waukesha, WI). The CT images were acquired on a GE Discovery CT 750HD (GE Healthcare) in helical mode at 120 kVp, 150 mA, 0.516 pitch, 0.8-s rotation time, 0.625-mm slice thickness, 0.625-mm spacing between slices, pixel size 0.39 mm \times 0.39 mm, and CTDIvol 41.07 mGy.

3.2.3.2 Patient MRI Protocol

MRI scans were performed using a 3.0T Discovery 750 MRI scanner (GE Healthcare) with 6-element flex coils and a flat insert table (GE Healthcare). Thirty slices with an axial field of view of

25.6 cm and slice direction in the superior-inferior direction (slice thickness of 4 mm) were selected to cover the spatial region encompassing the palatine process region cranially to the cricoid cartilage caudally for all scans. The T1-weighted MRI sequence was a three-dimensional spoiled gradient recalled echo sequence with flip angle 15°, repetition time 3.6 ms, echo time 1 ms, effective number of excitations 0.7, pixel bandwidth 325 Hz, acquisition pixel size 2 mm × 2mm (matrix size 128 x 128), and zero filling interpolation × 2. This resulted in reconstructed voxel sizes of 1 mm x 1 mm x 4 mm. The T2-weighted MRI sequence was a two-dimensional fast spin echo multi-slice sequence with flip angle 90°, repetition time ~3600 ms, echo time ~100 ms, effective number of excitations 1, pixel bandwidth 195 Hz, slice thickness 2.5 mm, gap 1.5 mm, acquisition pixel size 1 mm × 1 mm, and zero filling interpolation × 2.

3.2.4 Registration Techniques

3.2.4.1 Velocity

The images were first registered using manual alignment by shifting and rotating the secondary image. Then a region of interest that encompassed the whole porcine phantom or patient anatomy was drawn. Within this region of interest, the images were aligned first using Rigid 3 Passes. The Velocity rigid registration uses mutual information to align anatomy. Then MR Corrected Deformable was used to deformably align the images. Velocity's deformable image registration uses a cubic B-spline algorithm with a uniform knot vector and a steepest gradient descent optimizer. Mattes Mutual Information is used as the "good of match" driver for the registration. The MR correction applies a fade correction to the image to correct for shading artifacts typically caused by inhomogeneities in the magnetic field and then proceeds with the deformable image registration.

3.2.4.1 In-house Demons-Based Algorithm

The in-house deformable image registration is a dual-force Demons algorithm [59]. Before performing the deformable registration, we performed histogram equalization to match the contrast of the 2 images. The histogram equalization was performed locally by separating the images into small blocks. A multiresolution scheme was used to accelerate the registration and improve the robustness of the registration. The parameter settings for the deformable registration are specified in Table 3-1. These parameters were chosen based on our experience in optimizing the algorithm for the MR-MR registration. Refer to Wang *et al.* [59] for details about this deformable registration algorithm.

Table 3-1. Parameter Settings for the Dual-force Demons Deformable Image Registration

Parameter	Value
Number of bins for histogram equalization	256
Block size for histogram equalization	20
Multi-resolution levels	6
Number of iterations	200
Upper bound of step size	1.25
Gaussian variance for regularization	1.5

3.2.5 Statistical Methods

The applied deformation for the synthetic images was obtained from the deformation vector field used to generate the synthetic post-treatment image from the synthetic pre-treatment image. The applied deformation for the porcine phantom was calculated from the difference in marker location on the CT images. The applied deformation was summarized using the root mean square (RMSD) and maximum. The registration error was calculated as described above and was also summarized using the root mean square (RMSE) and maximum.

3.3 Results

3.3.1 Porcine Phantom

The four image sets were registered producing six different pairs of images. In our analysis using Velocity one of the gold markers was not mapped to a voxel. For the T1-weighted MR image registrations using the in-house demons-based algorithm, one of the gold markers was mapped to the registered image for only one of the registrations. For the T2-weighted MR image registrations using the in-house demons-based algorithm, one of the gold markers was not mapped to the registered image in two of the registrations. The markers in our study are represented by regions of interest (ROIs). Some markers have very small volume. The deformable mapping of these small ROIs could not produce a reasonable volume so that the software treated the deformed ROIs as noise and removed them. To avoid the confusion, we take out these small ROIs from our results.

For both the T1- and T2-weighted MR images, the RMSD was 5.0 mm in the left-right (LR) direction, 9.0 mm in the anteroposterior (AP) direction, and 6.1 mm in the superior-inferior (SI) direction. The T1-weighted MR images registered using Velocity had RMSEs of 1.8 mm in the LR direction, 1.5 mm in the AP direction, and 2.7 mm in the SI direction. The T1-weighted MR images registered using the in-house demons-based algorithm had RMSEs of 1.2 mm in the LR direction, 1.5 mm in the AP direction, and 2.1 mm in the SI direction. The T2-weighted MR images registered using Velocity had RMSEs of 1.3 mm in

the LR direction, 1.2 mm in the AP direction, and 1.6 mm in the SI direction. The T2-weighted MR images registered using the in-house demons-based algorithm had RMSEs of 0.81 mm in the LR direction, 1.1 mm in the AP direction, and 1.1 mm in the SI direction. Boxplots of the RMSE and RMSD in the LR, AP, and SI directions from Velocity are shown in Figure 3-4. The maximum registration errors from both registration systems are summarized in Table 3-2.

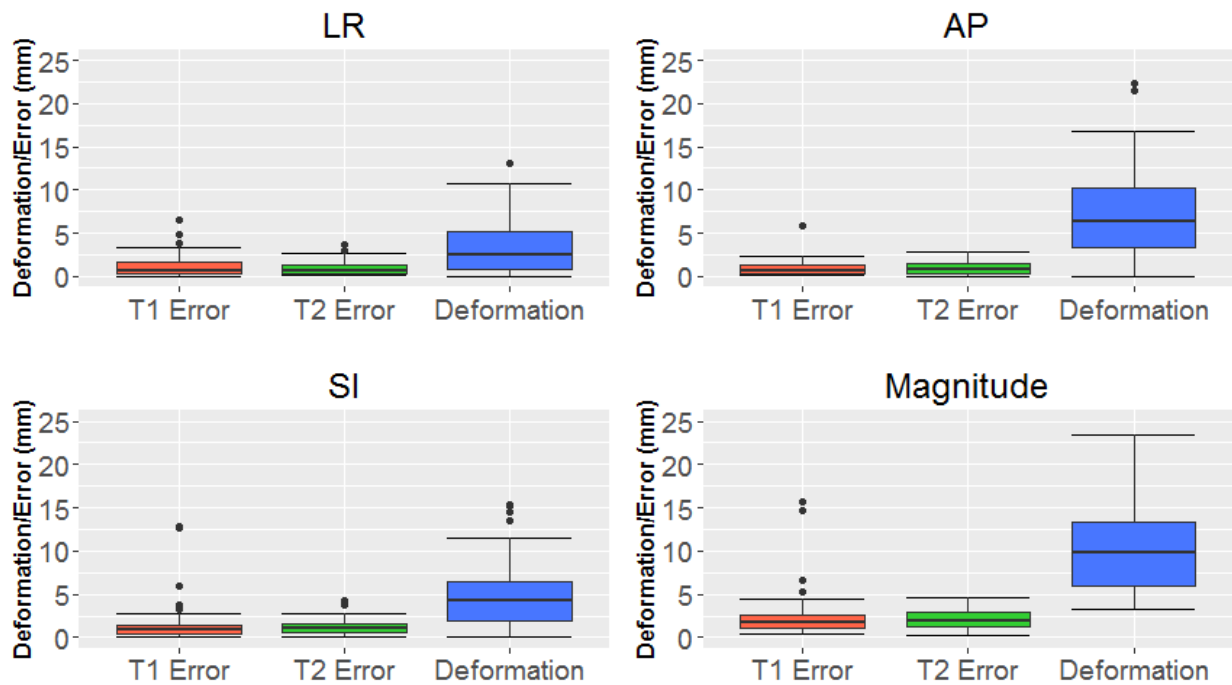


Figure 3-4: Porcine Phantom Registration Error.

The registration error using the T1-weighted MR images is shown in red, the registration error using the T2-weighted MR images is shown in green, and the applied deformation is shown in blue. Values are shown for the left-right (LR) direction, anteroposterior (AP) direction, superior-inferior (SI) direction, and magnitude.

3.3.2 Synthetic Images

For the T1-weighted MR images, the RMSD was 1.5 mm in the left-right (LR) direction, 2.1 mm in the anteroposterior (AP) direction, and 0.79 mm in the superior-inferior (SI) direction. The RMSE was 0.76 mm in the LR direction, 0.76 mm in the AP direction, and 0.69 mm in the SI direction. RMSD and RMSE for each gland can be found in Figure 3-5. The maximum registration error was 1.1-5.7 mm in the LR direction, 1.3-3.8 mm in the AP direction, and 1.7-10 mm in the SI direction for the five salivary glands. The applied deformation is larger for the left parotid than the right parotid for these synthetic images. This was a result of the patients included in this study. A larger sample size would likely not have seen this effect.

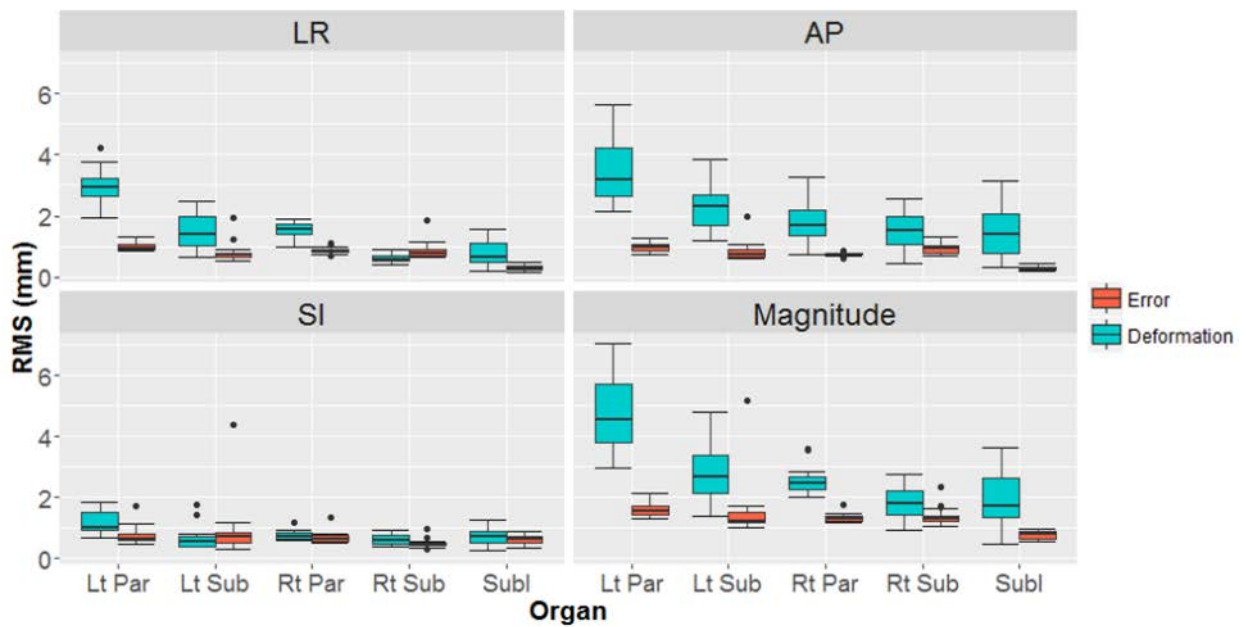


Figure 3-5: Root Mean Square (RMS) Applied Deformation and Registration Error in T1-Weighted Synthetic Images.

RMS registration error is shown in red and RMS applied deformation is shown in blue for the left-right (LR) direction, anteroposterior (AP) direction, superior-inferior (SI) direction, and magnitude (bottom right). Each boxplot shows RMS registration error and applied deformation for the left parotid (Lt Par), right parotid (Rt Par), left submandibular (Lt Sub), right submandibular (Rt Sub), and sublingual (Subl) glands. In each plot, the RMS registration error is shown to the right of the RMS applied deformation for each gland.

For the T2-weighted MR images, the RMSD was 1.1 mm in the LR direction, 3.4 mm in the AP direction, and 1.4 mm in the SI direction. The RMSE was 1.1 mm, 0.75 mm, and 0.81 mm in the LR, AP, and SI directions, respectively. RMSD and RMSE for each gland can be found in Figure 3-6. The maximum registration error was 1.9-6.0 mm in the LR direction, 1.5-4.3 mm in the AP direction, and 2.0-3.4 mm in the SI direction for the five salivary glands. The maximum registration error for each salivary gland is shown in Table 3-2.

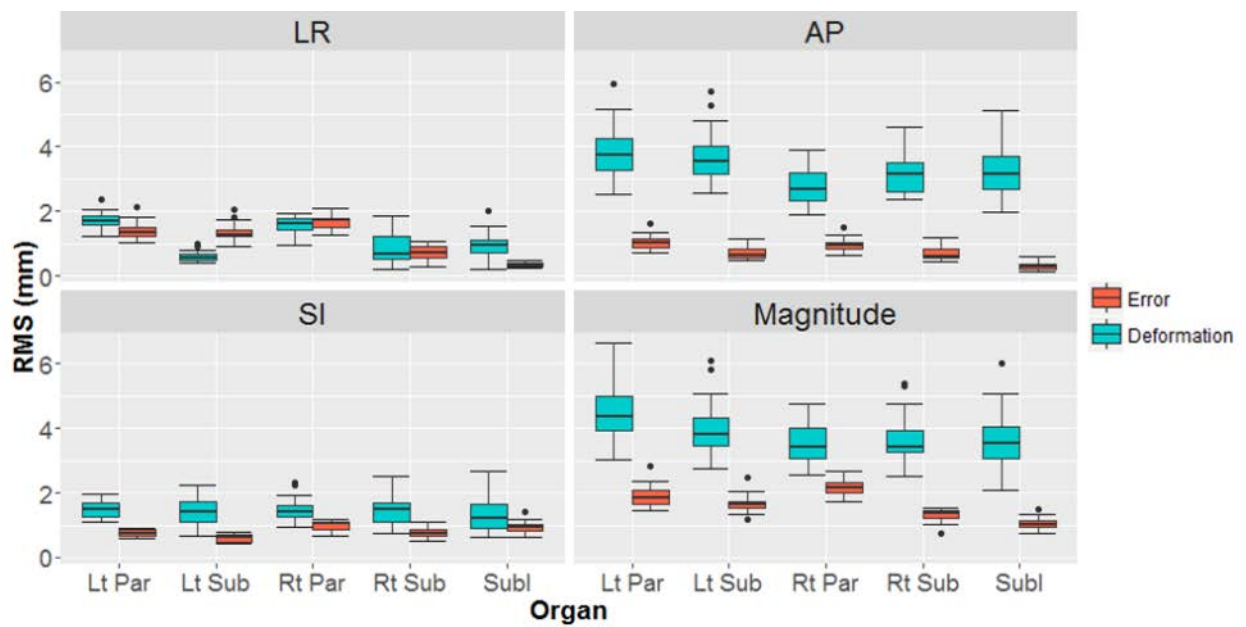


Figure 3-6: Root Mean Square (RMS) Applied Deformation and Registration Error in T2-Weighted Synthetic Images.

RMS registration error is shown in red and RMS applied deformation is shown in blue for the left-right (LR) direction, anteroposterior (AP) direction, superior-inferior (SI) direction, and magnitude. Each boxplot shows RMS registration error and RMS applied deformation for the left parotid (Lt Par), right parotid (Rt Par), left submandibular (Lt Sub), right submandibular (Rt Sub), and sublingual (Subl) glands. In each plot, the RMS registration error is shown to the right of the RMS applied deformation for each gland.

Table 3-2. Maximum Registration Error from Synthetic Images and the Porcine Phantom

Source	T1-weighted images (mm)			T2-weighted images (mm)		
	LR	AP	SI	LR	AP	SI
Left parotid gland	3.37	2.65	4.62	4.51	3.25	2.60
Right parotid gland	2.77	1.99	4.38	5.97	4.27	3.40
Left submandibular gland	4.90	3.80	10.2	3.66	1.95	2.04
Right submandibular gland	5.72	1.94	2.44	2.31	1.78	2.37
Sublingual gland	1.09	1.34	1.71	1.87	1.52	2.84
Porcine phantom, Velocity	6.5	5.9	12.9	3.7	2.9	4.3
Porcine phantom,	2.9	5.1	10.8	2.4	3.8	2.4
in-house algorithm						

LR, left-right; AP, anteroposterior; SI, superior-inferior.

3.4 Discussion

MRI use in the United States has increased more than three-fold over the past 20 years according to the Organisation for Economic Co-Operation and Development database [76]. This increased use includes longitudinal studies that require deformable image registration. The current study evaluated the performance of two image registration systems for MRI deformable image registration. The porcine phantom validated both image registration systems, Velocity and the in-house demons-based algorithm. Then in order to validate Velocity further, the in-house demons-based algorithm was used to generate synthetic images. The use of synthetic images relies on the accuracy of the in-house demons-based algorithm to create the inter- and intra-patient variation models. This system has been previously validated using CT [58, 59, 74]. The porcine phantom results demonstrated that this validation is also applicable to MRI. These evaluation measures, the porcine phantom and synthetic images, showed that both Velocity and the in-house demons-based image registration system performed well, with all RMSEs below 3 mm.

The RMSE was relatively stable, as shown by the increase in the applied deformation in the AP direction compared with the LR and SI direction for the synthetic images, even though the registration error stayed around the same values for the LR, AP, and SI directions. Furthermore, when applied deformations were pushed past physiological bounds in the porcine phantom, the registration errors were similar to the registration errors from the synthetic images.

The generally low RMSEs for both image registration systems are consistent with average registration errors reported in the literature when evaluated using CT [57-59, 69-74]. However, we did find occurrences of registration errors greater than 10 mm in both image registration systems. The maximum registration errors of 10.2 mm for the SI direction of the left submandibular gland on the T1-weighted synthetic images and 12.9 mm for the SI direction of the porcine phantom T1-weighted MR images in Velocity were higher than the maximum registration errors measured in CT-based registration with Velocity [57, 69-73]. These large registration errors may be due to the type of

surrounding tissue. Both of these points were in the vicinity of bone but at least 3 cm away. Singhrao et al used a head and neck phantom that included representative bony anatomy, and they found maximum errors between 6.5 and 8.3 mm for Velocity's deformable image registration with CT images [70]. This maximum error is closer to the error we found in the current study than are the maximum errors found in studies of anatomical regions outside the head and neck images [71, 72].

The porcine phantom results showed that the maximum registration errors for both the T1-weighted and T2-weighted MR images were lower than those observed for CT using the in-house demons-based algorithm [58]. This system has been previously validated and subsequently used for a variety of CT studies [77-79]. Results from the current study presented here imply that the in-house demons-based algorithm can be used reliably for MRI as well as CT.

However, there is a limitation with the use of a porcine phantom in this setup because we were only able to deform and not shrink the phantom. The synthetic images included shrinkage as part of the intra-patient variation model. The lack of shrinking in the porcine model limits its applicability in regions of interest that shrink or grow over time. Nevertheless, the current study expands on other studies that have evaluated deformable image registration using only a few markers [64, 65]. Our study design is similar to that used by Lian et al. [66] who used 10-15 markers, but the phantom in the current study was meat rather than tissue-equivalent bolus material, so that it included muscle, bone, and fat. Another limitation of this porcine phantom is the size of deformations that were applied. The deformations applied are directly linked to the spacing of the notches for the movable dividers in the container which was larger than the typical deformation seen in the synthetic images. Therefore, these deformations provide an extreme scenario. The low RMSE by both registration systems in this extreme case shows the good performance of these two systems for MRI deformable image registration.

Both studies utilized consistent imaging parameters and did not investigate the influence of acquisition parameters, such as coil placement, on the MRI deformable image registration. Therefore, these results demonstrate a controlled study which may not fully represent what is seen within clinics.

Additionally, image noise and artifacts can impact deformable image registration accuracy- high image noise would degrade accuracy as would the presence of artifacts- and their impacts were not evaluated in this study. However, it is not unreasonable to assume that these error estimates are applicable when imaging parameters are controlled, such as using a repeatable setup in a thermoplastic mask. Setup in a head and shoulder thermoplastic mask with dental stent and coils centered on the base of tongue region showed significantly improved image quality and reproducibility [80] and has consequently been used in a DCE-MRI longitudinal study for radiotherapy-induced mandibular changes [55].

The synthetic images experiment produced SI RMSDs that were typically less than the slice thickness, 4 mm. This is due to the patient population that was used to create the inter- and intra-patient variation models. Therefore, extending the results to different patient populations that have larger SI displacement must be done with caution. The porcine phantom results have larger SI displacement and demonstrated similar registration error. However, these larger SI displacements should be verified with patient data, such as synthetic images derived from patients with larger displacements, for full confidence in the registration error in different patient populations.

Additionally, the synthetic images experiment was limited to the salivary glands as these are often analyzed in normal tissue studies. However, the porcine phantom experiment is not site limited. While it was used in this study to support the synthetic image data, the results from this experiment are applicable to other body sites. Therefore, with similar synthetic image experiments, the results can be applied to other locations within the body.

3.5 Conclusions

Both Velocity and the in-house demons-based algorithm demonstrated low registration errors, with all RMSEs below 3 mm for RMSDs between 0.79 and 9.0 mm, indicating that these deformable image registration systems can be used for MRI longitudinal studies.

Chapter 4 : Dynamic Contrast-Enhanced Magnetic Resonance Imaging Feature

Stability

This chapter is based upon:

Joint Head and Neck Radiotherapy-MRI Development Cooperative. A Multi-Institutional Comparison of Dynamic Contrast-Enhanced Magnetic Resonance Imaging Parameter Calculations. *Scientific Reports* doi: 10.1038/s41598-017-11554-w. Volume 7, Article 11185. 2017. ©Nature Publishing Group.

This article is under a Creative Commons license (<http://creativecommons.org/licenses/by/4.0/>) which permits reproduction in any format.

4.1 Introduction

DCE-MRI is a noninvasive tool for examination of the microvasculature of tumors and normal tissue. Investigators have used DCE-MRI to assess therapeutic response of HNSCC and have shown associations between DCE-MRI metrics and changes in salivary glands and mandible [6-11, 55]. To the best of our knowledge, its use as a prognostic tool to inform treatment decisions for HNSCC has yet to be investigated in a large multisite prospective trial. Before such trials can begin, DCE-MRI inter-algorithm comparisons must be conducted to ensure consistency of output parameter maps for collating data during the multi-institution trial. Two quantitative metrics for DCE-MRI are the transfer constant for contrast agent transport from the blood plasma into the extravascular extracellular space (K^{trans}) and the volume fraction of the extravascular extracellular space (v_e). The calculation of these quantitative metrics can be impacted by the acquisition parameters. The accuracy and precision of these quantitative metrics can be influenced by arterial input function (AIF) quantification, temporal resolution in data acquisition, signal-to-noise ratio (SNR), and pharmacokinetic model selection [81-91]. For example, uncertainties in T1 map values and applied flip angle have been reported to cause errors of 88% in K^{trans} and 73% in v_e , while reduced temporal resolution by 7-fold have reported decreases in

K^{trans} up to 48% [88]. Therefore, acquisition parameters must be thoroughly tested and uniform across patients as they can dramatically impact measured DCE-MRI parameters.

The Tofts-Kermode pharmacokinetic model [92] is the most commonly used model for DCE-MRI analysis, but implementation of each algorithm differs in facets such as data preprocessing, approaches to numerical optimizations in kinetic analysis, and data postprocessing, which may impact the values of the output quantitative metrics. Several recent studies demonstrated significant inter-algorithm variability when evaluating DCE-MRI of the female pelvis, breast, and rectum [93-95]. Of these studies, the one by Huang et al. [94] demonstrated systematic differences in output parameter values between algorithms, which meant that results from different algorithms could be used together if correction factors were applied; the other studies, however, did not demonstrate any systematic errors. In addition, Cron et al. [96] found that the percentage of nonphysical values (e.g. v_e values greater than 1) in the quantitative metrics increased as noise increased when they tested three software packages. This noise dependence and inter-algorithm variance in quantitative DCE-MRI metrics are large obstacles to the clinical implementation of DCE-MRI and must be thoroughly investigated before proceeding with large multisite clinical trials using DCE-MRI in HNSCC patients.

In this study, we investigated the variability in K^{trans} and v_e across algorithms that are based on the Tofts-Kermode and extended Tofts pharmacokinetic models [97, 98]. For this purpose, we used digital reference objects (DROs) from the Radiological Society of North America Quantitative Imaging Biomarkers Alliance and DCE-MRI data from oropharyngeal squamous cell carcinoma patients who underwent multiple DCE-MRI scans during treatment with definitive chemoradiotherapy.

4.2 Methods and Materials

4.2.1 Algorithms

Eleven algorithms from six institutions and one commercial software package were analyzed. They consisted of seven Tofts-Kermode models (identified herein as algorithms 2, 3, 5, 6, 8, 10, 11) and

four extended Tofts models (algorithms 1, 4, 7, 9). Spatial averaging on the DCE-MRI images was used in algorithms 5, 6, 7, 8, and 9. All algorithms are currently used for research applications at the respective institutions. The algorithms are described in Table 4-1.

Table 4-1. Description of Algorithms

Institution	Model(s) Used
Massachusetts General Hospital	Tofts-Kermode (description in Appendix A)
MD Anderson Cancer Center	Tofts-Kermode and Extended Tofts (description in Appendix A)
Netherlands Cancer Institute	Tofts-Kermode and Extended Tofts [99]
nordicICE	Extended Tofts [100]
Oregon Health & Science University	Tofts-Kermode and Tofts-Kermode [83, 101, 102]
Princess Margaret Cancer Center	Tofts-Kermode and Extended Tofts [103, 104]
University of Texas at Austin	Tofts-Kermode [105, 106]

Algorithms are listed in alphabetical order not order displayed in figures.

4.2.2 DROs

DROs provided by the Radiological Society of North America Quantitative Imaging Biomarkers Alliance were used to assess algorithm performance. The DROs had six K^{trans} values ranging from 0.01 min^{-1} to 0.35 min^{-1} that were constant across the rows and five v_e values ranging from 0.01 to 0.5 that were constant down the columns, resulting in 30 different K^{trans} - v_e pairs, each encompassing 10 x 10 pixels. The K^{trans} and v_e values were used to generate synthetic image data using the Tofts-Kermode two-parameter model run in JSim, an open-source modeling system [92, 107]. One DRO without noise [108] and 28 DROs with noise (SNR 0.18-1.8) [109] simulated by varying the sampling interval, timing offset, S_0 , and sigma value were used to evaluate algorithm performance. For each K^{trans} - v_e pair, the output pixels from the algorithms were subjected to a threshold to non-physiologic pixels ($0 < K^{trans}$ output < 5 and $0 < v_e$ output < 1) and then averaged.

4.2.3 Patients

Fifteen patients diagnosed with human papillomavirus-positive oropharyngeal squamous cell carcinoma were included in this study under a protocol approved by the institutional review board at MD Anderson Cancer Center. All patients gave their study-specific informed consent. All methods were performed in accordance with the relevant guidelines and regulations. Patients underwent DCE-MRI scans from December 2013 to October 2014. The criteria for study inclusion were an age older than 18 years, histologically documented stage III or IV human papillomavirus-positive oropharyngeal squamous cell carcinoma according to the American Joint Committee on Cancer 7th edition staging criteria, eligibility for definitive chemoradiotherapy, and an Eastern Cooperative Oncology Group performance status of 0 to 2. Patients were excluded for any of the following reasons: definitive resection of a primary tumor, administration of induction chemotherapy before radiotherapy, a prior cancer diagnosis except that of appropriately treated localized epithelial skin cancer or cervical cancer, prior

radiotherapy to the head and neck, contraindications for gadolinium-based contrast agents, and claustrophobia.

Patient median age was 56 years (range, 47-68), with 14 men and 1 woman. All patients received radiotherapy at 70 Gy in 33 fractions. The majority of the patients (87%) received cisplatin-based chemotherapy concurrently with radiotherapy. Patient, disease, and treatment characteristics are listed in Table 4-2. Patient 12 did not have a primary tumor because he underwent bilateral tonsillectomy before scanning.

Table 4-2. Study Patient Demographics

Patient Number	Sex	Age (years)	Race/ Ethnicity	Smoking Status	Primary Tumor Site	TNM Category	Chemotherapy (weekly)
1	M	52	White	N	Base of tongue	T3N1M0	Cisplatin
2	M	53	White	Y	Base of tongue	T2N2aM0	Cetuximab
3	M	60	White	Y	Tonsil	T4N2bM0	Cisplatin
4	M	55	White	Y	Tonsil	T3N2bM0	Cisplatin
5	M	65	White	N	Base of tongue	T2N1M0	Cetuximab
6	M	57	Hispanic	Y	Tonsil	T2N2cM0	Cisplatin
7	M	60	White	Y	Base of tongue	T2N2bM0	Cisplatin
8	M	58	Black	Y	Base of tongue	T2N2cM0	Cisplatin
9	M	62	Asian	Y	Tonsil	T4N2cM0	Cisplatin
10	F	48	White	Y	Tonsil	T4N2bM0	Cisplatin
11	M	56	White	N	Tonsil	T2N2cM0	Cisplatin
12	M	68	White	Y	Tonsil	TxN2cM0	Cisplatin
13	M	47	White	N	Tonsil	T3N2bM0	Cisplatin
14	M	47	White	Y	Tonsil	T3N2bM0	Cisplatin
15	M	55	White	N	Base of tongue	T4N2bM0	Cisplatin

All patients underwent DCE-MRI scans within 1 week prior to treatment, 3-4 weeks after the start of treatment, and 6-8 weeks after the completion of treatment. The DCE-MRI scans were done using a 3.0T Discovery 750 MRI scanner (GE Healthcare) with six-element flex coils and a flat insert table (GE Healthcare). The same immobilization devices (individualized head and shoulder mask, customized head support, and intraoral tongue-immobilizing/swallow-suppressing dental stent) were employed in longitudinal scans to improve image co-registration and to reduce interval physiologic motion (e.g., swallowing).

Thirty axial slices with a field of view of 25.6 cm and thickness of 4 mm were selected to cover the spatial region encompassing the palatine process region cranially to the cricoid cartilage caudally for all scans. Prior to DCE-MRI, T1 mapping was performed using a total of six variable-flip-angle three-dimensional spoiled gradient recalled echo sequences (flip angles: 2°, 5°, 10°, 15°, 20°, and 25°; repetition time/echo time, 5.5/2.1 ms; number of effective excitations, 0.7; spatial resolution, 2 mm × 2 mm × 4 mm; scan time, 3 minutes). The DCE-MRI acquisition consisted of a three-dimensional fast spoiled gradient recalled echo sequence to gain sufficient SNR, contrast, and temporal resolution. The following scan parameters were used: flip angle, 15°; repetition time/echo time, 3.6/1.0 ms; number of effective excitations, 0.7; spatial resolution, 2 mm × 2 mm × 4 mm; temporal resolution, 5.5 s; number of temporal frames, 56; pixel bandwidth, 326 Hz; acceleration factor, 2; and scan time, 6 minutes. Gadopentetate dimeglumine (Magnevist; Bayer HealthCare Pharmaceuticals, Berlin, Germany) was administered intravenously to the patients at the end of the sixth frame (dose, 0.1 mmol/kg at a rate of 3 mL/second) followed by a 20-mL saline flush via a power injector (Spectris MR Injector; Medrad, Warrendale, PA) at a rate of 3 mL/second.

Variable-flip-angle images, DCE-MRI images, and a bootstrapped population AIF measured in a region of interest in the carotid artery [55] were distributed to each institution to use in their algorithm(s) to generate K^{trans} and v_e parameter maps for each patient.

Each patient had 6 ROIs—contralateral and ipsilateral parotid glands, contralateral and ipsilateral submandibular glands, sublingual glands, and a primary gross tumor volume (GTV-P)—contoured on his or her pretreatment images by a radiation oncologist with 7 years of experience (A.S.R. Mohamed). Midtreatment and posttreatment images were deformably registered to the pretreatment images using a commercially available software program (Velocity AI, version 3.0.1; Varian Medical Systems, Palo Alto, CA). The deformation vector fields were exported from the deformation software and used with an in-house MATLAB code (MATLAB 2014b; MathWorks, Natick, MA) to deform the ROIs and extract K^{trans} and v_e values from the six ROIs on each parameter map at the three time points. For each ROI, K^{trans} and v_e values were subjected to the same threshold constraints as in the DROs and then averaged.

4.2.4 Statistical Methods

A stratified permutation test was designed to determine whether the K^{trans} and v_e values from an algorithm for a specific DRO were generally ordered correctly in the DRO. Permutation tests work by rearranging data in many possible ways in order to estimate the sampling distribution for the test statistic. Algorithms were compared on a pairwise basis using a paired Wilcoxon rank-sum test to determine if the outputs of two algorithms were statistically different (R software package, version 3.3.1). Algorithms were split into two groups based on if spatial averaging was used on the DCE-MRI scans. The two groups were compared using a one-sided student's t-test to determine if lower error on the DROs was calculated when spatial averaging was used. All p-values were adjusted using the Benjamini-Hochberg correction for multiple comparisons.

For patient DCE-MRI data, consistency of trends across algorithms was assessed using linear mixed effects models (R lme4 package, version 1.1.12) constructed for the differences between the pretreatment and midtreatment, pretreatment and posttreatment, and midtreatment and posttreatment quantitative metrics, and percent change in these three time differences. Two mixed

effects models were created: one in which the algorithm was a fixed effect and the ROI was a random effect ($\Delta \sim \text{algorithm} + (1|\text{ROI})$) and one in which only the random effect of the ROI was included ($\Delta \sim 1 + (1|\text{ROI})$). A likelihood ratio test was performed for these two models to determine if the algorithm was a significant factor in the measured changes. All p-values were adjusted using the Benjamini-Hochberg correction for multiple comparisons (R, version 3.3.1). We used linear mixed effects models with likelihood ratio tests instead of ANOVA tests because in most comparisons we observed statistically different variances as determined using the Levene test, which violates one of the assumptions of ANOVA tests. Intraclass correlation coefficient is more appropriate for complete data sets⁴⁶, so it was not applicable for this data set.

For all ROIs, patients were categorized as above or below the median values from a given algorithm using three different metrics: (1) each time point, (2) difference between time points, and (3) percent difference between time points. Krippendorff's alpha was used to assess inter-algorithm reliability (R, irr package, version 0.84). We used Krippendorff's alpha to compare algorithms because of its ability to handle missing data, which occurred because for some algorithms, all K^{trans} and v_e values were outside the threshold for a given patient's ROI.

Trends within each algorithm were assessed using Spearman's rank correlation coefficient (R, version 3.3.1). Spearman correlations were conducted using three different sets of time points: (1) all three time points, (2) only the pretreatment and midtreatment time points, and (3) only the pretreatment and posttreatment time points were evaluated. All p-values were adjusted using the Benjamini-Hochberg correction for multiple comparisons. For all statistical tests, p-values below 0.05 after adjustment were considered significant.

4.3 Results

4.3.1 DROs

One of the Tofts-Kermode algorithms (algorithm 11) could not process the DROs because of the algorithm's structure. Therefore, the remaining 10 algorithms were used for DRO analysis. For the noiseless DRO, the stratified permutation test demonstrated that both K^{trans} and v_e were statistically significantly ordered correctly ($p < 0.05$) for all of the algorithms. Eighty-two percent of pairwise algorithm comparisons were statistically significantly different ($p < 0.05$) regarding K^{trans} , and 69% of the comparisons were statistically significantly different ($p < 0.05$) regarding v_e based on the Wilcoxon rank-sum test. Figure 4-1 shows the algorithm performance for the noiseless DRO. Most of the K^{trans} and v_e measured values in the noiseless DRO were close to the true simulated values: 96% of K^{trans} and 96% of v_e measured values were within 10% of the simulated values. More spread in the measured values was observed at higher simulated values of K^{trans} or v_e . Heat maps of the percentage error of K^{trans} and v_e measured values in comparison to the simulated values are shown in the appendix (Appendix A Figure A-1).

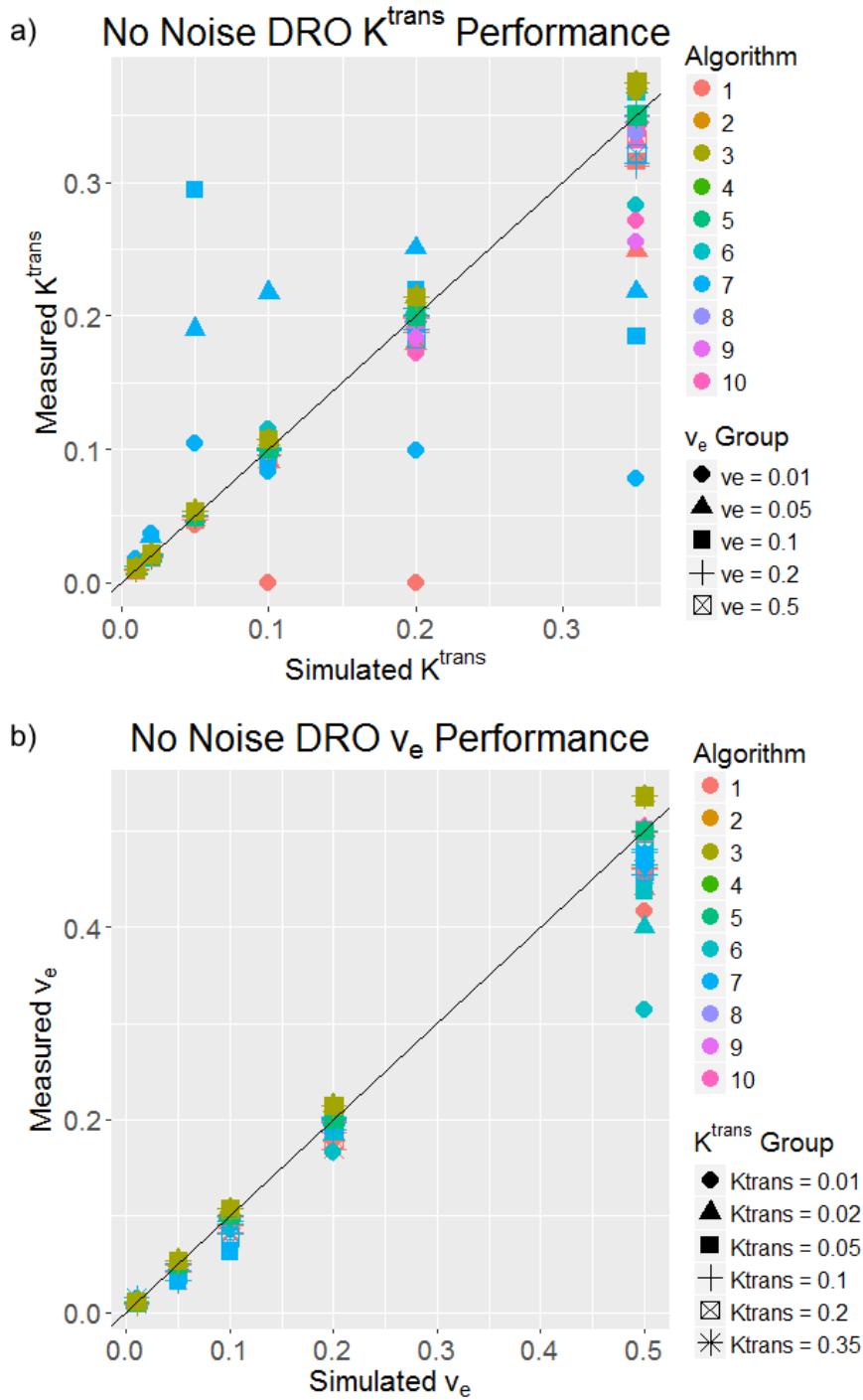


Figure 4-1: No Noise DRO Performance.

Plots of algorithm performance in a DRO with no noise for (a) K^{trans} and (b) v_e . The simulated values are on the x-axis, and the measured values from each algorithm are on the y-axis. The 45° line represents 100% accuracy of the measured values. Each color represents a different algorithm, and each shape represents a different v_e column in (a) and a different K^{trans} row in (b).

The stratified permutation test for the 28 DROs with noise demonstrated that in 86% and 84% of the cases (algorithm-DRO combinations), K^{trans} and v_e were statistically ordered correctly ($p < 0.05$) when one of the algorithms was excluded because of missing K^{trans} values and failure of the v_e test for all 28 of these DROs. Most of the test failures occurred at the lowest SNR (0.18). Eighty-four percent of the K^{trans} pairwise comparisons and 81% of the v_e pairwise comparisons were statistically significantly different ($p < 0.05$) based on the Wilcoxon rank-sum test results.

Heat maps of the percent error in K^{trans} and v_e relative to the simulated values in the 28 DROs with noise are shown in Figure 4-2. The maximum percent error in this figure was set to 100% and the minimum percent error was set to -100%. Therefore, any K^{trans} and v_e values greater than the maximum percent error are mapped to red. The only trend found was less error at higher K^{trans} and v_e simulated values although there is more spread in the measured values at these higher K^{trans} and v_e simulated values. Algorithms that used spatial averaging were found to have statistically significantly less ($p < 0.05$) K^{trans} and v_e calculated error than algorithms that did not have spatial averaging according to the student's t-tests.

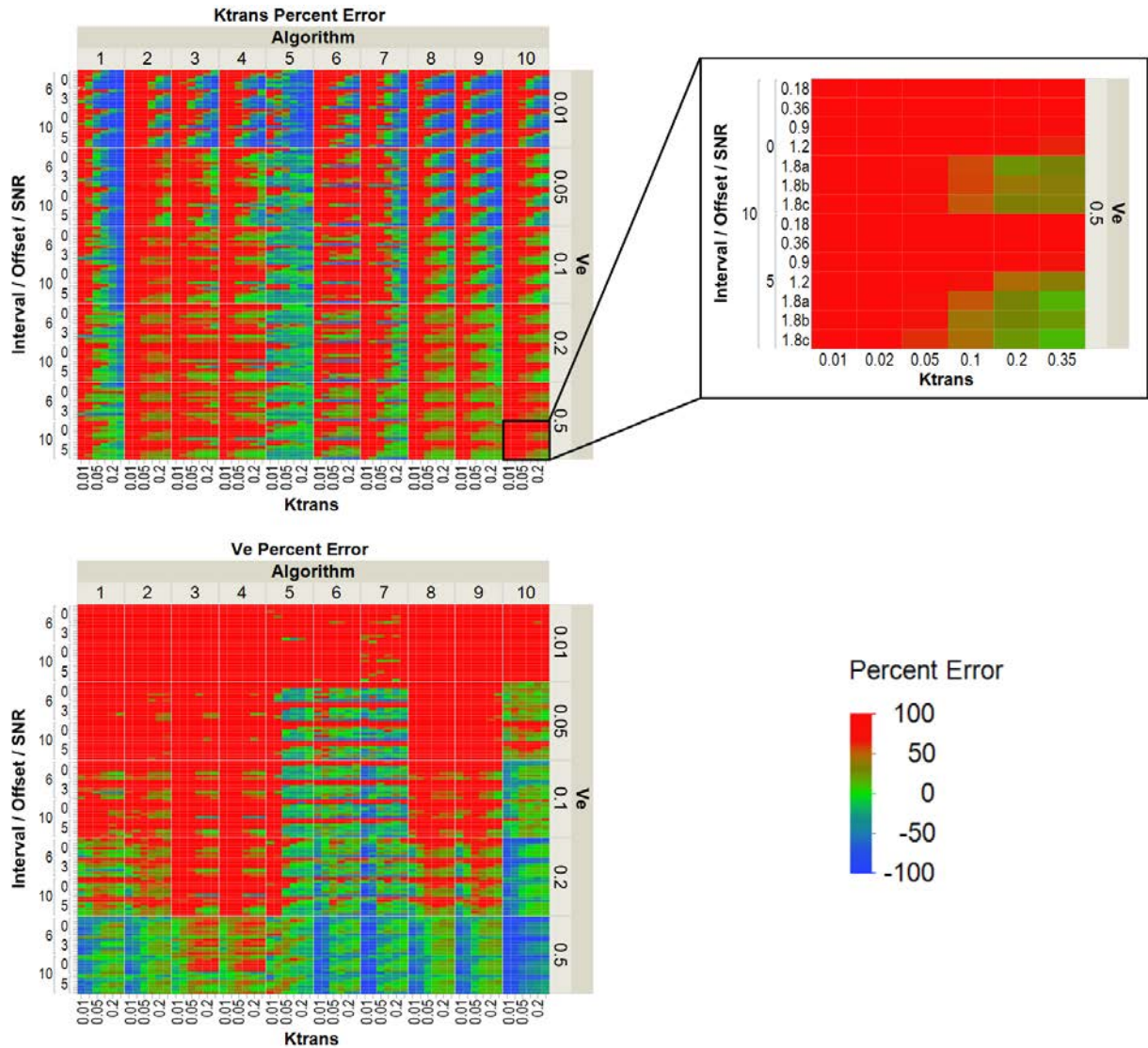


Figure 4-2: Heat Maps of DRO Error.

Heat maps of the percentage error for K^{trans} (top left) and v_e (bottom left) by algorithm in the 28 DROs with noise. The percentage error is defined using the formula $([\text{measured} - \text{simulated}]/\text{simulated} * 100)$. The left side of the heat map is grouped by the timing interval used for the DRO (6 or 10 s), the timing offset used for the DRO (0 or 3 s for the 6 s timing interval, 0 or 5 s for the 10 s timing interval), and the SNR (0.18-1.8). The inset (top right) shows the K^{trans} and SNR values for each block in the heat maps. The maximum percentage error is defined as 100%, and the minimum percentage error is set to -100%. Any errors greater than the maximum percentage are also mapped as 100% error in color. Each DRO is

differentiated by its sampling interval, timing offset, and SNR as determined by the S_0 and sigma value used to create the DRO.

We observed large variation in the percentage of values removed due to the threshold for K^{trans} and v_e for each algorithm. Some algorithms had almost no values removed, and some had a median of 70% of values removed.

These DRO results are for one method of excluding K^{trans} and v_e values. We also analyzed the data using the central 95% of the data for each K^{trans} - v_e pair with no threshold restrictions, which produced consistent test results.

4.3.2 Patients

The percentages of K^{trans} and v_e values removed from patient ROIs because they were outside the bounds of the threshold are shown in Figure 4-3 for the pretreatment, midtreatment, and posttreatment K^{trans} and v_e . As in the DROs, the percentages varied: some algorithms had low percentages removed, implying that they mostly produced realistic values, whereas some algorithms produced almost nothing but unrealistic values for certain patients. The average percentage removed for K^{trans} was 27%, 26%, and 22% for pretreatment, midtreatment, and posttreatment respectively. The average percentage removed for v_e was 46%, 49%, and 48% for pretreatment, midtreatment, and posttreatment respectively.

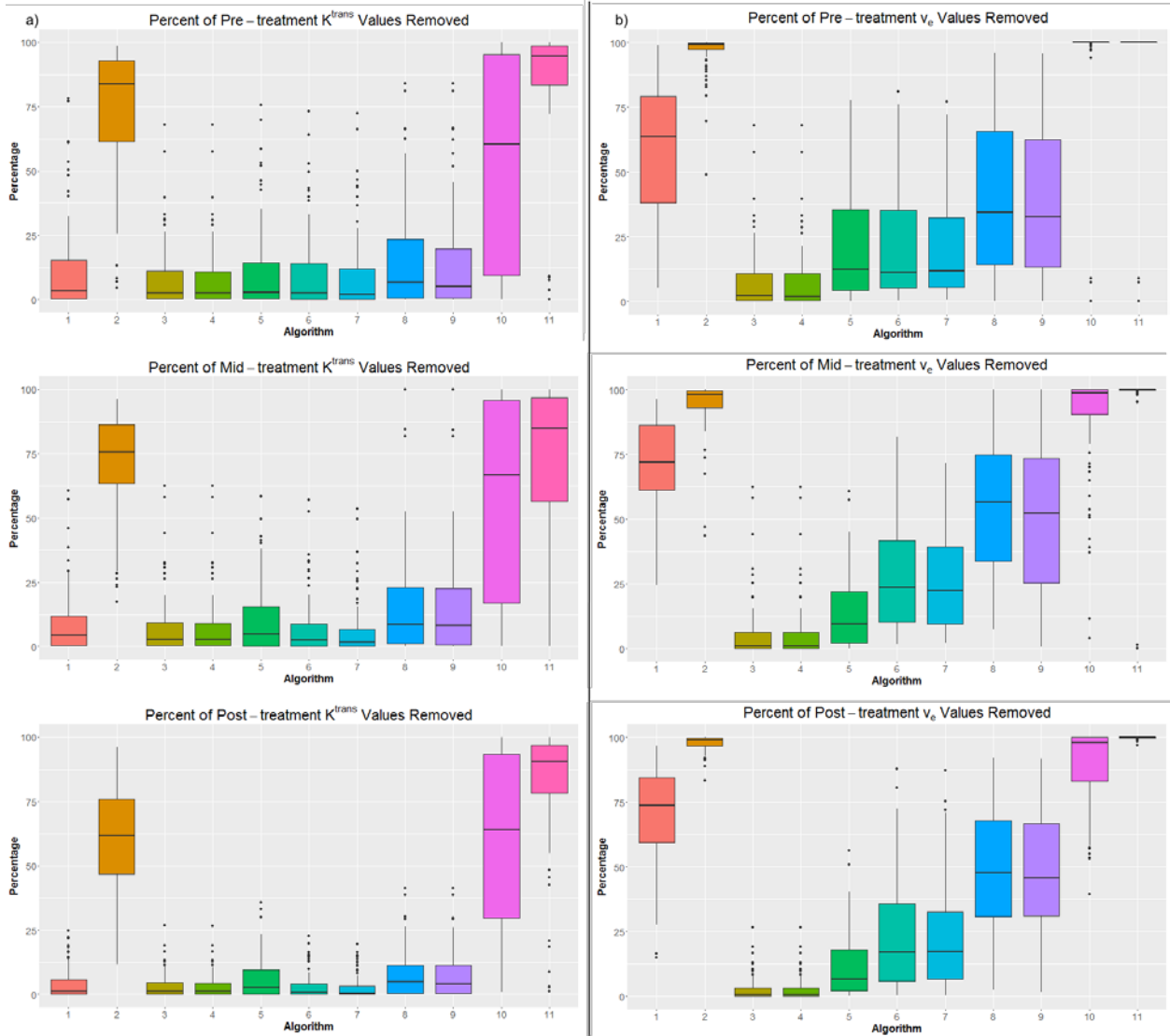


Figure 4-3: Percent of Values Removed.

Percentages of (a) K^{trans} and (b) v_e values removed from patient images. The box plots for each algorithm include the percentages removed for all patients and contours.

According to results of the likelihood ratio test, all differences were statistically significantly ($p < 0.05$) dependent upon the algorithm except for the pretreatment-to-posttreatment change in K^{trans} when all algorithms were included in the model. Algorithms were subset into Tofts-Kermode and extended Tofts groups. In the Tofts-Kermode group, three changes were not statistically significantly dependent on algorithm ($p < 0.05$): pretreatment-to-midreatment change in K^{trans} , midtreatment-to-posttreatment change in K^{trans} , and midtreatment-to-posttreatment change in v_e . In the extended Tofts group, algorithm was not a significant factor ($p < 0.05$) in pretreatment-to-posttreatment change. In all other changes, the algorithm was a significant factor. In all linear mixed effects models, the variance explained by the ROI was much smaller than the residual variance, suggesting that the ROI does not explain much of the variation seen in the linear mixed effects model. All organ variance was less than 30% of the residual variance; on average, it was 8% of the residual variance.

Figure 4-4 demonstrates an example of the difference in parameter values exported from different algorithms. The K^{trans} maps from the same axial slice of a patient are shown for all algorithms. It can be seen that some algorithms output mostly lower K^{trans} values while others output mostly higher K^{trans} values. In addition, some algorithms fit the noise data in voxels outside of the anatomy while other algorithms generated K^{trans} maps only within the anatomy.

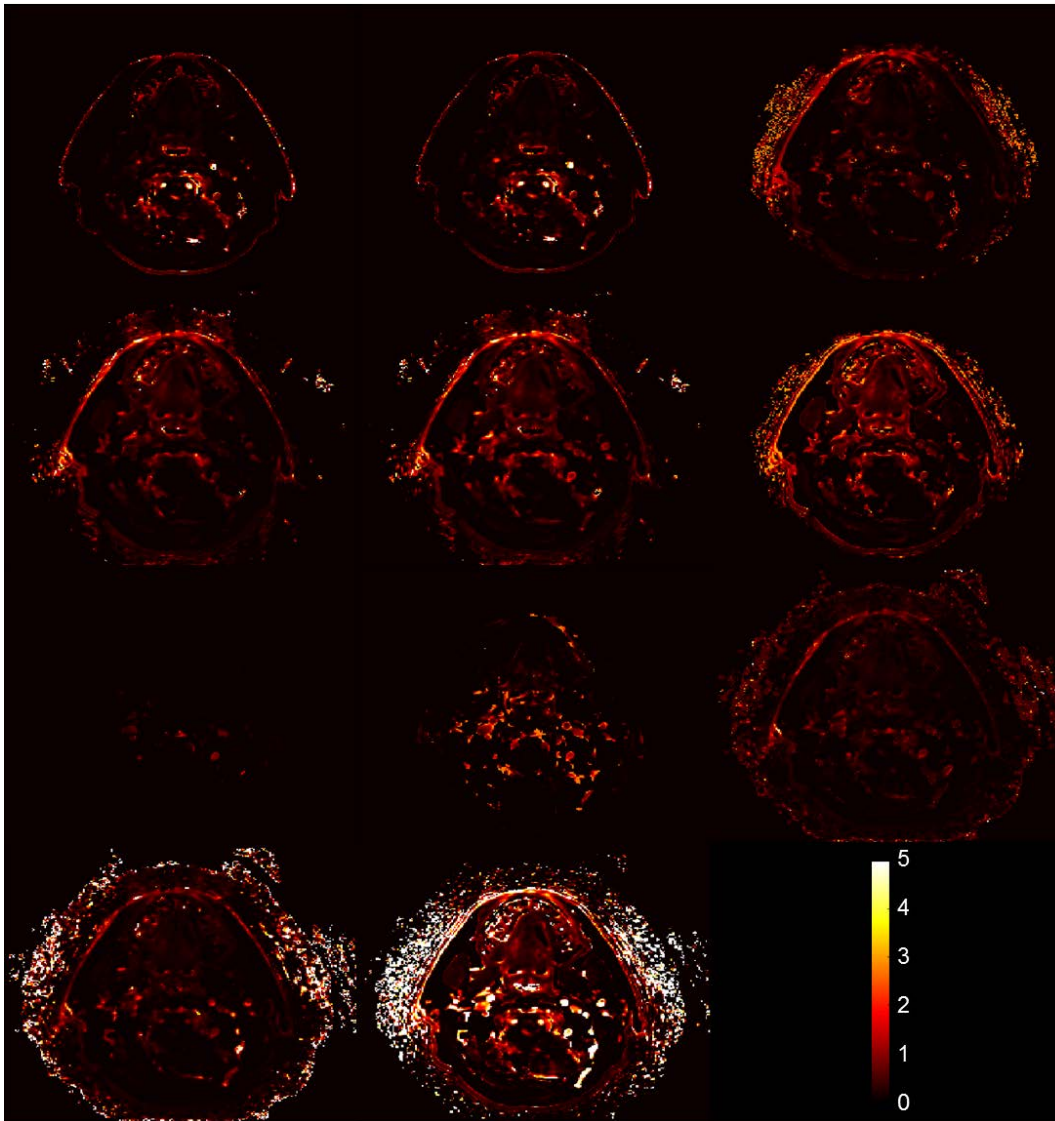


Figure 4-4: Differences in K_{trans} from Algorithms.

Illustration of differences in K_{trans} (min^{-1}) maps exported by different algorithms for one axial DCE-MRI slice.

Carletta's thresholds for good agreement between algorithms ($\alpha \geq 0.8$) and sufficient agreement for tentative conclusions ($0.800 > \alpha > 0.667$) were used [110] to assess the results of Krippendorff's alpha tests. The tests were run using all of the algorithms and also subsets of the algorithms, which were placed into Tofts-Kermode and extended Tofts groups. Of all of these tests, only those in the extended Tofts group had alphas that fell in range for tentative conclusions: 7 of the 108 tested correlations in this group had alphas in this tentative conclusions range. No alphas were in the good agreement range. An illustration of this inconsistent sorting of patients is shown in the appendix (Appendix A Figure A-3). Carletta's thresholds for good agreement and tentative conclusions are weaker than those suggested by others. Krippendorff [111] and Neuendorf [112] suggested using higher standards, which would remove all the metrics found to be partially reliable across algorithms.

Few statistically significant Spearman correlations ($p < 0.05$) were observed: 8% of all tested K^{trans} correlations and 29% of all tested v_e correlations across all algorithms. The only trend in these correlations across algorithms was a statistically significant Spearman correlation of v_e in the GTV-P.

4.4 Discussion

Use of DCE-MRI is increasing in oncology research and investigators have performed many promising studies indicating correlations between predicted therapeutic outcome and DCE-MRI metrics [9, 10]. However, many different DCE-MRI platforms were employed in these studies, and no studies have demonstrated whether data and conclusions regarding HNSCC can be aggregated. We addressed this issue by analyzing the same sets of DRO and HNSCC patient data with a subset of the currently used algorithms that are based on the Tofts-Kermode or extended Tofts model, as these pharmacokinetic models are the ones most commonly employed in DCE-MRI.

The key results from this study are that algorithms were able to determine high values from low values on DROs, but workflow differences may obscure the ability to discern values across algorithms in patients. This may be specifically related to T1 mapping which was not controlled in the

patient portion of this study. Specifically, trends among algorithms from the same institution (institution supplied both Tofts-Kermode and extended Tofts algorithms) were consistent, but not across institutions. This highlights the effect of preprocessing, also shown by the impact of spatial averaging on the calculated error. Therefore, translatability of DCE-MRI across algorithms is not currently feasible.

A digital phantom was used to assess algorithms with a known “ground truth”. The DROs we used had SNRs of 0.18 to 1.80 in the noisy DROs. Although these SNR values and K^{trans} and v_e values within the DRO are below that typically found in head and neck cancer cases [113-116], the DROs were used due to their availability and Quantitative Imaging Biomarkers Alliance-backed quality. The DROs, however, do not come with instructions for interpretation of results, which makes conclusions difficult especially for the DROs that contain very high noise.

The good algorithm performance for the noiseless DRO is consistent with the results reported by Huang et al. [94] and suggest that the algorithms tested here are constructed properly. However, the error increased dramatically when high levels of noise were added to the images. Our assessment using percentage error may explain why the error appeared extremely high in the low K^{trans} and v_e regions as a small absolute error in this region will appear with a high percentage error. Heat maps of the error with the noisy DROs are shown in the appendix (Appendix A Figure A-2) to remove this discrepancy in percentage error between low and high values.

The difference between algorithms was significant for DROs according to the Wilcoxon rank-sum test results, which is consistent with the results reported by Beuzit et al. [95], who used SNRs of 10 and 100 and still found significant differences between different software packages. A limitation of this test is that if the differences between two algorithms are small but all of one sign (such that all values from one algorithm are higher than all values from another algorithm), the differences will be statistically significant. This does not appear to be the cause of the statistically significant differences

observed here because each algorithm has its own error signature, and we could not identify a systematic error in any of the algorithms.

The DRO results demonstrated the potential of DCE-MRI quantitative metrics for clinical application, an illustration of which was the patient data set we used. The significance of including algorithm in the linear mixed effects models was consistent with the Wilcoxon rank-sum test results for the DROs. The small variance explained by the ROI compared with the residual variance in the linear mixed effects models was surprising. If associations can be found using DCE-MRI, different trends between normal tissue and tumor would particularly be expected, yet the ROI provided little explanation of the variance in the data in the linear mixed effects models.

A majority of the algorithms tested produced statistically significant Spearman correlations of v_e in the GTV-P. The agreement of Spearman correlations across algorithms within the GTV-P but not within normal tissue may be due to a difference in contrast-induced signal change, as the GTV-P has a much higher signal change than does normal tissue in DCE-MRI. This means that the GTV-P has higher K^{trans} and lower error in the presence of noise based on the DRO data. However, this agreement of Spearman correlations of v_e in the GTV-P is contradicted by the Krippendorff's alpha results for the GTV-P. Only the midtreatment K^{trans} value in the extended Tofts group had an alpha in the range where tentative conclusions can be drawn. This discrepancy may be explained by small interpatient variability in the K^{trans} and v_e values, which limited the algorithms' ability to separate patients into above or below the median. However, the Spearman rank correlation coefficient identifies trends and is not as affected by interpatient differences in values as Krippendorff's alpha if the trend is consistent.

The Krippendorff's alpha results demonstrated that different algorithms do not consistently classify patients' K^{trans} and v_e values, change in values, or percent change in values. These results indicate that there is currently no clinical level at which these quantitative metrics can be used across algorithms to quantify patients. Based on the algorithms' performance for the DROs in the stratified permutation test in our study, this result from Krippendorff's alpha tests is surprising. However, small

interpatient variation in the K^{trans} and v_e values may have caused the low inter-algorithm reliability. This low inter-algorithm reliability, even within the Tofts-Kermode and extended Tofts groups, contrasts with the results described by Huang et al. [94]. They found good parameter agreement for percentage change when they grouped algorithms by pharmacokinetic model and that all of the algorithms provided good prediction of response to therapy as assessed using univariate logistic regression. This difference may have resulted from the imaging technique used, tissue of interest, and/or patient distribution of K^{trans} and v_e values.

Uncertainties in DCE-MRI exist due to AIF selection and imaging parameters [81, 82, 85-87, 91], but we did not explore them in this study because they were controlled: we examined each algorithm with the same patient DCE-MRI images, variable-flip-angle images, and AIF. In previous studies, T1 mapping and AIF selection impacted K^{trans} and v_e values [81-86, 91, 117-119]. The agreement between algorithms that we observed may have been lower if we had included all of the differences typically seen in a multisite clinical trial, including different scanners, scanning protocols, AIFs, and DCE-MRI algorithms at each institution. In our relatively controlled study, we observed statistically significant differences in both DRO and patient data among the algorithms. It must be acknowledged that there is no “ground truth” against which these algorithms can be compared, and it is unclear whether there was a true therapeutic effect that should have been identified by DCE-MRI in the patient data. Even if there was no net effect across this population of patients, however, it is clear that different approaches to DCE-MRI analysis has significant impact on within-patient trends.

We chose the upper bound for K^{trans} since one of the algorithms in this study used 5 min^{-1} , providing a feasible physical upper limit. We chose the lower bound for K^{trans} because when a given pixel or voxel has a poor fit within an algorithm, it is often given a value of 0 or a negative value. Accordingly, we excluded these values from analysis. We chose the bounds for v_e based on the physical limits given by its definition as a fractional space. Furthermore, poor fits in an algorithm are often mapped to 0 or 1. Therefore, we excluded these values. While 0 is a physically realistic value for K^{trans}

and 0 and 1 are physically realistic values for v_e , these values must be excluded owing to a high proportion of bad pharmacokinetic model fits mapped to these values. The high percentage of values that must be removed represents an area of improvement for future algorithms. Cron et al. [96] demonstrated that as noise in DCE-MRI scans increases, the percentage of nonphysical K^{trans} and v_e values increases. Thus, voxel-based analysis of DCE-MRI quantitative metrics may not be reliable, so global metrics, such as average, of regions must be used for studies. For regions in which a high percentage of values are excluded, the average value extracted is not a reliable metric, as it comes from only a small subregion which is not representative of the whole region. This issue can be mitigated on the imaging end by increasing the SNR at the cost of the increased scan time, poorer temporal resolution, spatial resolution, or coverage, and potentially on the software end by improving how algorithms handle noise through the use of DROs.

In summary, we showed that rigorous standardization and careful quality assurance of software programs, including comparison of parameter calculations with standard data sets, are needed for collating pharmacokinetic analysis of DCE-MRI data among different algorithms. This must include assessment of the impact of image noise on quantitative metric error. Authors recently reported the need for careful quality assurance for functional MRI [120]. Efforts like those by the Quantitative Imaging Biomarkers Alliance to standardize DCE-MRI acquisition parameters represent a natural step forward for quality assurance and serve as the foundation for the current quality assurance work used in the present study.

To support these efforts, we provided our data set in a repository to allow for their use as perpetual head and neck cancer patient-derived standards for future DCE-MRI software and/or algorithm development [121] in addition to the extant DRO library maintained by one of the authors (D. Barboriak). To that end, we recommend the following:

1. Consistent use of the same software for DCE-MRI analysis within a given study and for cross-comparisons between studies.

2. Specification and setting of acquisition parameters before proceeding with clinical trials as with the present data set.
3. Before performing multi-institution clinical trials, confirmation that DCE-MRI parameter values are consistent across institutions.
4. Inclusion of reference to a DRO with clinically relevant SNRs to benchmark performance of DCE-MRI software using clear evaluation criteria.

Clinically, our DRO data point to the fact that algorithms differed substantially despite reliance on the same basic underlying pharmacokinetic model(s), performing relatively stable in low-noise conditions. This, coupled with the inter-algorithm variability observed with the *in vivo* head and neck cases (which were performed in immobilization on a single MRI platform with standard AIF selection) suggests that, at present, any clinical trial desiring to implement DCE-MRI, should at a minimum, use a single pre-specified DCE-MRI software workflow, and eschew use of multiple algorithms. This also means that DCE-MRI findings from one software are broadly uninterpretable in a differing platform at present.

Until quantitative metrics can be reliably calculated across algorithms, patient-derived DCE-MRI analyses with different algorithms cannot be aggregated. Semiquantitative metrics, such as the area under the curve, have been shown to be more reproducible than quantitative metrics and may be the best interim option for use in prognostic studies using different algorithms [122]. Further refinement is required before DCE-MRI software-derived parameters can be used as a routine cross-institutional metric for multi-site clinical trials.

Chapter 5 : Computed Tomography Radiomics Feature Dependence on Tube Voltage

5.1 Introduction

CT imaging protocols can impact radiomics feature values. Many of the reconstruction parameters have been investigated. Several recent studies have also investigated the effect of some acquisition parameters. Fave et al. conducted a preliminary investigation of the effect of tube voltage on radiomics features [123]. Images at different tube voltages were simulated by applying an offset to patient images that were acquired at 120 kVp. They found that tube voltage did not affect feature values as the inter-patient variation was significantly higher than the intra-patient variation. Mackin et al. evaluated the effect of tube current changes using a phantom with inserts of different textures. They concluded that tube current did not have a significant effect on radiomics features in textured objects [38]. Berenguer et al. used a similar radiomics phantom as that used by Mackin et al. and found that the tube current had no effect on over 70% of radiomics features using cut-offs for the metrics of coefficient of variation, quartile coefficient of dispersion, and intraclass correlation coefficient [124]. They found that tube voltage affected more features than tube current, as only 42%-68% of radiomics features passed the cut-offs for the same metrics. In that study, the robustness of features was computed based solely on changes due to different parameter settings; in this study, we related the robustness of features due to changes in tube voltage to patients.

Until the recent study by Berenguer et al.[124], the effect of tube voltage on radiomics feature values had not been thoroughly investigated using multiple acquired CT scans. Tube voltage experiments require additional CT scans; this increases the dose to the patient, unlike the voxel size, image thickness, and convolution kernel, which can be analyzed by creating additional reconstructions after one CT scan has been acquired. While most clinics use 120 kV for their CT examinations, there

have been several studies that have recommended adjusting the tube voltage based on the patient size or the purpose of the examination [125-131].

In the current study, we investigated the effect of tube voltage on radiomics feature variability using a phantom with different cartridges by comparing the induced change to the variability between patients in two different cohorts. This was done using a phantom developed by Mackin et al. [132]. We evaluated features using four different pre-processing techniques and compared them in two patient cohorts using the same pre-processing techniques to determine the relative effect of tube voltage-induced radiomics feature changes. On the basis of the results of studies by Fave et al. and Mackin et al. [38, 123], we hypothesized that tube voltage would not have a significant effect on radiomics features measured in textured materials.

5.2 Methods

The credence cartridge radiomics (CCR) phantom developed by Mackin et al. [132] was used to evaluate the effect of tube voltage on radiomics features. For this study, acrylic, wood, cork, dense cork, and rubber cartridges were analyzed. These cartridges were organized by the amount of texture, as determined by the standard deviation of the CT numbers within the material. The order of the materials, from the least texture to the most texture, was determined using the standard deviation (sd): acrylic (sd: 4), cork (sd: 34), wood (sd: 36), dense cork (sd: 47), and rubber (sd: 100). The difference between a material with low texture and high texture is demonstrated in Figure 5-1: acrylic has no texture pattern, while cork has a rough texture pattern. Images of the other cartridges can be seen in the report by Mackin et al. [132].

5.2.1 Effective Atomic Number of Phantom Materials

Materials of similar effective atomic number will have similar relative contributions of photoelectric and Compton interactions, so the impact of x-ray energy on the attenuation coefficient

(and therefore Hounsfield Units) will be similar. Therefore it was necessary to estimate the effective atomic number for each material in the phantom.

The effective atomic number was calculated for each material [133]. The weight fractions for each element in the material were taken from the literature for cork, dense cork, and wood [134, 135]. The weight fractions for the other materials could not be found in the literature and instead were calculated based on their molecular formulas: acrylic = $C_5H_8O_2$ and rubber = C_5H_8 .

5.2.2 Phantom Scans

The CCR phantom was imaged on two CT scanners: a GE Lightspeed RT (GE Healthcare) and a Philips Brilliance Big Bore (Philips Healthcare, Eindhoven, The Netherlands). On each scanner, the impact of tube voltage was assessed by keeping the CT dose index ($CTDI_{vol}$) constant. A recent study showed that the tube current does not affect feature values [38]; thus, the tube current changes that were necessary to keep the $CTDI_{vol}$ constant as the tube voltage changed was not a confounding factor. The imaging protocols used were designed to align as much as possible with the controlled protocol used by Ger et al. on 100 CT scanners [136].

For the GE CT scans, the $CTDI_{vol}$ was set as close to 13.3 mGy as possible (mean, 13.27 mGy; standard deviation, 0.12 mGy). The acquisition parameters were as follows: 0.75 pitch, 1.0 s rotation time, 2.5 mm image thickness, 0.976 mm x 0.976 mm pixel size, body bowtie filter, and 4 x 1.25 collimation. The tube voltage varied from 80 to 140 kVp in increments of 20, with tube current values of 315, 175, 105, and 75 mA for the four tube voltage settings, respectively.

For the Philips CT scans, the $CTDI_{vol}$ was set as close to 13.3 mGy as possible (mean, 11.6 mGy; standard deviation, 3.4 mGy). The acquisition parameters were as follows: 0.938 pitch, 1.0 s rotation time, 3 mm image thickness, 0.976 mm x 0.976 mm pixel size, body bowtie filter, and 16 x 1.5 collimation. The tube voltage varied among three settings of 90, 120, and 140 kVp, with tube current values of 265, 250, and 150 mA for the three tube voltage settings, respectively.

The half-value layer for the GE CT with the body bowtie filter at 80, 100, 120, and 140 kVp was measured to be 5.5 mm Al, 6.7 mm Al, 7.8 mm Al, and 8.8 mm Al, respectively. The half-value layer for the Philips CT with the body bowtie filter at 90, 120, and 140 kVp was measured to be 7.1 mm Al, 8.7 mm Al, and 9.8 mm Al, respectively.

5.2.3 Patient Scans

For this study, we retrospectively reviewed the images and medical records of NSCLC and HNSCC patients with a waiver of informed consent from the Institutional Review Board at the University of Texas MD Anderson Cancer Center. Both of these cancer types have been investigated in previous studies, and for both, radiomics features added to models with conventional prognostic factors (e.g., age) have demonstrated promising prognostic ability [15, 17, 20, 22, 29].

Twenty patients with NSCLC were selected. Their mean age was 67 years (range, 52-78 years), mean height 170 cm (range, 154-182 cm), mean weight 72.9 kg (range, 41.0-97.6 kg), and mean tumor volume 77 cm³ (range, 11-389 cm³). The non-contrast CT scans were acquired on a GE Discovery CT scanner with the following parameters: 120 kVp, 300 mA, 1.35 pitch, 0.5 s rotation time, 2.5 mm image thickness, and 0.976 mm x 0.976 mm pixel size. Patients' tumors were contoured by the treating radiation oncologist.

Thirty patients with HNSCC were selected. Their mean age was 64 years (range, 50-87 years), mean height 175 cm (range, 149-193 cm), mean weight 80.5 kg (range, 43.9-114.9 kg), and mean tumor volume 13 cm³ (range, 1.2-91 cm³). The contrast-enhanced CT scans were acquired using a GE LightSpeed CT scanner with the given parameters: 120 kVp, 220 mA, 1.375 pitch, 1.0 s rotation time, 1.25 mm image thickness, and 0.488 mm x 0.488 mm pixel size. Patients' tumors were contoured by a radiation oncologist.

5.2.4 Radiomics Feature Analysis

Images were analyzed using IBEX, an open-source radiomics tool [137, 138]. The CCR phantom scans were contoured using the location of a small radiopaque marker on the edge of the phantom; the x, y, and z location of the marker was entered into an in-house Python (version 2.7) script, creating an 8 x 8 x 2 cm³ region of interest (ROI) for each cartridge. Forty-nine features were calculated in IBEX: 11 intensity histogram, 22 gray level co-occurrence matrix [139], 11 gray level run length matrix [140, 141], and five neighborhood gray tone difference matrix features [142] (Table 5-1). Each feature was calculated four times: (1) with no image pre-processing other than thresholding, (2) with thresholding and a Butterworth smoothing filter with an order of 2 and a cut-off of 125, (3) with thresholding and 8-bit depth resampling, and (4) with thresholding, Butterworth smoothing, and 8-bit depth resampling. Pre-processing has been demonstrated to affect the prognostic value of features [143]. The settings used for the features and the pre-processing were based on the work of Fave et al., who found that these were the most prognostic and appropriate for noise levels in CT [143]. No voxel resampling was used for this analysis as all images on each scanner had the same voxel size.

The lower threshold for the NSCLC patients was -100 HU, and the upper threshold was 200 HU. For the HNSCC patients, only a lower threshold of -100 HU was applied. These thresholds were chosen to remove air and bone from the ROIs. An upper threshold was not applied to the HNSCC patients as this sometimes removed contrast from the ROIs. For the phantom images, no thresholding was applied. The settings used for each feature are described in detail in Fave et al.'s supplemental material [14]. For gray level co-occurrence and gray level run length matrix features, the feature value used in the analysis was averaged over the angles.

Table 5-1. Radiomics Features Analyzed

Gray Level Co-occurrence	Gray Level Run Length	Neighborhood Gray Tone Difference	Intensity Histogram
Auto Correlation	Gray Level Nonuniformity	Busyness	Energy
Cluster Prominence	High Gray Level Run Emphasis	Coarseness	Entropy
Cluster Shade	Long Run Emphasis	Complexity	Maximum
Cluster Tendency	Long Run High Gray Level Emphasis	Contrast	Mean
Contrast	Long Run Low Gray Level Emphasis	Texture Strength	Median
Correlation	Low Gray Level Run Emphasis		Minimum
Difference Entropy	Run Length Nonuniformity		Standard Deviation
Dissimilarity	Run Percentage		Uniformity
Energy	Short Run Emphasis		Kurtosis
Entropy	Short Run High Gray Level Emphasis		Skewness
Homogeneity	Short Run Low Gray Level Emphasis		Variance
Homogeneity 2			
Information Measure Correlation 1			
Information Measure Correlation 2			
Inverse Difference Moment Norm			
Inverse Difference Norm			
Inverse Variance			
Max Probability			
Sum Average			
Sum Entropy			
Sum Variance			
Variance			

5.2.5 Statistical Methods

The Spearman correlation of feature values with tube voltage was calculated with the null hypothesis that there is no correlation. The Spearman correlation was computed for each feature in a material with a given pre-processing technique using R (R software package, version 3.4.3). As the Spearman correlation cannot handle ties, the p-value was computed using permutations through the R software package coin (version 1.2-2). To determine how common statistically significant correlations were, we tallied the total number of instances in which a given feature had a statistically significant correlation ($p < 0.1$). This p-value was chosen due to the small number of tube voltage data points that could be acquired which limits small p-values from being achieved. Multiple hypothesis testing was not used for the Spearman correlation tests in order to give a “worst case” estimate as multiple hypothesis testing would shift more instances to non-significant p-values.

The variability in radiomics features induced by changing the tube voltage was evaluated by comparing it to the variability in radiomics features in patients:

$$Patient - normalized phantom range = \frac{Phantom Range_{f,m} / \overline{x_{f,m}}}{s_{p_f} / \overline{x_{p_f}}} \quad (5-1),$$

where $Phantom Range_{f,m}$ is the range of values across the different tube voltage scans for a given feature f and phantom material m ; $\overline{x_{f,m}}$ is the mean of values across the different tube voltage scans for a given feature and phantom material; s_{p_f} is the patient standard deviation for a given feature; and $\overline{x_{p_f}}$ is the patient mean for a given feature. Larger values of the patient-normalized phantom range showed that the variability in feature values due to changes in tube voltage were large in comparison to inter-patient variation. Thus, this feature dependence on tube voltage could affect their use in prognostic model building.

Paired Student's t-tests were used to compare the results of the different pre-processing techniques, patient cohorts, and vendors. These tests were initially run as two-sided tests with the null

hypothesis that the difference between the means is 0. If a statistical difference was found, one-sided t-tests were then run to determine which of the two subjects (the two pre-processing techniques, patient cohorts, or vendors in the specific t-test) in the comparison was larger. A p-value of 0.05 was used to determine statistical significance.

5.3 Results

5.3.1 Effective Atomic Number

The calculated effective atomic number values are summarized in Table 5-2. In addition, this table contains effective atomic numbers for bone, fat, and muscle for comparison to the cartridge material values [133].

Table 5-2. Estimated Atomic Number of Phantom Cartridges and Human Tissues

Cartridge or Tissue	Effective Atomic Number
Acrylic	6.56
Wood	7.16
Cork	6.86
Dense cork	6.61
Rubber	5.37
Bone	12.31 [133]
Fat	6.46 [133]
Muscle	7.64 [133]

5.3.2 Spearman Correlation of Features with Tube Voltage

The Philips scanner data only contained three data points of different tube voltages; thus, it was difficult to obtain statistically significant Spearman correlations. Table 5-3 summarizes the percentage of total features within each feature group with statistically significant ($p < 0.1$) Spearman correlations. This table only includes the GE scanner data. The material-, scanner-, and feature-specific Spearman rho and p-value data are provided in Appendix B. All of the feature groups (gray level co-occurrence, gray level run length, neighborhood gray tone difference, and intensity histogram) had features that were correlated with tube voltage. The gray level run length matrix features were most often correlated with tube voltage from the four categories.

Table 5-3. Percentage of Features with Significant Spearman Correlation

Feature Group	Features with Significant Spearman Correlation (%)
GLCM	29.5
GLRLM	55.5
NGTDM	22.0
IH	51.4

GLCM: gray level co-occurrence matrix; GLRLM: gray level run length matrix; NGTDM: neighborhood

gray tone difference matrix; IH: intensity histogram

The strength of the correlation was material dependent. Figure 5-1 shows axial slices of the acrylic and cork cartridges at each tube voltage from the GE scans analyzed using only thresholding. Visually, there is little to no discernable change in these images. However, the feature plots next to these images in the figure show that the features changed with the tube voltage. In addition, the change and trend with the tube voltage depended on the cartridge material. The four features in the figure are from different feature groups; thus, each group was represented. While the Spearman correlation p-value was 0.17 for all four features for acrylic, the trend is clear. Because of the limited number of permutations that are possible with only four distinct tube voltages, low p-values were difficult to achieve. In contrast, all p-values for cork were much higher. However, while the Spearman correlation was high for acrylic, there was always one tube voltage data point that did not follow the trend. The tube voltage that this occurred at was not consistent across the four features.

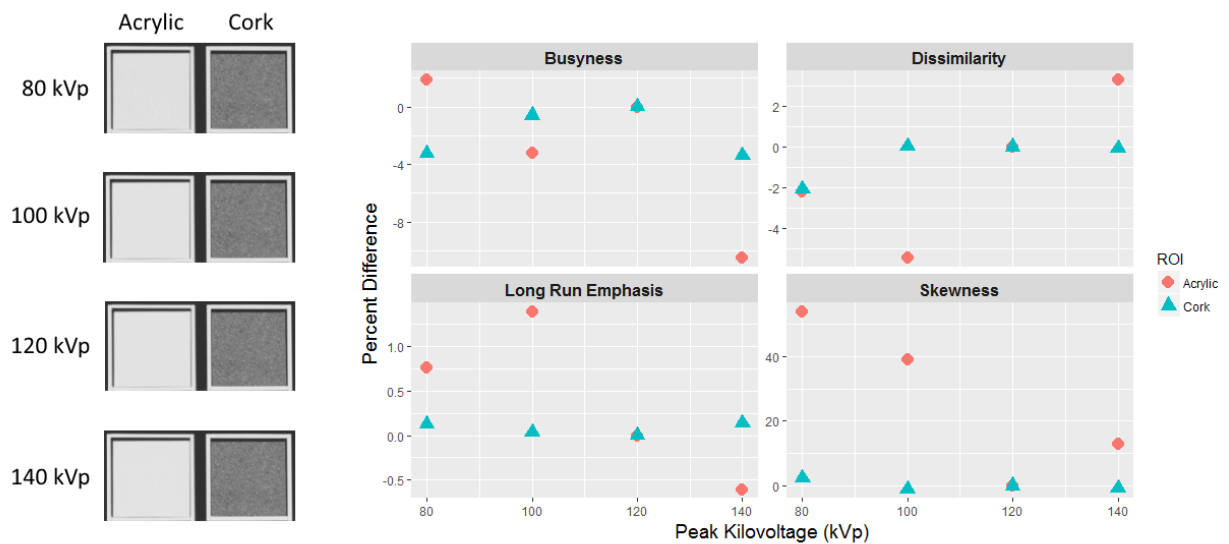


Figure 5-1: Acrylic and Cork Changes with Tube Voltage.

Axial slices of the acrylic and cork cartridges at 80, 100, 120, and 140 kVp are shown from the data set on the GE scanner. In these slices it can be seen that acrylic has low texture while cork has a rough texture pattern. Next to the axial slices are feature plots using thresholding pre-processing, with one feature from each feature group. The percentage difference of the feature value relative to the feature value at 120 kVp is shown for acrylic (red circles) and cork (blue triangles). There was a change in the features with tube voltage, but it was material dependent. For busyness, acrylic has a Spearman rho of -0.8 with a p-value of 0.17, and cork has a rho of -0.2 with a p-value of 0.73. For dissimilarity, acrylic has a rho of 0.8 with a p-value of 0.17, and cork has a rho of 0.2 with a p-value of 0.73. For long run emphasis, acrylic has a rho of -0.8 with a p-value of 0.17, and cork has a rho of 0.2 with a p-value of 0.73. For skewness, acrylic has a rho of -0.8 with a p-value of 0.17, and cork has a rho of -0.4 with a p-value of 0.49. Window width: 1600, window level: -300.

5.3.3 Patient-Normalized Phantom Range

We computed the patient-normalized phantom range using data from both scanners as the phantom range not impacted by the scarcity of data unlike the Spearman correlations described above. Figure 5-2 shows a heat map of the patient-normalized phantom range for the GE scanner when the pre-processing technique was only thresholding. The patient-normalized phantom ranges, calculated using the two different patient populations, are shown for each material, ordered from the least to the most texture. The heat maps for the other pre-processing techniques and for the Philips scanner are provided in the Appendix B.

According to Figure 5-2, the gray level run length matrix features performed the worst, indicating that they are the most sensitive to changes in tube voltage. In addition, the gray level co-occurrence matrix features of auto correlation, sum average, and sum variance consistently had higher patient-normalized phantom range values across the materials than the other gray level co-occurrence matrix features. The intensity histogram features of mean, median, and minimum also performed poorly and had high patient-normalized phantom range values for all materials and patient cohorts, as shown in Figure 5-2. However, these three intensity histogram features had very little patient variation (i.e., small s_{p_f} in equation 1); almost any change due to tube voltage in the phantom would cause these features to have high patient-normalized phantom range values. For example, kurtosis and mean both had very small ranges across different tube voltage scans, but the spread of kurtosis values among patients was large while it was small for mean; thus, kurtosis had much lower patient-normalized phantom range values than mean.

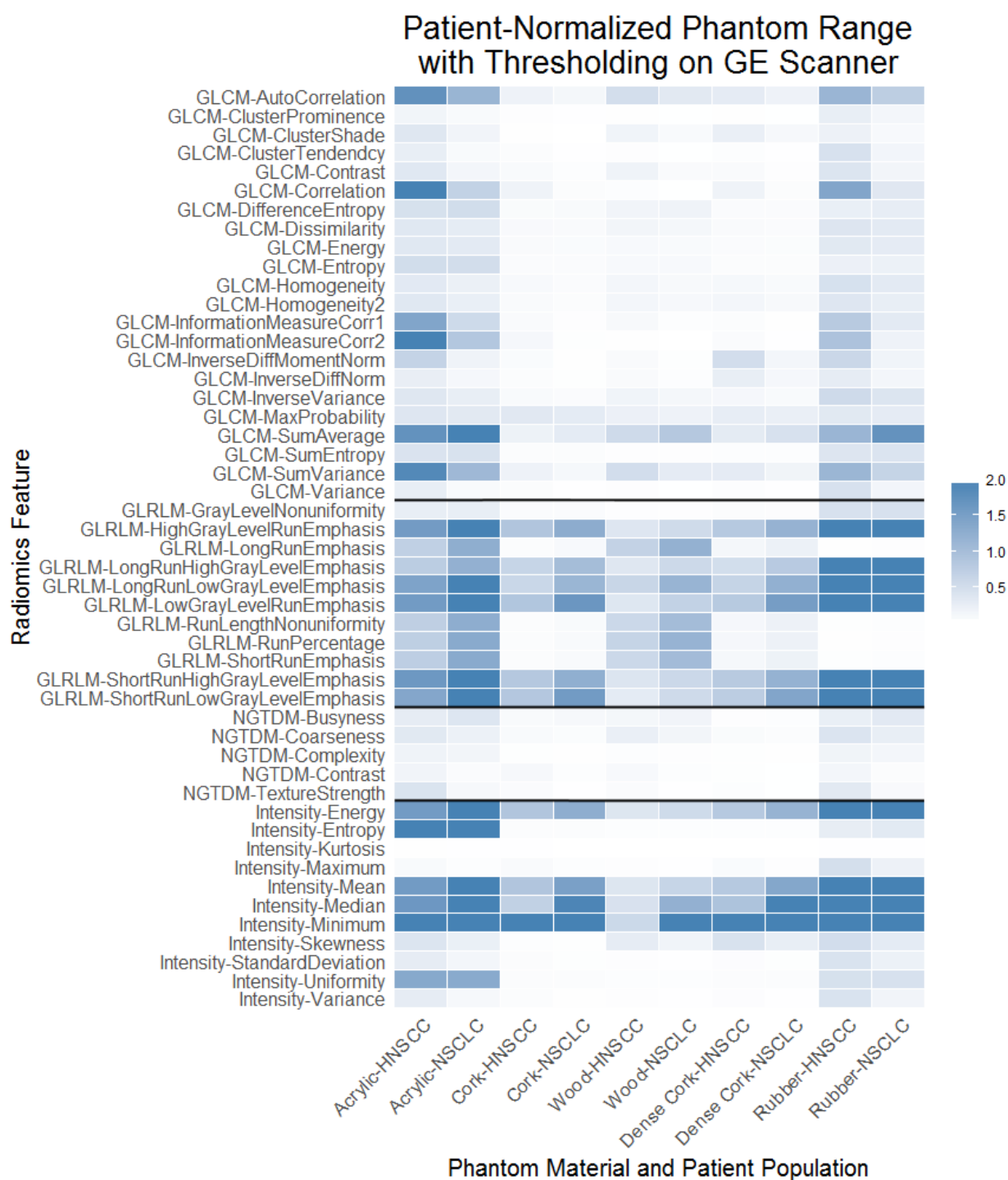


Figure 5-2: Heat map of Patient-Normalized Phantom Range Values.

The patient-normalized phantom range values are shown for each feature on the basis of each material and patient population. The values are between 0 and 2; any values above 2 were mapped to the maximum color. The materials along the x-axis are listed in order from least to most texture on the

basis of the measured standard deviation of the CT numbers of the material. The patient-normalized phantom range that was calculated using the two patient cohorts appears together for each material. Each feature along the y-axis is identified first by the acronym for the feature group that it is part of. Each feature group is shown together with black lines separating the feature groups within the heat map. Textured materials have lower patient-normalized phantom range values. In addition, the gray level run length matrix features can be seen to have the highest values in comparison to the other feature groups. GLCM: gray level co-occurrence matrix; GLRLM: gray level run length matrix; NGTDM: neighborhood gray tone difference matrix; HNSCC: head and neck squamous cell carcinoma; NSCLC: non-small cell lung cancer.

There was also a clear difference in patient-normalized phantom range values across the different materials: features from low-texture materials had higher patient-normalized phantom range values. According to the results of one-sided paired t-tests, the order of the materials, from statistically higher to lower patient-normalized phantom ranges, was acrylic, rubber, cork, dense cork, and wood. Acrylic did not have statistically significantly ($p = 0.13$) higher patient-normalized phantom ranges than rubber, and dense cork did not have statistically significantly ($p = 0.17$) higher patient-normalized phantom ranges than wood. Interestingly, rubber, the material with the highest texture, had higher patient-normalized phantom ranges than cork, dense cork, and wood. This is discussed further in the Discussion.

The patient-normalized phantom range values were statistically significantly ($p < 0.05$) different when bit-depth rescaling, smoothing, and thresholding were all used for pre-processing. The patient-normalized phantom range values for this pre-processing technique were statistically significantly higher than were those when the other pre-processing techniques were used. The other pre-processing techniques did not produce patient-normalized phantom range values that were statistically significantly different.

The patient-normalized phantom range values, when computed using the HNSCC patient cohort, were statistically significantly ($p = 0.003$) lower than were those using the NSCLC patient cohort. The patient-normalized phantom ranges were not statistically significantly ($p = 0.75$) different between the GE and Philips scanners.

5.4 Discussion

In this study, we investigated the effect of tube voltage on radiomics texture values. The tube voltage used in patient scans may vary because of differing protocols and scanners; thus, its effect must be investigated before these patients can be combined into a single cohort. The Spearman correlation test showed that about half of the features in several feature groups were correlated with tube voltage,

which was expected because of the association between HUs and tube voltage [144]. Thus, it was important to determine the relative impact of this change in feature values due to tube voltage compared to the inter-patient variation in order to evaluate if the feature value change could impact patient studies. Overall, radiomics features were more robust to changes in tube voltage when measured in materials with more texture; thus, in radiomics studies, differences in the tube voltage of patients' CT scans is not an important factor.

The cartridge materials used had effective atomic numbers close to those of human tissues. Acrylic, cork, rubber, and wood had effective atomic numbers close to those of fat and muscle. In addition, the calculated values were close to those reported in the literature. The calculated effective atomic number for acrylic was 6.56, which is slightly higher than that found in the literature (6.43-6.47) [145, 146]. The calculated effective atomic number for wood roughly agreed with the Monte Carlo calculations by Marashdeh et al. [135], who found that for wood particles less than 50 μm in size, the effective atomic number was about 7 for energies between 45 keV and 60 keV. Elias et al. showed that the percentage of carbon in the rubber can affect the effective atomic number, with it ranging from 4.99 to 8.66 [147]. Our calculated value fell within this range, implying that it is a reasonable estimate. In addition, the cork and rubber cartridges have been previously shown to have textures similar to those of NSCLC tumors [132].

The gray level run length feature group performed the worst of the feature groups with most of its features consistently having high patient-normalized phantom range values. This may be due to the effect of tube voltage on image contrast, which would change the proportion of voxels in each gray level. The gray level run length matrix sums over the same range of values as the minimum and maximum are specified values, but the proportion of values in the lower or higher HU range may differ. This would affect the features that stress high or low gray levels, such as low gray level run emphasis (with high patient-normalized phantom range values, as shown in Figure 5-2), which is defined as

$$LGRE = \frac{1}{n_r} \sum_{i=1}^M \frac{p_g(i)}{i^2} \quad (5-2),$$

where n_r is the total number of runs, i is the gray level, M is the total number of gray levels, and

$$p_g(i) = \sum_{j=1}^N p(i, j) \quad (5-3)$$

is the sum distribution of the number of runs with gray level i , run length j , maximum run length N , and run-length matrix $p(i, j)$. It can clearly be seen how a scan at a different tube voltage in which there is a shift in voxels within a certain gray level would be affected in this formula: M stays the same, but $p_g(i)$ differs, while the emphasis of i^2 in the denominator remains the same.

The results of this study were similar to those from Mackin et al., who found that mAs did not affect radiomics features when measured in materials that had texture (i.e., radiomics features that were not measured in acrylic) [38]. Measuring features in the acrylic cartridge represents measuring the noise characteristics due to the different parameters. Therefore, both of these studies showed that the inherent texture of a material is more important than the noise created by varying acquisition parameters. These results are somewhat consistent with those from the study by Fave et al. [123], who simulated patient scans at different tube voltages and found that intra-patient variation was significantly lower than inter-patient variation for all features tested. In our study, the difference between scans at different tube voltages was not given by a HU shift applied to all voxels, which may explain the finding that some features changed significantly in comparison to inter-patient variation.

These results are also somewhat consistent with Berenguer et al. [124]. They found that 43% of features had a coefficient of variation of less than 10% and that 55% of features had a coefficient of variation of less than 15% because of changes in tube voltage. In our study, 77% and 75% of the features measured on the GE and Philips scanners, respectively, had a coefficient of variation of less than 10%, and 84% and 81% of the features had a coefficient of variation of less than 15%. We examined 196 features (four pre-processing techniques and 49 features), while Berenguer et al. examined 177 features without pre-processing. In addition, we calculated gray level co-occurrence features and neighborhood gray tone difference features in 2D, while they calculated them in 3D. These differences may account for the larger number of features that had a smaller coefficient of variation in

our data set. Another potential cause for the difference was that we only had a few data points because of the limited tube voltage settings on the scanners; therefore, the coefficient of variation may not be a stable metric in such cases. Our results also mostly agree with Buch et al. [148]. Our specific additional contributions were that our study used a validated phantom that had similar features as patients [132] and a common, open-source radiomics tool.

In this study, rubber had the highest texture. We had assumed that the cartridge material referred to as “rubber” was pure rubber when calculating the effective atomic number. However, it was ground-up tires; thus, less than 50% was rubber, and 17% was metal [149]. This high proportion of metal in the material explains why the rubber cartridge was not statistically significantly different from the acrylic cartridge. The high atomic number components caused the changes in the tube voltage to dramatically affect features measured in the rubber cartridge.

In previous studies in which patient cohorts were used to compare to phantom results, only one patient cohort was typically used. In this study, we used two distinct patient cohorts and found them to have statistically different results. The HNSCC patient cohort likely produced smaller patient-normalized phantom range values because these patient scans had contrast; thus, there may have been a larger patient spread in radiomics feature values.

There were a few limitations to this study. First, constant $CTDI_{vol}$ was difficult to achieve using the parameters that could be changed on the Philips machine while trying to keep the pitch, rotation time, image thickness, and pixel size constant; thus, the variability of the $CTDI_{vol}$ was much larger for the Philips scanner than for the GE scanner. Secondly, it was conducted in a phantom; thus, the true implications for patient scans are not known. Conducting this study in patients would require an unnecessarily excessive radiation dose. The effective atomic number of the phantom cartridges was similar to that of human tissue; thus, reasonable conclusions can be drawn for patients’ tumor and normal tissues on the basis of the phantom results. In addition, because of the construction of the phantom, the conclusions in this study only apply to non-contrast CT scans. The effect of changing tube

voltage on iodine-containing tissue would be different. The phantom design was also not specific to a body part and did not include build-up which could affect beam hardening and scatter, which could consequently affect texture. Additionally, only one scan was taken at each tube voltage setting. The tube voltage and current can differ slightly between scans; however, in previous experiments, we found that the coefficient of variation of features acquired under the same settings was typically below 1%.

Lastly, only two scanners were evaluated. Mackin et al. found that different scanners produced different feature values, but the features typically clustered by vendor [132]. Our study was focused on the magnitude of feature changes induced by altering the tube voltage; therefore, a large sample of scanners was not necessary as each scanner represented its vendor. The two scanners used in this study were from two different vendors and thus had different tube voltage settings. However, the same trends were found on both scanners, as shown by the t-test on the patient-normalized phantom range values and the heat maps. Specifically, we found that features in high-textured materials were more robust to tube voltage changes than were features in low-textured materials. This study and that by Berenguer et al. [124] found that the scanner did not affect the results. Thus, it is reasonable to presume that the same trend would be found for other GE and Philips scanners and scanners from other vendors.

5.5 Conclusions

In this study, we found that changes in the tube voltage had less effect on the measured radiomics feature values of phantom cartridges with high texture than on those with low texture (i.e., cork vs acrylic). High-texture cartridges are more representative of patient tumor tissue; thus, features measured in tumors are also not expected to be significantly affected by tube voltage. Several features had consistently high patient-normalized phantom range values, indicating that they were unreliable and should be used with caution in studies that include patients scanned with different tube voltages. Based on the results from this study, we recommend using the most common tube voltage setting for

the particular anatomical site of interest when conducting prospective studies, as this eliminates any effect of tube voltage on radiomics feature values shown by the patient-normalized phantom range values. However, in retrospective studies, a patient would not have to be excluded from the cohort due to differences in acquisition tube voltage. This is because the differences in radiomics feature values will be small, and therefore patient stratification will not be affected, as long as non-robust features are not used.

Chapter 6 : Impact of Head and Neck Artifacts on Computed Tomography

Radiomics Features

This chapter is based upon:

Ger, RB, Craft, DF, Mackin, DS, Zhou, S, Layman, RR, Jones, AK, Elhalawani, H, Fuller, CD, Howell, RM, Li, H, Stafford, RJ, Court, LE. Practical Guidelines for Handling Head and Neck Computed Tomography Artifacts for Quantitative Image Analysis. *Computerized Medical Imaging and Graphics* doi: 10.1016/j.compmedimag.2018.09.002. Volume 69, Pages 134-139. 2018. ©Elsevier.

The permissions for reuse of these materials were obtained from Elsevier.

6.1 Introduction

While radiomics studies have identified several imaging features that are associated with prognosis, these findings can be affected by a variety of factors. The impact of many characteristics of imaging protocols, such as voxel size, tube current, tube voltage, and kernel, has been studied thoroughly [32, 34, 35, 37, 38, 150]. However, the effects of factors intrinsic to the patient have not been investigated. For example, CT scans of the head and neck cover the oral cavity, where many patients have metal dental fillings that cause streak artifacts. As radiomics is based on the assumption that gene expression at a microscopic level is discernible on a macroscopic level in the voxels, it is likely that measuring the radiomics features of the structures affected by a streak artifact would not provide any valuable information about that structure. Another type of artifact observed in CT scans, beam hardening, can affect images containing bone. Because there are many bones in the area of interest in head and neck examinations, this area may be particularly prone to the effects of these small artifacts. As a result, patients whose structure of interest is affected by streak or beam-hardening artifacts are often excluded from the large data sets required to achieve sufficient statistical power for radiomics studies. Therefore, finding a way to include as many patients as possible is needed.

We aimed to test the impact of these artifacts and if needed, methods for compensating for these artifacts in head and neck radiomics studies. First, we determined whether streak artifacts do in fact alter radiomics feature values, and, if so, whether the simple technique of removing the slices affected by the streak artifact produced feature values similar to those in regions unaffected by the artifact. Second, we aimed to determine whether a buffer region is needed between bone and other structures to ensure that the measured feature values are not affected by beam-hardening artifacts.

6.2 Methods and Materials

6.2.1 Streak Artifact

6.2.1.1 Impact of Streak Artifacts on Feature Values

The impact of streak artifacts on feature values was investigated using a cohort of 458 patients with HNSCC. All procedures were performed in accordance with the Declaration of Helsinki on Ethical Issues with a waiver of informed consent from the Institutional Review Board at the University of Texas MD Anderson Cancer Center. Only the patients whose CT images exhibited a visible streak artifact on slices showing the GTV were selected, resulting in the selection of 108 patients. The 108 patient cohort had a mean age of 58 years (range: 30-80 years), mean height of 173 cm (range: 149-191 cm), and mean weight of 77.9 kg (range: 46.0-136.0 kg). In order to evaluate the impact of streak artifacts on the radiomics features (gray-level co-occurrence matrix features, gray-level run length matrix features, neighborhood gray tone difference matrix features, and intensity histogram features), a new ROI was created from which the GTV slices with the streak artifact were removed. Radiomics features were extracted for the 2 ROIs: the original GTV and the modified GTV. A pairwise *t*-test was used to determine if there was a significant ($p < 0.05$) difference in the measured features. This difference in measured features would indicate the streak artifacts having an impact. Streak artifacts were identified manually using a window width of 400 and a window level of -200. An example of a streak artifact is shown in Figure 6-1.

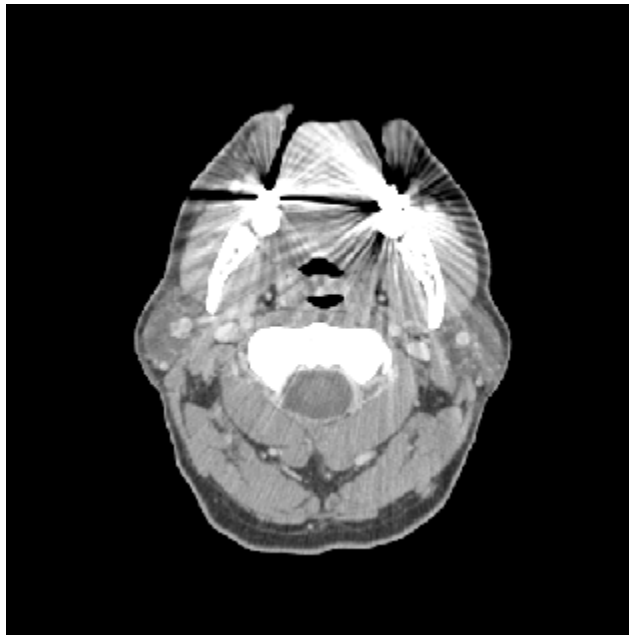


Figure 6-1: Streak Artifact.

Example of identified streak artifact using a window width of 400 and window level of -200.

6.2.1.2 Impact of Removing Slices

A potential approach to removing the effect of streak artifacts on radiomics features is to exclude the affected slices from the feature calculations. In order to study the impact of excluding slices on feature values, we selected CT images for 30 HNSCC patients from the 458 patient cohort with no streak artifacts or with GTVs located far from the artifact and whose imaging parameters were the same. The 30 patients' mean age was 64 years (range: 50-87 years), mean height, 175 cm (range: 149-193 cm), and mean weight, 80.5 kg (range: 43.9-114.9 kg). Slices of the GTV were removed in 2 ways: (1) sequentially, from superior to inferior and (2) in the order given by a random number generator. A new ROI was created for each slice removed from the GTV until only 1 slice remained in the GTV (e.g., if the total GTV was 10 slices, the first ROI would consist of the full 10 slices, the next ROI would contain 9 slices, the next ROI would contain 8 slices, etc.)

To determine the volume that could be removed before feature values changed, we developed a range variation metric based on the range of values across the volumes. First, for each feature and patient, the scaled range, SR , was calculated as

$$SR_i = \frac{range(IF_i)}{FIF_i} \quad (6-1),$$

where i is the patient number (1 to 30), IF_i is the feature value for patient i , and FIF_i is the feature value at full volume for patient i . Next, the scaled range for each patient, SR_i , was used as an input in the range variation metric to determine whether the given feature was robust to the removal of slices from the GTV:

$$Range\ Variation = \frac{mean(SR_i)}{SD(FIF) / mean(FIF)} \quad (6-2),$$

where the numerator is the mean of the scaled ranges and the denominator is the standard deviation of the feature values at full volume across all patients divided by the mean of the feature values at full volume for all patients. The range variation metric thus represents the average effect of removing parts of the tumor divided by the variability in the patient population.

To evaluate the robustness of features to the removal of slices with streak artifacts, we divided the data set into 4 volume range groups. In each group, all contours within a specified volume range were evaluated: 75%-100%, 50%-100%, 25%-100%, and 0%-100%. For example, in the 75%-100% group, all ROIs for a given patient that had at least 75% of the original GTV remaining after removing slices were evaluated. A cutoff of 0.5 was used to determine if features were robust. Features with a range variation above 0.5 were deemed not robust. A pairwise *t*-test was conducted to compare the feature values at 100% GTV and 50% GTV.

6.2.1.3 Feature Extraction

Images were analyzed using IBEX, an open-source radiomics tool [137, 151]. The GTV was contoured on patient images by a radiation oncologist (H. Elhalwani). Twenty-two gray-level co-occurrence matrix features [139], 11 gray-level run length matrix features [140, 141], 5 neighborhood gray tone difference matrix features [142], and 11 intensity histogram features were calculated in IBEX. Each feature was calculated with 4 different preprocessing techniques because different preprocessing techniques have been shown to have different predictive power in non-small cell lung cancer patient survival for individual features [143]; therefore, we tested the following combinations of preprocessing techniques: (1) thresholding, (2) thresholding and Butterworth smoothing (order 2, cutoff 125), (3) thresholding and 8-bit depth resampling, and (4) thresholding, Butterworth smoothing, and 8-bit depth resampling. The lower bound of the threshold was -100 Hounsfield units, and no upper limit was used. The preprocessing and feature group settings were the same as those described in Fave et al.'s supplemental material [14]. These features and preprocessing were chosen as they have been used in prognostic studies and for each feature, at least one of the preprocessing techniques has been shown to be correlated with overall survival, local recurrence, or distant metastases [14]. Illustrations of the effects of each of these preprocessing techniques is demonstrated in Fave et al. [143].

6.2.2 Bone Artifact

6.2.2.1 Phantom Design and Analysis

A modified version of the radiomics phantom described by Mackin et al. [132] was created to analyze the impact of bone artifacts on feature values (Figure 6-2). The phantom contained a cylindrical insert made up of 5 different materials: cork, dense cork, hemp seeds, rubber, and solid water. The hemp seeds were held in a 3D-printed cartridge made of polylactic acid (PLA). There was a central hole with a diameter of 2.2 cm through the insert materials. A rod of solid polyvinyl chloride (PVC) was placed in the hole to simulate bone [152]. A rod of 3D-printed PLA was placed in the hole to simulate water [153, 154]. The feature values obtained with the PLA rod in place were considered the true values, as this rod induced no artifact. These values were used as the reference for comparison of the feature values extracted with the PVC rod, which simulated bone, in place. The phantom was encased in a $28 \times 21 \times 22\text{-cm}^3$ buildup of high-density polystyrene; the size was based on the mean physical dimensions of a European female chest [155].

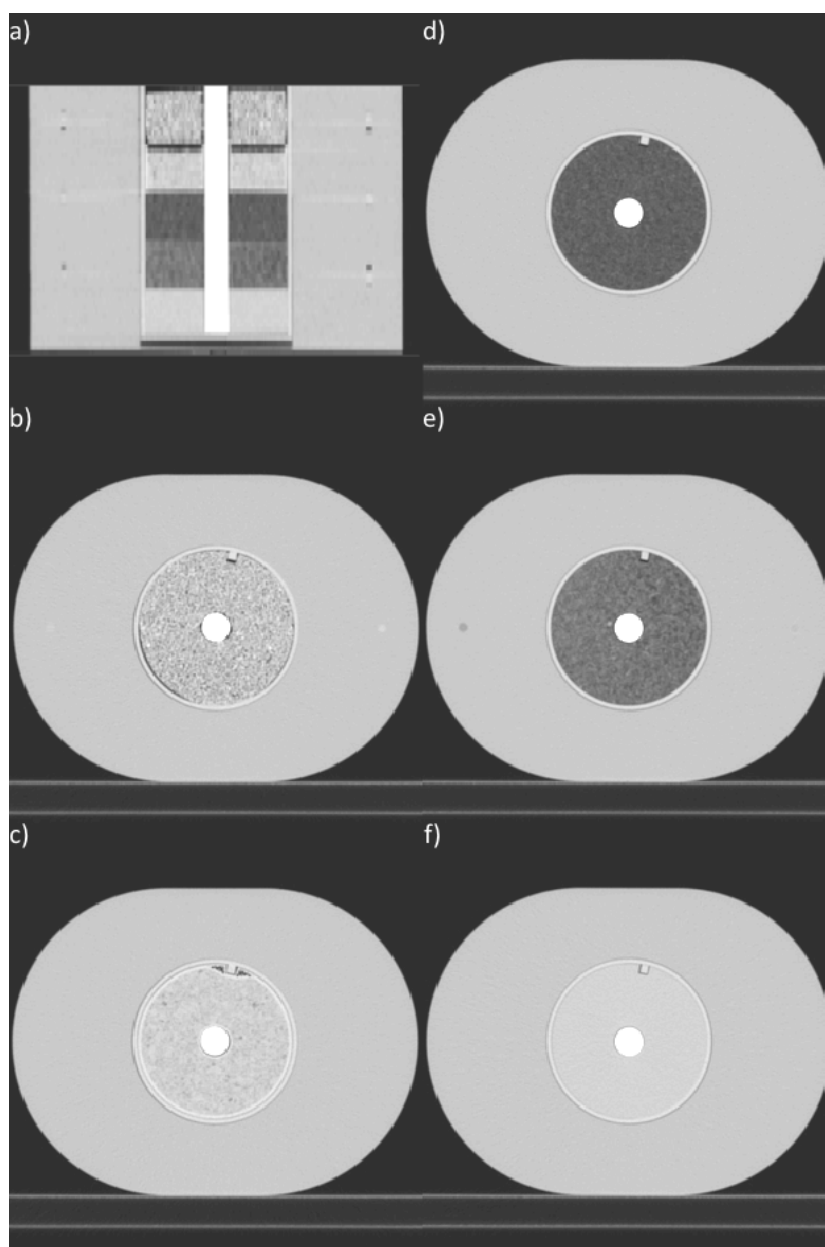


Figure 6-2: CT Images of Bone Phantom.

(a) Coronal slice of the full phantom with the polyvinyl chloride rod in the center of the cartridge materials and high-density polystyrene surrounding the cartridges. The full phantom had dimensions of $28 \times 21 \times 22 \text{ cm}^3$, the mean dimensions of a European female chest. Axial slices showing inserts of (b) rubber, (c) hemp seeds, (d) cork, (e) dense cork, and (f) solid water. Window width: 1600, window level: -300.

The phantom was scanned on a Philips Brilliance Big Bore CT scanner (Philips Healthcare, Eindhoven, The Netherlands). Scans were acquired with the following parameters: 120 kVp, 200 mAs, 0.938 pitch, 1.0 s rotation time, 3-mm image thickness, 0.976 mm × 0.976-mm pixel size, and kernel C. The phantom was scanned 5 times each with the PVC and PLA inserts.

Wilcoxon rank sum tests were used to determine whether the feature values measured with the PLA insert were different from the feature values measured with the PVC insert (i.e., if contours next to bone were affected). For features that were significantly ($p < 0.05$) different, we computed the average absolute difference between the measurement with the PLA insert and the measurement with the PVC insert and compared it to the standard deviation for the feature measured in a cohort of HNSCC patients. The HNSCC patients were the same 30 patients as in the slice-removal study described in section 6.2.1.2.

6.2.2.2 Feature Extraction

The phantom was semiautomatically contoured using the location of a radiopaque marker on the superior edge of the phantom. The marker location was input into an in-house MATLAB (MathWorks) script to create a cylindrical ROI for each material with an outer diameter of 8.2 cm, an inner diameter of 2.2 cm, and height of 2 cm. Additional ROIs for each material were created with inner diameters increasing in 2-mm steps up to 3.4 cm. The same features and preprocessing techniques were used as for the streak artifact study described in section 6.2.1.2; however, for the phantom images, no threshold bounds were applied.

6.3 Results

6.3.1 Streak Artifacts

6.3.1.1 Impact of Streak Artifacts on Feature Values

On average, 3.0 cm³ of GTV had to be removed to eliminate streak artifacts (standard deviation: 4.0 cm³, range: 0.11-28 cm³). Table 6-1 shows the percentage of features for which the measured value in the original GTV (with artifact) and the modified GTV (without artifact) differed significantly. Only for gray-level run length matrix features preprocessed using thresholding and intensity features preprocessed using thresholding, smoothing, and 8-bit depth resampling were fewer than 70% of the features affected by the streak artifact. For all other feature categories and combinations of preprocessing techniques, at least 73% of feature values were affected by the streak artifact.

Table 6-1. Percentage of features with significantly different values with streak artifacts and with artifact slices removed

Feature category	Preprocessing technique			
	Thresholding	Thresholding and 8-bit depth resampling	Thresholding and Butterworth smoothing	Thresholding, smoothing, and 8-bit depth resampling
GLCM (N = 22)	95%	91%	95%	91%
GLRLM (N = 11)	17%	91%	73%	91%
NGTDM (N = 5)	80%	100%	80%	100%
Intensity (N = 11)	82%	82%	73%	64%

GLCM: gray-level co-occurrence matrix, GLRLM: gray-level run length matrix, NGTDM: neighborhood gray tone difference matrix

6.3.1.2 Impact of Removing Slices

The number of features that were not robust across the volume range for each slice removal and preprocessing technique is shown in Table 6-2. Almost all features were robust with removal of up to 50% of the original GTV. When modified GTVs were allowed to go down to only 1 slice, almost no features were robust. The features that were not robust are listed in the Appendix C Table C-1. The range variation for each feature and preprocessing for the four volume groupings is contained in spreadsheets in the Supplemental Material of Ger et al. [156].

Table 6-2. Number of features that were not robust across the volume range for each slice removal and preprocessing technique

ROI Volume Range	Sequential Slice Removal				Random Slice Removal			
	Thresholding	Thresholding and 8-bit depth resampling	Thresholding and Butterworth smoothing	Thresholding, smoothing, and 8-bit depth resampling	Thresholding	Thresholding and 8-bit depth resampling	Thresholding and Butterworth smoothing	Thresholding, smoothing, and 8-bit depth resampling
75% - 100%	0	1	0	0	0	0	0	0
50% - 100%	4	3	4	3	2	2	2	1
25% - 100%	20	16	22	20	9	5	9	5
0% - 100%	48	48	48	48	47	44	47	46

ROI, region of interest

The pairwise *t*-test between features at 100% GTV and 50% GTV showed that 76% of features under the sequential slice removal technique and 80% of features under the random slice removal technique were not significantly ($p < 0.05$) different.

6.3.2 Bone Artifact

Table 6-3 shows the number of features that significantly differed in the phantom with a PLA or PVC rod. Adding Butterworth smoothing and 8-bit depth resampling had no consistent effect on the differences. For example, adding smoothing reduced the number of significantly different gray-level co-occurrence matrix features in cork but increased the number in rubber. For the features with significant differences, the average difference divided by the standard deviation across the 30 HNSCC patients was 0.57 (range: 0.0044-3.6) for cork, 0.28 (range: 0.0011-1.7) for dense cork, 2.0 (range: 0.0083-36) for hemp seeds, 0.14 (range: 0.0014-0.93) for rubber, and 3.0 (range: 0.0031-36) for solid water. The results were similar for the various ROIs with different inner diameters.

Table 6-3. Number of features with significantly different values when measured in a phantom with a PLA or PVC rod

Feature	Preprocessing technique	Cork	Dense Cork	Hemp Seeds	Rubber	Solid Water
GLCM (N=22)	Thresholding	12	5	7	1	13
	Thresholding, smoothing	4	4	18	11	15
	Thresholding, 8-bit depth resampling	11	7	6	1	12
	Thresholding, smoothing, 8-bit depth resampling	4	3	18	10	13
GLRLM (N=11)	Thresholding	7	6	6	3	6
	Thresholding, smoothing	7	6	7	7	10
	Thresholding, 8-bit depth resampling	5	5	4	3	10
	Thresholding, smoothing, 8-bit depth resampling	5	5	8	5	6
NGTDM (N=5)	Thresholding	0	0	0	0	3
	Thresholding, smoothing	0	4	1	0	1
	Thresholding, 8-bit depth resampling	0	2	0	0	1
	Thresholding, smoothing, 8-bit depth resampling	0	1	1	0	3
Intensity (N=11)	Thresholding	8	3	7	3	8
	Thresholding, smoothing	5	5	11	8	9
	Thresholding, 8-bit depth resampling	7	4	5	3	7
	Thresholding, smoothing, 8-bit depth resampling	6	5	9	8	9

GLCM: gray-level co-occurrence matrix, GLRLM: gray-level run length matrix, NGTDM: neighborhood gray tone difference matrix

6.4 Discussion

In this study, we showed that streak artifacts affect radiomics feature values, suggesting that regions containing such artifacts should not be included in radiomics data sets. We demonstrated that a simple technique, removing the slices with the artifact, can be used to remove up to 50% of the original GTV from the ROI while retaining similar feature values. Additionally, while the presence of bone within the image can affect some feature values, the effect is typically smaller than the spread in values in the patient population and can, therefore, be ignored.

The choice of 0.5 as the cutoff of the range variation metric to determine feature robustness was arbitrary. However, we found that using different cutoffs did not change the conclusions: feature values were generally robust when up to 50% of the GTV was removed, but having only a small fraction (less than 50%) of the GTV remaining in the ROI caused very large differences in feature values. Additionally, a 50% cutoff for robustness means that only a few patients would be lost from the data set; only 15 (3%) of 458 HNSCC patients had artifact on more than 50% of the GTV.

The majority of feature values were not significantly affected by the PVC rod simulating bone in the phantom. When features were affected, the effect was typically smaller than the standard deviation in values from the patient population. Interestingly, we found that the mean HU value was always significantly higher when the PVC rod was in the phantom than when the PLA rod was in the phantom, whereas the median HU was only higher for the PVC rod than the PLA rod when 8-bit depth resampling was not used. The presence of PVC caused this difference no matter the distance measured from the interface of the cartridge material and PVC. However, the average difference was typically one-tenth of the standard deviation of the value from the patient cohort for these features. Therefore, the presence of bone may cause changes in feature values, but the proximity of contours does not.

A method similar to that used for determining the impact of bone artifacts could have also been used for the streak artifacts. For example, a phantom could have been scanned with and without a metal bb inside of it. Since a phantom only simulates patient tissues and is not a true match and we

had a large sample of patients with streak artifacts within their GTV, we chose to analyze the patient set instead of a phantom. This gave us the most realistic scenario for implementation on patient data sets. We did not have a large sample for the bone artifacts and thus chose to study the impact within a phantom.

There were several limitations to this study. First, only one technique was investigated for dealing with streak artifacts. Metal artifact reduction techniques may also be able to solve the problem of streak artifacts and may be able to do so without having to exclude any patients. Second, the bone artifact study was conducted in a phantom. While the rubber and cork cartridges have been shown to produce feature values similar to those produced by non-small cell lung cancer tumors [132], the size and shape of the phantom simulated the dimensions of a human chest, not a head. A potential avenue for future investigation could be to create an anthropomorphic head and neck phantom using real human bone. This would allow us to investigate whether features are more susceptible to beam-hardening artifacts right inside the skull than near a cylinder of bone-like material. Additionally, many of these features have been found to have some correlation with volume [143]. This may contribute to features becoming less robust as more volume was removed. Finally, both of these studies were conducted on CT scans from a single vendor so the results may not be universally applicable.

6.5 Conclusion

We demonstrated that streak artifacts affect the measured radiomics feature values. In order to deal with this effect, we suggest simply removing the slices with the artifact. Using this method, feature values are robust when up to 50% of the original GTV is removed. Excluding patients in whom more than 50% of the GTV is affected by the artifact only causes about 3% of patients to be excluded. Additionally, we demonstrated that contours can abut bone if needed, as most features are not affected by the presence of nearby bone.

Chapter 7 : Inter-Scanner Variability of Radiomics Features on Computed Tomography Scanners

This chapter is based upon:

Ger RB, Zhou, S, Chi, PM, Lee, HJ, Layman, RR, Jones, AK, Goff, DL, Fuller, CD, Howell, RM, Li, H, Stafford, RJ, Court, LE, Mackin DS. Comprehensive Investigation on Controlling for CT Imaging Variabilities in Radiomics Studies. *Scientific Reports* doi: 10.1038/s41598-018-31509-z. Volume 8, Pages 13047. ©Nature Publishing Group.

This article is under a Creative Commons license (<http://creativecommons.org/licenses/by/4.0/>) which permits reproduction in any format.

7.1 Introduction

Many radiomics studies are conducted at one facility. However, as the field of radiomics has grown, researchers have sought larger patient cohorts by combining data from multiple facilities. This means that patients are scanned on different CT scanners using different protocols, which may affect radiomics features [157]. Due to this there have been studies analyzing the uncertainty due to different protocol parameters. These uncertainty studies often involve only a few scanners at one facility, which provides valuable information about imaging variability, but these results may not be generalizable to a larger population of CT scanners at multiple facilities. Mackin et al. created a radiomics phantom to investigate the imaging variability among 17 scanners using the routine chest protocol on each [132]. They found that radiomics feature value differences due to the different scanners were similar to the inter-patient radiomics feature variability among NSCLC patients and thus recommended that these imaging differences be considered in future studies.

In this study, we aimed to obtain a large sample of CT scanners for an in-depth analysis of imaging variability to determine how retrospective radiomics studies should select patients and how

prospective radiomics studies should design CT protocols. The large sample would allow for the conclusions to be applied generally to all CT scanners. Local protocols were used, as many studies use retrospective data and it is of interest whether protocol differences will cause large radiomics feature value differences, thus causing patient stratification to be dominated by scan protocol and not true patient radiomics feature values. Also, a controlled scan was used to see whether imaging differences could be minimized using a harmonized protocol across different vendors.

7.2 Methods

7.2.1 Methods and Materials

We used an updated version of the Credence Cartridge Radiomics phantom originally described by Mackin et al. [132] in 2015. This version of the phantom, shown in Figure 7-1, is comprised of six round cartridges encased in high-density polystyrene buildup. The six cartridges were held within the buildup in an acrylic case with a notch designed to keep the cartridges in the same position. This case can be seen in Figure 7-1 as the bright line around each cartridge before the buildup. The size of the buildup, 28 cm × 21 cm × 22 cm, is based on the mean physical dimensions of a European woman's chest [155]. The six cartridges are each comprised of different materials: 50% acrylonitrile butadiene styrene (ABS), 25% acrylic beads, and 25% PVC pieces (percentages are by weight); 50% ABS and 50% PVC pieces; 50% ABS and 50% acrylic beads; hemp seeds encased in polyurethane; shredded rubber; and dense cork. These materials were chosen to produce a range of radiomics feature values similar to those of NSCLC tumors for the original materials [132], the new materials followed the same analysis as the original materials. Additional details on the differences between this phantom and the original phantom are described in the Discussion.

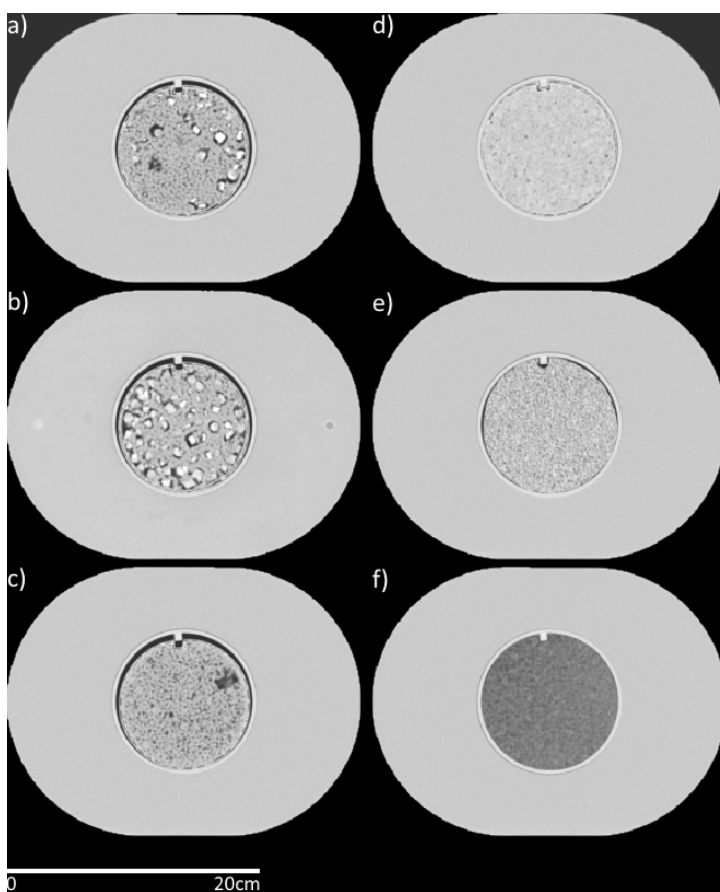


Figure 7-1: Radiomics Phantom used for Inter-Scanner Analysis.

Axial views from a computed tomography scan of the radiomics phantom used. The cartridges are (a) 50% acrylonitrile butadiene styrene (ABS), 25% acrylic beads, and 25% polyvinyl chloride (PVC) pieces (percentages are by weight), (b) 50% ABS and 50% PVC pieces, (c) 50% ABS and 50% acrylic beads, (d) hemp seeds in polyurethane, (e) shredded rubber, and (f) dense cork. The high-density polystyrene buildup is seen outside the cartridges with dimensions of 28 cm × 21 cm × 22 cm. The cartridges had a diameter of 10.8 cm. Window width: 1600, window level: −300.

7.2.2 CT Scans

A controlled CT scan was acquired using the following parameters for each scanner: tube voltage, 120 kV(p); tube current, 200 mA·s; helical scan type; spiral pitch factor, 1.0; 50-cm display field of view; and image thickness, 3 mm (except for GE scanners, which used an image thickness of 2.5 mm). The acquisition parameters were designed to give about 13 mGy CTDI_{vol} (average 16 mGy, standard deviation 4 mGy) in order to produce the same noise characteristics. A recent study by Mackin et al. [38] showed that features were not affected by noise levels in the image, thus variations in CTDI_{vol} should not impact the radiomics features. The convolution kernel was standard for GE; C for Philips; B31f, B31s for Siemens; and FC08 for Toshiba. These kernels were chosen to minimize the difference in radiomics feature values across vendors as described in Mackin et al.'s abstract [158]. Also, the local chest protocol and local head and neck protocol were used to acquire scans of the phantom. For the local protocols, no parameters were changed in order to estimate the variability in protocols across institutions and scanners. The parameters for each of the local protocol scans is supplied in the Supplemental Material of Ger et al. [136].

7.2.3 Patient Scans

A phantom alone cannot provide insight into the impact of feature variability within a patient study. Thus, we have included patient cohorts to determine the size of the imaging variability with respect to inter-patient variability, providing an estimate on the impact of the imaging variability for each feature.

For this study, we retrospectively reviewed the images and medical records of 20 patients with NSCLC and 30 patients with HNSCC with a waiver of informed consent from the Institutional Review Board at the University of Texas MD Anderson Cancer Center. These two cohorts of patients were used to compare the imaging variability to inter-patient variability. Radiomics features have been shown to improve the patient outcome models for both of these patient types [15, 17, 20, 22, 29].

The NSCLC cohort had 10 men and 10 women, mean age of 67 years (range, 52-78 years), mean weight of 72.9 kg (range, 41.0-97.6 kg), and mean height of 170 cm (range, 154-182 cm). The CT scans were acquired on a GE Discovery CT scanner (GE Healthcare) at 120 kVp, 300 mA, 0.5 s rotation time, 2.5-mm image thickness, 1.35 pitch, and 0.976 mm × 0.976 mm pixel size.

The HNSCC cohort had 25 men and 5 women, mean age of 64 years (range, 50-87 years), mean weight of 80.5 kg (range, 43.9-114.9 kg), and mean height of 175 cm (range, 149-193 cm). The CT scans were acquired using a GE LightSpeed CT scanner (GE Healthcare) at 120 kVp, 220 mA, 1.0 s rotation time, 1.25-mm image thickness, 1.375 pitch, and 0.488 mm × 0.488 mm pixel size. For both patient cohorts, the tumors were contoured by a radiation oncologist.

7.2.4 Radiomics Feature Extraction

The phantom was semi-automatically contoured using an in-house MATLAB (version 2016b, MathWorks) script. A cylindrical ROI was created for each cartridge. Each ROI was 8.2 cm in diameter. The ROIs for the cartridge with 50% ABS and 50% acrylic beads and the cartridge with hemp seeds in polyurethane each had a height of 1.9 cm. All other ROIs each had a height of 2 cm. Mackin et al. showed that the size of the ROI did not impact conclusions of a phantom study [38], therefore we maximized the acceptable region within each cartridge. The ROIs were automatically placed into IBEX, an open-source radiomics tool [137, 151], and then viewed to determine acceptability. Generated contours were scrutinized and edited as needed.

Forty-nine features were calculated using IBEX: 22 gray level co-occurrence matrix features [139], 11 gray level run length matrix features [140, 141], 11 intensity histogram features, and five neighborhood gray tone difference matrix features [142] (Table 7-1). Four different preprocessing techniques were used for each feature: (1) thresholding; (2) thresholding and 8-bit depth resampling; (3) thresholding and a Butterworth smoothing filter (order of 2, cut-off of 125); and (4) thresholding, 8-bit depth resampling, and Butterworth smoothing [143]. The thresholds for the NSCLC patient cohort

were a lower threshold of –100 HU and a higher threshold of 200 HU. A lower threshold of –100 HU was used for the HNSCC patient cohort with no upper threshold. No thresholding was applied to the phantom images. The settings for each feature were the same as those listed by Fave et al. in the Supplemental Material [14]. For the local scans, the pixel size was resampled to 1 mm × 1 mm using trilinear interpolation as suggested by the results from Mackin et al. [37]. For features that have been previously found to correlate with volume, the updated formulae were used as described by Fave et al. [143].

Table 7-1. Radiomics Features Analyzed

Gray Level Co-occurrence Matrix	Gray Level Run Length Matrix	Intensity Histogram	Neighborhood Gray Tone Difference Matrix
Auto Correlation	Gray Level Nonuniformity	Energy	Busyness
Cluster Prominence*	High Gray Level Run Emphasis	Entropy	Coarseness
Cluster Shade*	Long Run Emphasis	Kurtosis	Complexity
Cluster Tendency	Long Run High Gray Level Emphasis	Maximum	Contrast
Contrast	Long Run Low Gray Level Emphasis	Mean	Texture Strength
Correlation	Low Gray Level Run Emphasis	Median	
Difference Entropy	Run Length Nonuniformity	Minimum	
Dissimilarity	Run Percentage	Skewness*	
Energy	Short Run Emphasis	Standard Deviation	
Entropy	Short Run High Gray Level Emphasis	Uniformity	
Homogeneity	Short Run Low Gray Level Emphasis	Variance	
Homogeneity 2			
Information Measure			
Correlation 1			
Information Measure			
Correlation 2			
Inverse Difference			
Moment Norm			
Inverse Difference			
Norm			
Inverse Variance			
Max Probability			
Sum Average			
Sum Entropy			
Sum Variance			
Variance			

*Indicates features that were subsequently not used due to sensitivity of region of interest placement within the phantom material

7.2.5 Statistical Methods

7.2.5.1 Feature Stability

The features were tested for reproducibility by moving the ROIs on one controlled scan of the phantom. The ROIs were shifted 10 times within the acceptable region of the cartridges. The coefficient of variation was calculated for each feature. Features for which more than 50% of instances (with four preprocessing types and six cartridges, there were 24 total instances for each feature) had a coefficient of variation above 10% were removed from further analysis. It was important to remove these features as features that are very sensitive to the positioning of the ROI may not properly represent the imaging variation and may only represent placement of the ROI on the different scans.

7.2.5.2 Resampling the z Dimension

For the local protocol scans, the image thickness was not consistent. The impact of the image thickness on feature value was evaluated by computing the Pearson correlation for each ROI-feature combination. Additionally, the impact of resampling the image thickness was investigated by resampling the z dimension from 1 mm to 7 mm in 1 mm increments. Features were acquired using all z dimension resampling values and without resampling the z dimension. The intra-class correlation coefficient (ICC) was computed for each feature using the eight resampling options to determine if resampling changed the feature values and thus reduced the correlation of feature values with image thickness. The ICC (2,1) (two-way random effects, absolute agreement, single rater/measurement) and ICC (3,1) (two-way random effects, consistency, single rater/measurement) as described by Shrout and Fleiss [159] were computed in R (version 3.4.3) using the psych package (version 1.7.8) [160]. For these tests, features were calculated with thresholding preprocessing on the local chest protocol scans. The other preprocessing techniques and the head protocol scans were not used as this step was simply to determine the relationship between image thickness and feature values, and the additional preprocessing and protocol scans produced redundant data.

7.2.5.3 Imaging Variability

Our goal was to determine how the manufacturer and scanner uncertainties contribute to the overall variability in the feature values. To determine these uncertainties, we first built a linear mixed-effects model, which estimates the contribution of the manufacturer, the additional scanner-wise variability within a given manufacturer, the cartridge material, and the residual to the variability in the measurements. The standard deviations of the distributions are used to provide estimates of the variability contributed from the manufacturer, scanner, cartridge material, and residual. The term scanner is used here to indicate an individual scanner (e.g., multiple of the same type of scanner from the same manufacturer are each considered distinct). There are many factors that could affect the images from a particular scanner, including the quality assurance (QA) technique/periodicity, scanner maintenance, and scanner design. Thus, radiomics features calculated from images taken using CT scanners of the same manufacturer/model may be different. The term residual typically implies a small contribution. However, for this study the term is simply used to represent anything that is not included within the formula (i.e., anything that is unknown).

A linear mixed-effects model was created for each scan type (controlled, local chest, and local head and neck protocol):

$$f_{m,i} = \mu + \alpha_m + \beta_i + g(t) + \varepsilon_{m,i} \quad (7-1),$$

where f is the feature, μ is the mean, m is the cartridge material, i is the scanner, α is the material-wise contribution, β is the scanner-wise contribution, $g(t)$ is the fixed effect of the impact of image thickness on feature value, and ε is the residual. β_i is normally distributed with a mean of $\gamma_{v,i}$ and a variance of $\sigma_{\beta,m}^2$ ($\sigma_{\beta,m}^2 = \sigma_{\beta}^2 \times \hat{f}_m^2$). $\gamma_{v,i}$ is the vendor-wise contribution which is normally distributed with a mean of 0 and a variance of $\sigma_{\gamma,m}^2$ ($\sigma_{\gamma,m}^2 = \sigma_{\gamma}^2 \times \hat{f}_m^2$). \hat{f}_m is the mean feature value for the cartridge material. $\varepsilon_{m,i}$ is normally distributed with a mean of 0 and variance of $\sigma_{\varepsilon,m}^2$ ($\sigma_{\varepsilon,m}^2 = \sigma_{\varepsilon}^2 \times \hat{f}_m^2$). The model computes a significance test before producing the results. If the standard deviation due to one component is much

smaller than the others, it is set to 0 and combined into the residual. The linear mixed-effects models were computed in R (version 3.4.3) using the lme4 package (version 1.1-17).

Imaging variability was measured using the uncertainties from the linear mixed-effects models. Currently, most studies do not apply corrections for the manufacturer and scanner. The total imaging variability was calculated to estimate the impact of continuing to not apply corrections. It was calculated as follows:

$$IV_{total} = \frac{\sigma_{t,m}/\widehat{f_m}}{\sigma_p/\mu_p} \quad (7-2),$$

where σ_p is the standard deviation of the feature value for patients, μ_p is the mean feature value for patients, and $\sigma_{t,m}$ is the total standard deviation from the model, given by

$$\sigma_{t,m} = \sqrt{\sigma_{\beta,m}^2 + \sigma_{\gamma,m}^2 + \sigma_{\varepsilon,m}^2} \quad (7-3).$$

This metric (equation 7-2) includes a comparison to the patients to gauge the impact of the imaging variability in a patient setting.

The residual imaging variability was calculated to estimate the imaging variability that would exist in cohorts that include CT images from different scanners even if corrections could be applied based on the manufacturer and individual scanner, as follows:

$$IV_{residual} = \frac{\sigma_{\varepsilon,m}/\widehat{f_m}}{\sigma_p/\mu_p} \quad (7-4).$$

We repeated this modeling process for the three scan types (controlled, local chest, and local head and neck protocols) and compared the results. To determine if the controlled scan significantly reduced the variability, we performed one-sided pairwise t-tests comparing σ_{β} , σ_{γ} , and σ_{ε} between the controlled protocol and both local protocols.

7.2.5.4 Quality Assurance Using a Radiomics Phantom

The feasibility of creating a credentialing phantom for radiomics studies, similar to the credentialing of institutions for National Institutes of Health radiation therapy studies, was investigated. Ideally, the credentialing phantom would be small for ease of transport and use. Therefore, the ability of each cartridge was tested for its use in QA checks to determine which CT scanners do not fall within the credentialed standard population of scanners. The spread of feature values from different scanners should be small relative to the inter-patient spread, therefore, the patient standard deviations were used to determine if scanners fell close enough to the population scanner value or not. The controlled scans were used for this analysis. For each feature, the patient standard deviation was scaled to account for differences in means between the patient and phantom populations.

$$\sigma_{scaled} = \frac{\sigma_p}{\mu_p} \times \hat{f} \quad (7-5)$$

For each scanner, the number of features that fell outside 1/3 of the scaled patient standard deviation from the mean feature value was tallied. The idea of the bounds was to determine if criteria could be established such that a certain number of features would fall within the bounds in order for the given scanner to pass the QA test. Therefore, the bounds were set as follows:

$$Lower\ bound = \hat{f} - \frac{1}{3} \sigma_{scaled} \quad (7-6)$$

$$Upper\ bound = \hat{f} + \frac{1}{3} \sigma_{scaled} \quad (7-7)$$

7.3 Results

7.3.1 Scanners

The phantom was scanned on 100 scanners: 51 GE scanners (GE Healthcare), 20 Philips scanners (Philips Healthcare), 17 Siemens scanners (Siemens Healthineers, Erlangen, Germany), 11 Toshiba scanners (Canon Medical Systems USA, Tustin, CA, USA), and one Philips and Neusoft Medical System scanner (Shenyang, China). Ninety-four scanners had a controlled protocol scan that could be

used: 48 GE, 18 Philips, 17 Siemens, and 11 Toshiba scanners. However, the kernel used for the Toshiba scans switched from FC18 (six scanners) to FC08 (five scanners) halfway through owing to a study that found the FC08 kernel to match the GE standard kernel best [158]. To determine whether both Toshiba kernels could be used in the analysis, k-means clustering was performed. The scanners did not cluster by kernel type. While the best match should always be used to minimize discrepancies, in this study the kernel differences among the Toshiba scanners was not a driving force in the variability and therefore, kernel did not matter for Toshiba and all Toshiba scans were included in the analysis. Ninety-three scanners had a local chest protocol scan that could be used: 47 GE, 19 Philips, 17 Siemens, and 10 Toshiba scanners. Eighty-eight scanners had a local head protocol scan that could be used: 46 GE, 18 Philips, 14 Siemens, and 10 Toshiba scanners. The various reasons that scans could not be used were as follows: the field of view did not encompass all the cartridges, the scan extent did not cover the length of the phantom, and the scan was acquired with variable image thickness. Head and neck protocols could be acquired only on CT scanners used for radiation therapy purposes; on diagnostic scanners, a head scan, typically brain, was acquired (both head and neck and head protocols are referred to as “head protocols” hereafter).

We were able to ascertain that at least 96% of scanners followed AAPM or ACR recommendations for QA. Additionally, at least 49% of scanners were ACR accredited, 20% of scanners were in the radiation therapy department of scanners at ACR accredited facilities, and 6% of scanners were currently undergoing ACR accreditation.

The local chest protocol scans had image thicknesses ranging from 1 to 5 mm. The local head protocol scans had image thicknesses ranging from 0.5 to 5 mm. Histograms of the distributions are shown in Figure 7-2.

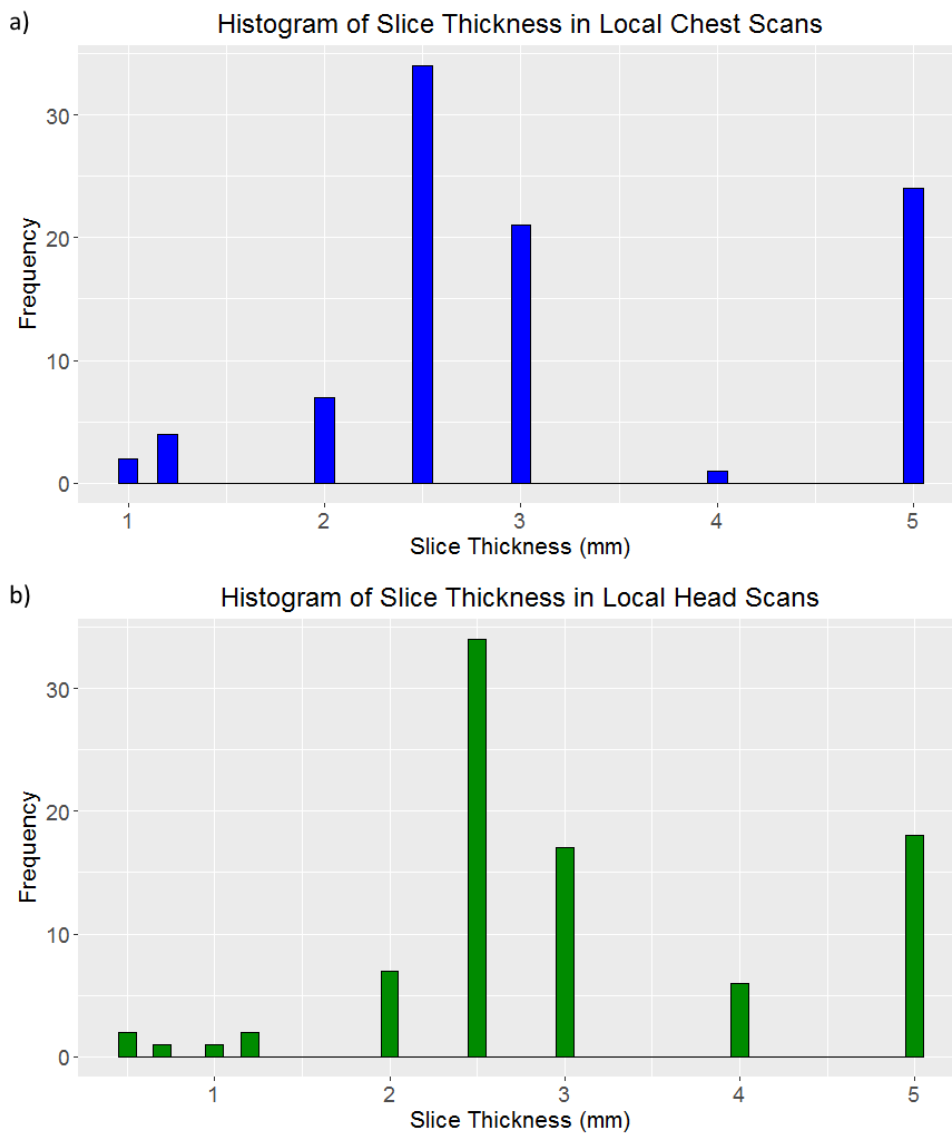


Figure 7-2: Image Thickness Histograms.

Histograms of image thicknesses across the scans taken using (a) the local chest protocol and (b) the local head protocol.

7.3.2 Feature Stability

Three features had a coefficient of variation greater than 10% in more than 50% of instances (with 24 total instances for each feature): the features of cluster prominence, cluster shade, and skewness. These features were not included in subsequent analysis. The coefficient of variation exceeded 10% for auto correlation and sum variance in 42% of instances and for long run low gray level emphasis, low gray level run emphasis, short run low gray level emphasis, and the minimum in 46% of instances. All other features had a coefficient of variation greater than 10% in less than 25% of instances; the majority of features had a coefficient of variation greater than 10% in 0% of instances.

7.3.3 Resample the z Dimension

Figure 7-3 shows the absolute value of the Pearson correlation coefficient of each ROI for the correlation of each feature with the image thickness. The mean absolute value of the Pearson correlation coefficient was 0.42. The correlation values had similar ranges for all the feature categories except for the gray level run length matrix category, which had lower correlation values. The mean absolute value of the Pearson correlation coefficient increased to 0.46 when gray level run length matrix features were not included. A second version of Figure 7-3 without the ABS cartridges is reproduced in Appendix D Figure D-1. For this analysis the mean absolute value of the Pearson correlation coefficient was 0.39. Without the gray level run length matrix features, the mean absolute value of the Pearson correlation coefficient was 0.41.

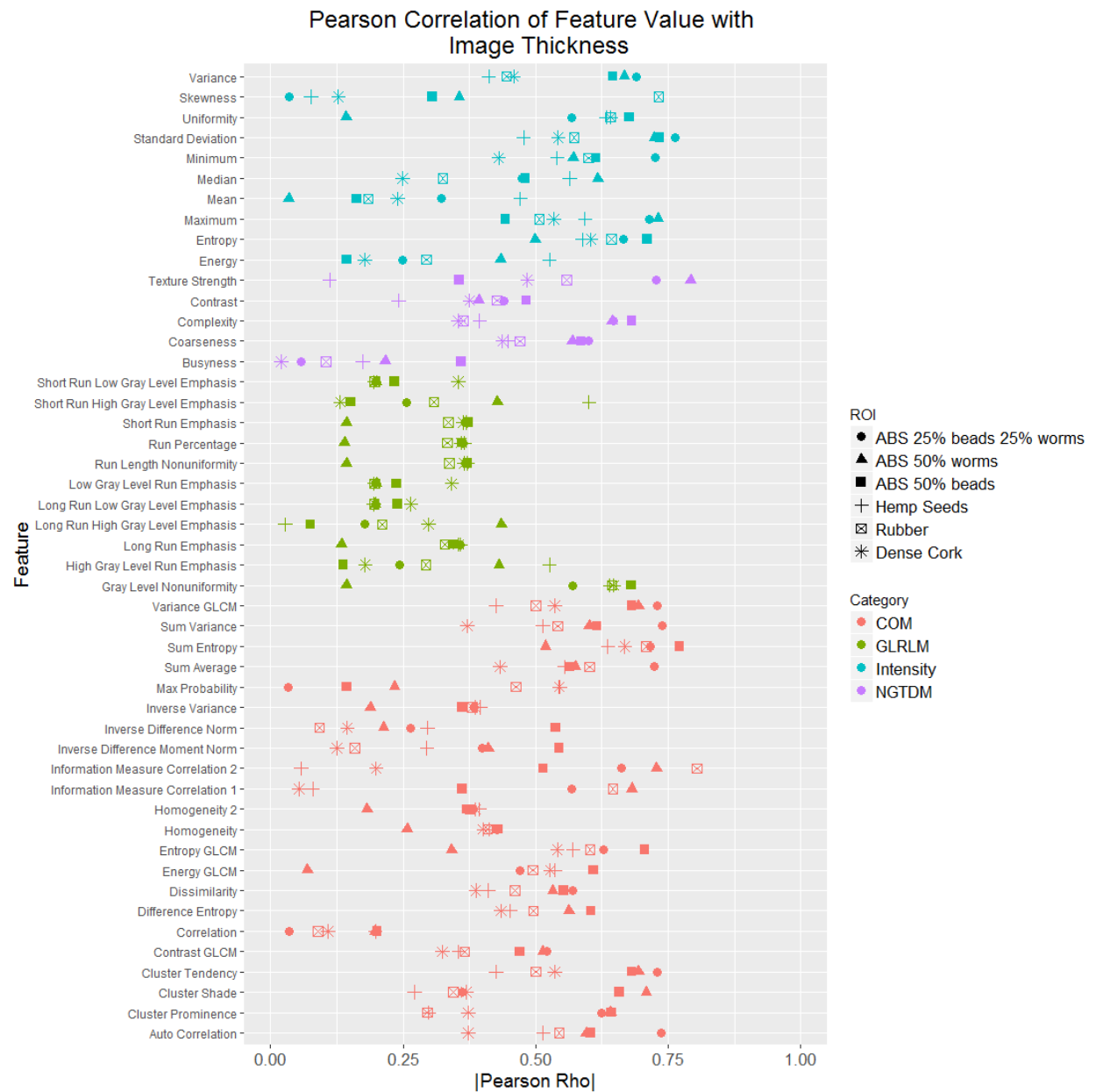


Figure 7-3: Pearson Correlation of Feature Value with Image Thickness.

Absolute value of the Pearson correlation rho for the correlation between feature value and image thickness for each region of interest (ROI). Each ROI is a different shape. Each category of feature is a different color. The correlation varies between and within features depending on the ROI. COM: gray level co-occurrence matrix, GLCM: gray level co-occurrence (used when there are features with the same name in different categories to differentiate them), GLRLM: gray level run length matrix, NGTDM: neighborhood gray tone difference matrix, beads: acrylic beads, worms: polyvinyl chloride pieces.

To determine the level of reliability based on the ICC values, the guidelines from Koo and Li were followed [161]. ICC values less than 0.5 signify poor reliability, those between 0.5 and 0.75 signify moderate reliability, those between 0.75 and 0.9 signify good reliability, and those greater than 0.9 signify excellent reliability. When comparing feature values across different resampling techniques using ICC (2,1) (two-way random effects, absolute agreement, single rater/measurement), we found that 35 features had excellent reliability, seven features had good reliability (entropy, max probability, low gray level run emphasis, short run low gray level run emphasis, busyness, complexity, and contrast), and four features had moderate reliability (information measure correlation 1, information measure correlation 2, long run low gray level emphasis, and texture strength). When ICC (3,1) (two-way random effects, consistency, single rater/measurement) was used, we found that 39 features had excellent reliability, five features had good reliability (information measure correlation 2, max probability, low gray level run emphasis, short run low gray level run emphasis, and texture strength), one feature had moderate reliability (long run low gray level emphasis). Thus, feature values did not change with resampling; therefore, for the linear mixed-effects analysis, no resampling in the z dimension was done for the local chest and local head protocols. Additionally, these results paired with the Pearson correlation results implied that there was a relationship with image thickness that needed to be included in the modeling.

7.3.4 Imaging Variability

The variability due to the material was 0 in every model. The relative proportions of σ_B (scanner-wise variability), σ_V (manufacturer-wise variability), and σ_ϵ (residual variability) were calculated for each feature. Plots of the proportion of each of these variabilities using thresholding and bit depth rescaling are shown in Figure 7-4 for the controlled protocol and local head protocol. All other plots (other preprocessing and chest protocol) are in Appendix D Figures D-2 through D-11. Figure 7-4 shows that the contribution from σ_V is reduced when the controlled protocol is used. The mean total variability

for the controlled protocol was 0.43 compared with that of the local chest protocol and was 0.48 compared with that of the local head protocol. The average proportion of total variability was 0.29, 0.27, and 0.43 for the manufacturer, scanner, and residual respectively based on the head protocol scans. The average proportion of total variability was 0.30, 0.27, and 0.44 for the manufacturer, scanner, and residual respectively based on the chest protocol scans. The average proportion of total variability was 0.20, 0.25, and 0.55 for the manufacturer, scanner, and residual respectively based on the controlled protocol scans. The details of this are shown in Figure 7-4.

The residual contribution was not always small; it was often the largest component. This is particularly evident for the controlled protocol where the residual should have a large relative contribution since factors that were contributing to the variability have been accounted for in the design of the protocol. The manufacturer contribution was not always larger than the scanner contribution to the total variability, as can be seen in Figure 7-4, thus demonstrating that the variability among scanners of the same manufacturer can vary more than different manufacturers.

If it was possible to correct for the manufacturer and individual scanner, then, when using a controlled protocol, only the residual variability would remain. In that situation, the mean controlled residual variability would be 0.36 compared with the chest protocol total variability and 0.40 compared with the head protocol total variability. This is the theoretical best possible improvement that can be achieved until we have an in-depth understanding of the components hidden in the residual. In comparison to the controlled protocol, this is an additional 7-8% reduction in variability ($100 \times$

$$\left(\text{mean} \left(\frac{\text{total variability controlled protocol}}{\text{total variability local protocol}} \right) - \text{mean} \left(\frac{\text{residual variability controlled protocol}}{\text{total variability local protocol}} \right) \right).$$

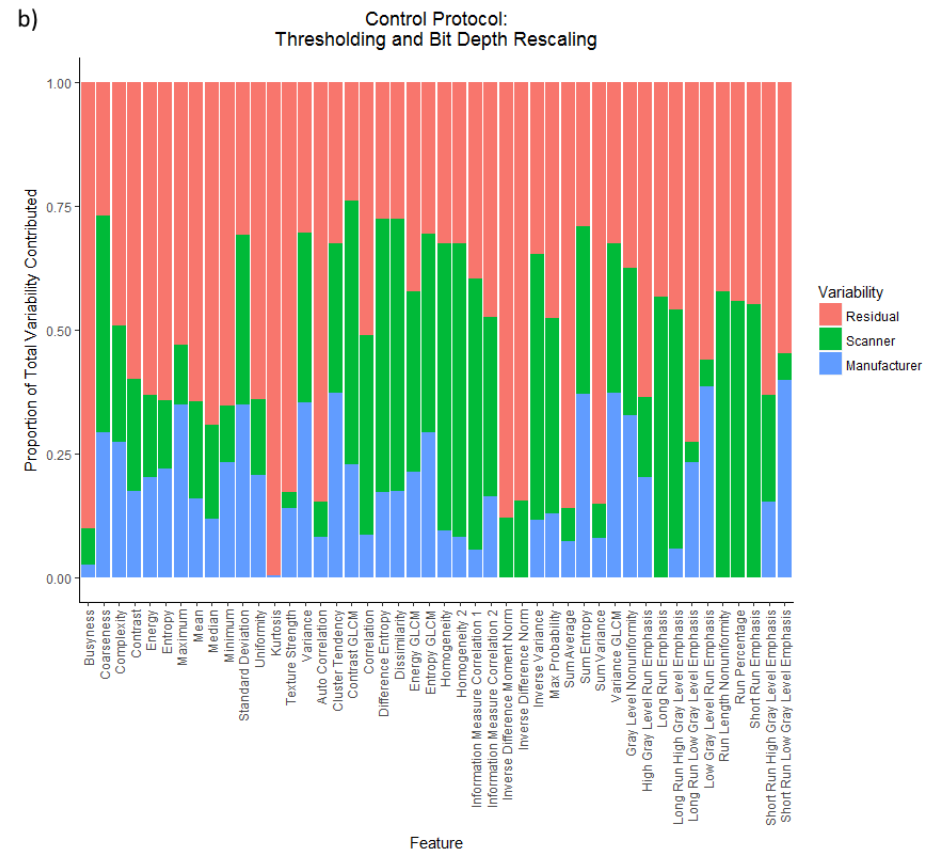
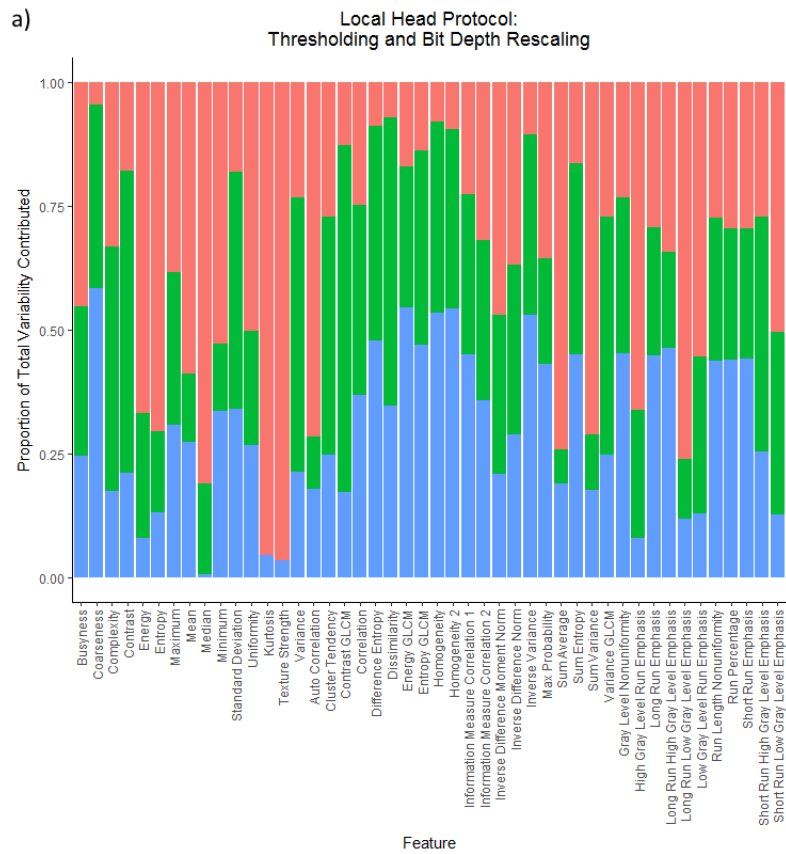


Figure 7-4: Bar Plots of Variation from Head and Controlled Protocols.

Bar plots of the relative contributions of the scanner-wise variability (green), manufacturer-wise variability (blue), and residual variability (red) for each feature using thresholding and bit depth rescaling calculated on (a) the local head protocol and (b) the controlled protocol. The contribution of the manufacturer was much larger for many features in the local head protocol than in the controlled protocol. The total variability for the controlled protocol compared with that of the head protocol was 0.48.

The linear mixed-effects models produced a spectrum of variabilities, from high to low. For ease of summary, a cutoff has been established. Spreadsheets with the data are in the Supplemental Material of Ger et al. to allow for different cutoffs to be used in future studies [136]. For IV_{total} and $IV_{residual}$ (equations 7-2 and 7-4), a cutoff of 1/3 was used to create a binary of significance (i.e. significant or not). This was done for each feature to indicate that the imaging variation was negligible relative to inter-patient variability or imaging variability was significant relative to inter-patient variability. The total numbers of features in each category that had IV_{total} or $IV_{residual}$ values greater than 1/3 are displayed in Table 7-2.

Two gray level run length matrix features and one intensity feature were always above the cutoff: long run low gray level emphasis, low gray level run emphasis, and the minimum. Short run low gray level emphasis was also often above the cutoff. While only features that passed the feature stability test were included in the analysis, we were interested in examining if these features' poor performance in the IV_{total} and $IV_{residual}$ tests could be attributed to other causes. Therefore, we re-examined the feature stability and found that these features were not as stable as many of the other features that also passed the test. There was no clear way to determine the cutoff for the feature stability test, but this indicates that the poor performance in the IV_{total} and $IV_{residual}$ tests could be due to sensitivity of these features to the ROI placement.

Overall, there was very little to no improvement in the number of features above the cutoff when $IV_{residual}$ was computed compared with IV_{total} . There were fewer features above the cutoff for the controlled protocol compared with the local protocols except when thresholding, smoothing, and bit depth rescaling were used.

Table 7-2. Number of features for each protocol and preprocessing type that have imaging variability compared to inter-patient variability from linear mixed-effects models above the cutoff

Protocol	Feature Group	Thresholding				Thresholding and Smoothing			
		Total Variability		Residual Variability		Total Variability		Residual Variability	
		NSCLC	HNSCC	NSCLC	HNSCC	NSCLC	HNSCC	NSCLC	HNSCC
		Patients	Patients	Patients	Patients	Patients	Patients	Patients	Patients
Controlled Protocol	GLCM (N = 20)	1	1	1	1	0	2	0	2
	GLRLM (N = 11)	3	3	3	3	2	2	2	2
	Intensity (N = 10)	1	1	1	1	1	1	1	1
	NGTDM (N = 5)	0	0	0	0	0	0	0	0
Local Chest Protocol	GLCM (N = 20)	3	4	2	3	3	3	2	3
	GLRLM (N = 11)	3	3	3	3	3	3	3	3
	Intensity (N = 10)	1	1	1	1	1	1	1	1
	NGTDM (N = 5)	0	0	0	0	0	0	0	0
Local Head Protocol	GLCM (N = 20)	2	4	1	3	2	3	2	3
	GLRLM (N = 11)	3	3	3	3	3	3	3	3
	Intensity (N = 10)	1	1	1	1	1	1	1	1

NGTDM (N = 5)	0	0	0	0	0	0	0	0
------------------	---	---	---	---	---	---	---	---

Protocol	Feature Group	Thresholding and Bit Depth Rescaling				Thresholding, Smoothing, and Bit Depth Rescaling			
		Total Variability		Residual Variability		Total Variability		Residual Variability	
		NSCLC	HNSCC	NSCLC	HNSCC	NSCLC	HNSCC	NSCLC	HNSCC
		Patients	Patients	Patients	Patients	Patients	Patients	Patients	Patients
Controlled Protocol	GLCM (N = 20)	0	0	0	0	1	3	1	3
	GLRLM (N = 11)	3	3	3	3	3	3	3	3
	Intensity (N = 10)	1	1	1	1	1	1	1	1
	NGTDM (N = 5)	0	0	0	0	0	0	0	0
Local Chest Protocol	GLCM (N = 20)	2	4	2	2	2	2	2	2
	GLRLM (N = 11)	3	3	3	3	2	2	2	2
	Intensity (N = 10)	1	1	1	1	1	1	1	1
	NGTDM (N = 5)	0	0	0	0	0	0	0	0
Local Head Protocol	GLCM (N = 20)	1	4	0	2	1	2	1	2
	GLRLM (N = 11)	3	3	3	3	1	1	1	1

Intensity (N = 10)	1	1	1	1	1	1	1	1
NGTDM (N = 5)	0	0	0	0	0	0	0	0

GLCM: gray level co-occurrence matrix, GLRLM: gray level run length matrix, NGTDM: neighborhood gray tone difference matrix, NSCLC: non–small cell lung cancer, HNSCC: head and neck squamous cell carcinoma. Total

variability: $IV_{total} = \frac{\sigma_{t,m}/\widehat{f_m}}{\sigma_p/\mu_p}$, residual variability: $IV_{residual} = \frac{\sigma_{\varepsilon,m}/\widehat{f_m}}{\sigma_p/\mu_p}$, with a cutoff of 1/3.

Twenty of the 24 pairwise t-tests of σ_β , σ_v , and σ_ϵ between the controlled protocol and local chest protocol and between the controlled protocol and local head protocol were significant ($p < 0.05$). All comparisons between the controlled and local head protocol were not significant when thresholding and smoothing were applied as the preprocessing. Additionally, σ_ϵ was not significantly different between the controlled and local head protocol when thresholding, smoothing, and bit depth rescaling were applied as the preprocessing. Appendix D Table D-1 shows the p-values for all comparisons.

Since there was a disproportionately high number of GE scanners, the linear mixed-effects models were also run with only the GE scanners. A pairwise t-test was run on σ_β and σ_ϵ between the models with all of the scanners and the models with only the GE scanners. There was a significant difference ($p < 0.05$) for 11 of the 24 comparisons between variabilities calculated from linear mixed-effects models with all scanners and models with GE scanners only. Appendix D Table D-2 shows the p-values for all comparisons.

7.3.5 Quality Assurance Using a Radiomics Phantom

The three cartridges with ABS had noticeable changes over the course of the study. The mean values of the cartridges over time are shown in Appendix D Figure D-12. The three cartridges with ABS displayed a downward trend in mean value over time, while the other cartridges did not show any trend with time. Therefore, the three ABS cartridges were excluded from the QA analysis with a radiomics phantom.

The gray level run length matrix features had a disproportionately high number of scanners outside the established bounds; therefore, these features were not included in the QA analysis. Thus, 35 features with four preprocessing types were included in the QA test. Histograms of the number of scanners with the percentage of features outside the bounds set using the scaled patient standard deviation showed that many scanners had more than 20% of features outside the bounds, as shown in

Appendix D Figure D-13 for each of the rubber, dense cork, and hemp seed cartridges using the HNSCC and NSCLC patient cohorts.

Not all features may be useful, as not all features have been correlated with patient outcomes. Therefore, a subset of features with associated preprocessing type were selected on the basis of studies by Fave et al. and Fried et al. [14, 22]. The features and the preprocessing types that were correlated with patient survival on univariate analysis were included, which resulted in 26 features. Like the gray level run length matrix features, the features of auto correlation, correlation, sum average, sum variance, and the median had a disproportionately high number of scanners outside the bounds. Excluding the features that were shown to not be robust and excluding the gray level run length matrix features reduced the feature set to 16 features with their associated preprocessing types. These 16 features are listed in the Appendix D Table D-3. Figure 7-5 shows histograms for percentages of features outside the bounds (similar to Appendix D Figure D-13, but with the reduced set of features). More scanners had low percentage of features outside $1/3$ of the scaled patient standard deviation in the NSCLC patient cohort than in the HNSCC patient cohort; this is discussed further in the Discussion section. One scanner consistently had the highest percentage of features outside the bounds. However, aside from this scanner, the scanners with the highest percentages of features outside the bounds were not consistent across the different materials.

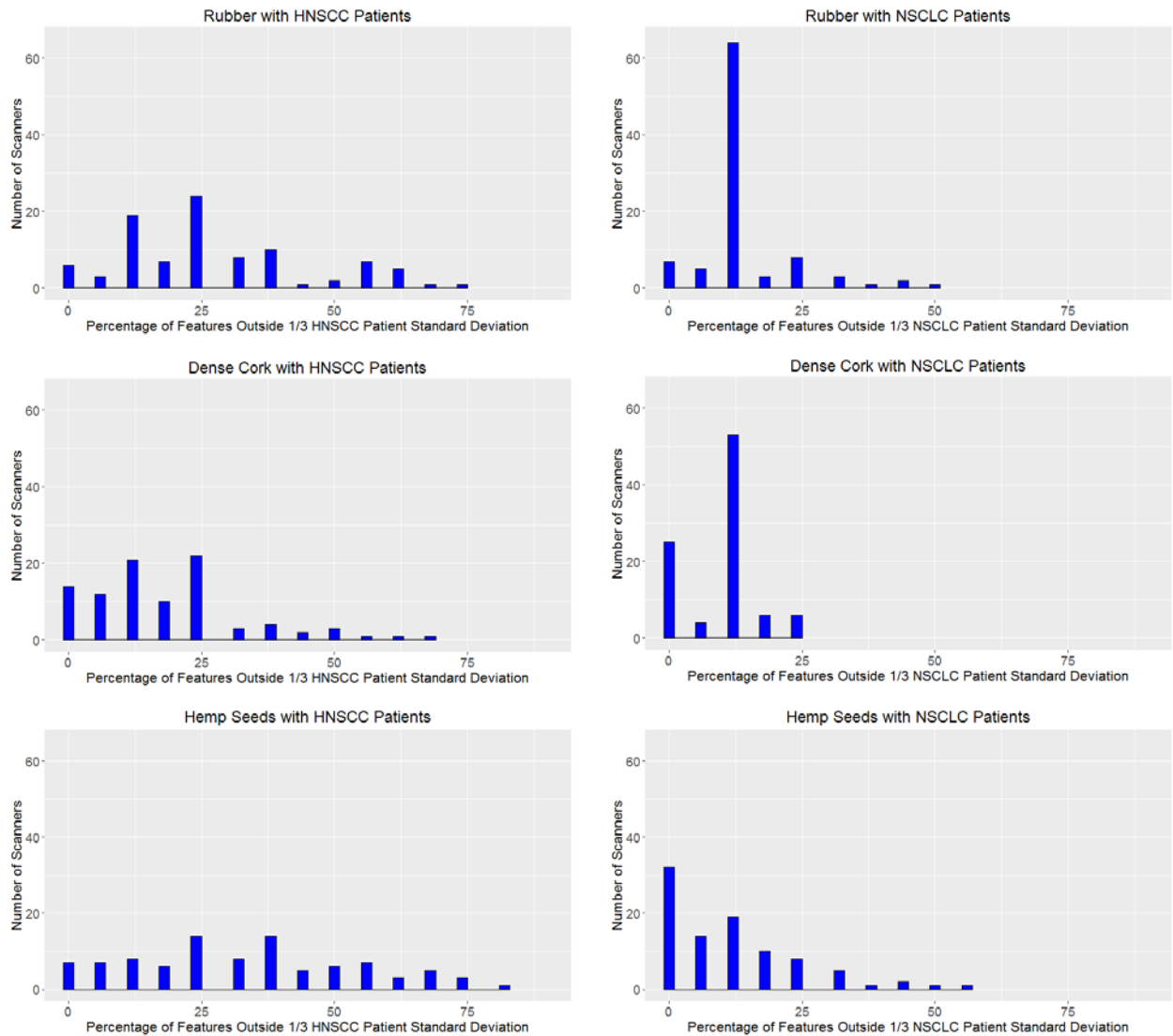


Figure 7-5: Histograms of Number of Scanners that have a Percentage of features outside of patient bounds.

The percentages of features outside 1/3 of the scaled patient standard deviation for rubber, dense cork, and hemp seeds in the head and neck squamous cell carcinoma (HNSCC) patient cohort and the non-small cell lung cancer (NSCLC) patient cohort using the features correlated with patient survival in previous studies without non-robust features. More scanners had fewer features outside 1/3 of the patient standard deviation in the NSCLC patient cohort than the HNSCC patient cohort.

7.4 Discussion

This study showed that imaging variability exists but is not large compared with inter-patient variability for most features. A controlled scan can be helpful for reducing these uncertainties in prospective studies, as there was statistically significantly less variability in the controlled protocol scans than in the local protocol scans. The controlled protocol reduced the total variability by over 50% compared with both local chest and local head protocol scans. It is theoretically possible to correct for the manufacturer and the individual scanner. One possible way to do this is to use a phantom on each scanner to correct for all the factors that could impact the output of a scanner. If this were done perfectly, the imaging variability could be reduced by an additional 7-8% compared with the reduction due to implementing a controlled protocol.

The controlled protocol implemented in this study specified kernels for each manufacturer. Solomon et al. and Winslow et al. compared kernels on Siemens and GE [162, 163]. Both found that the GE standard kernel was the closest match to the B31f or B31s kernel on Siemens, which agrees with our controlled protocol. Additionally, Shafiq-ul-Hassan et al. recently demonstrated the feasibility of correcting for the different kernels, achieving improvements in feature robustness by 30-78% [164]. Our goal in this study was to harmonize the kernels across manufacturers such that the kernel did not affect the imaging variability. However, including this new correction technique may reduce imaging variability further.

Gray level run length matrix features had high feature variability when ROIs were moved. Additionally, these features had the highest imaging variability. We believe that these results are due to the current construction of these features. Examining low gray level run emphasis demonstrates this issue. Low gray level run emphasis is defined as

$$LGRE = \frac{1}{n_r} \sum_{i=1}^M \frac{p_g(i)}{i^2} \quad (7-8)$$

where n_r is the total number of runs, M is the total number of gray levels, i is the gray level, and

$$p_g(i) = \sum_{j=1}^N p(i, j) \quad (7-9)$$

is the sum distribution of the number of runs with gray level i , run length j , maximum run length N , and run-length matrix $p(i,j)$. A slight shift in the distribution of gray levels within the ROI can significantly impact the feature value as the range of the summations remain the same but $p(i,j)$ changes, thus impacting the feature value. Thus we recommend that these features not be used until these issues can be resolved. This problem may be why gray level run length matrix features have not come out in the final models in many studies.

Many of the features showed a correlation between feature value and image thickness that must be considered. Also, the slope of the fixed-effects term for the image thickness was generally the same for a given feature across all models, even in the controlled protocol scans where there were only two image thickness values, indicating the strength of this relationship. This agrees with several studies that have demonstrated the relationship between radiomics features and image thickness [32-34, 165]. However, the high ICC values indicate that the feature value correlation with image thickness cannot be fixed by resampling the image and thus cannot be fixed for retrospective scans for this particular phantom study. When the range of resampled image thickness values was decreased (i.e. not including thicknesses above 5mm), the ICC values remained high. Noise characteristics were not included in this part of the study which can affect feature values as thicker slices can introduce less noise than thinner slices. Even given the limitations of this study, these results indicate that this effect cannot be compensated for after reconstruction with resampling for this phantom study. This is in contrast to the studies by Shafiq-ul-Hassan et al. and Larue et al. who found that resampling an arbitrarily chosen standard voxel size improved feature reproducibility [32, 36]. Therefore, in this study there is a need to control the image thickness as resampling to a variety of image thickness values did not change the feature value, and thus, we recommend controlling image thickness in prospective studies to eliminate this feature value dependence. If the image thickness cannot be completely controlled, the range of image thicknesses used within a study cohort should be limited to reduce this effect.

The importance of a controlled protocol for prospective studies was also demonstrated through the linear mixed-effects models. There was significantly less variability in the controlled protocol scans compared with the local protocol scans. Furthermore, the total variability (Table 7-2) does not include the contribution from the fixed-effect term for image thickness, which would increase imaging variability. Reducing the uncertainty is a crucial step in moving forward with radiomics studies, as reduced uncertainty allows more levels of stratification in prognostic models and enables the movement towards individual prognostic models instead of sorting patients into groups. The manufacturer-wise variation was reduced when a controlled scan was implemented because imaging parameters were harmonized. Many local protocols use the standard kernel, but this kernel is not the best match across different manufacturers. The controlled scan also demonstrated more benefit than post-processing correction for the manufacturer and individual scanner. Radiomics has traditionally been conducted on standard of care imaging, but the large improvements of a controlled protocol demonstrated in this study show the potential importance of such a controlled scan. Thus, efforts should be made to implement a controlled protocol for prospective radiomics studies, and only patients whose imaging parameters match the controlled protocol should be selected in retrospective studies. Studies by Mackin et al. [38] and Fave et al. [150] have shown that tube current and tube voltage do not significantly impact the majority of radiomics features. Therefore, the reconstruction settings dominate the imaging variability and most of the benefit of the controlled scan can be achieved using an additional radiomics reconstruction resulting in no extra dose to the patient.

This study uses the second version of the radiomics phantom. The lessons learned from the first phantom, used in several studies [38, 132], led to this new, improved phantom. The buildup was one considerable difference between the phantoms. Buildup was added to make the phantom more realistic. Also, only the rubber and cork cartridges were kept from the first phantom, as features measured from these cartridges more closely matched NSCLC patient features than did features from other cartridges in the first phantom. In this phantom, we added hemp seed and ABS cartridges, and we

have learned that for future phantoms, ABS cartridges should not be used, as they change over time. The cartridges that were added matched features calculated from patients better and produced a more realistic range of textures. While three of the cartridges changed over time and thus are not optimal options for future work, removing these from the linear mixed-effects models did not change conclusions.

Almost all of the scanners in this study followed established QA protocols. However, in spite of this there were still large imaging variabilities. Therefore, there may be a need for radiomics QA and we demonstrated the potential for a radiomics QA process. The different materials identified different scanners with large percentage of features outside the established bounds, which indicates that a radiomics QA phantom may not be feasible with only one material. The choice of 1/3 in establishing the bounds was arbitrary. The cutoff for the percentage of features failed that would be acceptable to pass the QA process depends on the bounds chosen. When the features found to be correlated with patient survival by Fave et al. and Fried et al. [14, 22] were used, the histograms of the number of scanners with features outside the bounds was reduced, likely because those features are more robust. While studies have found that a radiomics signature developed from NSCLC patients can be used to predict survival in head and neck cancer patients [19, 26], there are distinct feature clusters for the lung and the head and neck cancer patient cohorts [20]. Our patient sets also showed different feature distributions for lung and head and neck patient cohorts, which contributed to the difference in QA results. Therefore, for QA purposes, a distinct radiomics signature should be selected for each cancer site to be credentialed.

There are several limitations to this study. First, the phantom was not imaged by a single user; therefore, there may be some added variability due to different users. Secondly, the phantom materials are not the same as human tissue. Dense cork and rubber have been previously shown to have radiomics feature spectrums similar to those of NSCLC patients [132], and these cartridges have effective atomic numbers close to those of human tissues [133, 134, 147]. Using patients for these

studies is not feasible; therefore, these materials are a close match to human tissues, and results derived from them can be applied to patient CT scans. Additionally, the same phantom was used for chest and head scans. The dimensions of the phantom were designed for chest imaging. Visual inspection of the images did not yield any artifacts specific to the head protocols. While not optimized for head imaging, this phantom still provides valuable information on the radiomics feature variability of these protocols.

Also, there was not an even distribution of scanners by manufacturer. There was a disproportionately high number of GE CT scanners, and it is unknown whether our sample of scanners accurately represents the distribution of scanners in clinical use, as these data are not available. When GE scanners alone were run through the linear mixed-effects model, some variabilities were statistically significantly different between the GE scanners alone and between all scanners. This difference may point to there being scanner-wise variability differences between manufacturers which was not accounted. This was due to the limited number of scanners outside GE which is a limitation of this study. The sample of scanners selected were acquired in Dallas, San Antonio, Houston, Galveston, Baton Rouge, and New Orleans thru proximity and personal contacts. As this sample only constitutes scanners from Texas and Louisiana, the manufacturer distribution may look different in other parts of the USA or in other countries. Additionally, the patient scans used were from selected scanners using well-specified imaging parameters. This may not represent the true inter-patient variation that may exist in a large radiomics study. However, as these patient scans were well controlled, this provides a conservative estimate of the imaging variability effect within patient cohorts. The results from IV_{total} and $IV_{residual}$ are promising given that this may be a conservative estimate and within a larger patient cohort even fewer features may be adversely affected due to larger inter-patient variation.

7.5 Conclusion

A controlled protocol substantially reduces imaging variability compared with local protocols, as the controlled protocol can reduce the total variability by more than 50%. Thus, controlled protocols should be used for radiomics studies. Most of this benefit can be achieved by an extra radiomics reconstruction resulting in no additional dose to the patient. Correcting for the manufacturer and individual scanner can also yield an additional benefit.

Chapter 8 : PET Imaging Protocol Effect on Radiomics Feature Values

8.1 Introduction

Variability in imaging protocols can add noise to radiomics data in patient studies. For PET images, acquisition and reconstruction parameters have been shown to affect radiomics features. In particular, the number of iterations, matrix size, and smoothing filter produce variability in radiomics features [39-49]. In general, these studies have been performed using only one scanner and have investigated only a few of the parameters that can be altered in the imaging protocol. Those studies that used a phantom often used one with uniform spheres, such as the National Electrical Manufacturers Association phantom, which may not be representative of the texture within patients' tumors. Thus, although these studies have provided valuable insight into particular issues, they may not be generalizable.

In this study, we aimed to fill this gap by using a phantom that provided radiomics feature values similar to those found in patients. We used scanners from several different vendors and investigated the effects of changing all of the parameters that could be changed for reconstructions. Filling this gap allows for more precise inclusion criteria in patient studies in order to reduce the noise in radiomics features to produce the best possible prediction studies.

8.2 Methods

8.2.1 Phantom Scans

PET images of a 3-dimensional Hoffman brain phantom were acquired on GE Discovery 710 (GE Healthcare, Chicago, IL), Siemens mCT (Siemens Healthineers, Forchheim, Germany), and Philips Vereos (Philips Healthcare, Eindhoven, The Netherlands) PET scanners. A standard-protocol scan was acquired on each machine, and then each parameter that could be changed was altered individually. For example, to assess the impact of time per bed position, the other standard-protocol parameters were held constant while the time per bed position was set to 2 minutes for one reconstruction, 3 minutes

for another reconstruction, 4 minutes for another reconstruction, and 5 minutes for another reconstruction. The parameters that could be changed and the settings investigated for each scanner are listed in Table 8-1.

Table 8-1. Parameters Changed to Investigate Impact on Radiomics Features

Parameters	Scanner		
	GE Discovery 710	Siemens mCT	Philips Vereos
Field of view (cm)	25, 50, 70		
Filter cutoff (mm)	1, 3, 5, 8, 10	1, 3, 5, 8, 10	None, 1, 3, 5, 8, 10
Iterations × subsets	1 × 4, 2 × 8, 4 × 8, 2 × 18, 4 × 32	Non-TOF: 1 × 4, 2 × 8, 4 × 8, 2 × 12, 4 × 24 TOF: 1 × 21, 2 × 21, 3 × 21, 4 × 21	1 × 4, 2 × 8, 4 × 8, 2 × 20, 3 × 15, 4 × 32
Matrix size	128, 192, 256	128, 200, 256, 400, 512	
Time per bed position (min)	2, 3, 4, 5	2, 3, 4, 5	2, 3, 4, 5
Type of reconstruction	VPFX, VPFX-S, VPHD, VPHD-S, QCFX-S, QCHD-S	Backprojection, backprojection TOF, iterative, iterative TOF, TRUEX, TRUEX TOF	
Z smoothing	None, light, standard, heavy		

TOF: time of flight

Types of reconstruction are proprietary names used by each vendor.

The standard-protocol settings for the GE scanner were 70 cm field of view, 5 mm filter cutoff, 2 iterations and 18 subsets, 192 matrix size, standard z smoothing, 6 minutes per bed position, and VPFX-S reconstruction. The standard-protocol settings for the Siemens scanner were 82 cm field of view, 5 mm filter cutoff, 2 iterations and 21 subsets, 200 matrix size, 5 minutes per bed position, and TRUEX time-of-flight (TOF) reconstruction. As the Siemens scanner also allows for continuous bed motion, this type of acquisition was also explored and treated as an additional scanner. The standard-protocol settings were the same as the fixed number of bed positions acquisition but with 0.4 mm/s as the bed speed. The standard-protocol settings for the Philips scanner were 60 cm field of view, 3 iterations and 15 subsets, 128 matrix size, no smoothing filter, and 5 minutes per bed position. The term “standard protocol” here means that it was the baseline acquisition. The time per bed position was longer than that used clinically and for the Siemens scanner, the standard acquisition used at MD Anderson is continuous bed motion. The phantom was injected with 2.53-2.75 mCi of F-18 fluorodeoxyglucose for each scan and then imaged about 30 minutes later. The weight was set to 20 kg to obtain standardized uptake values (SUVs) in the phantom that were similar to the SUVs in patient tumors.

8.2.2 Patients

Data from a patient cohort were used to provide context to the variability observed between scanners. For example, an interscanner variation of 0.4 for a given feature with a phantom does not necessarily represent the impact of interscanner variation in a patient study. However, if the interscanner variation is computed relative to interpatient variation, the impact on patient studies can be directly observed. In order to make interscanner comparisons that were relative to interpatient variation in this study, PET studies of 224 patients with non-small cell lung cancer (NSCLC) were retrospectively analyzed. The requirement for informed consent was waived by the Institutional Review Board at The University of Texas MD Anderson Cancer Center. This cohort consisted of 84 women and

140 men with an average age of 65 years (range, 39–89 years), average height of 171 cm (range, 147–195 cm), average weight of 82 kg (range, 39–151 kg), and average tumor volume of 90 cm³ (range, 0.4–920 cm³).

8.2.3 Feature Extraction

Each phantom scan was semiautomatically contoured with 10 cylindrical regions of interest (ROIs) using in-house developed MATLAB (MathWorks, Natick, MA) scripts. Each ROI had a diameter of 19.4 cm and a height of 1 cm. Some slices of the phantom with contours are shown in Figure 8-1. A threshold of 0.4 SUV was used before feature calculation on the phantom images to remove background noise or activity that had leaked to the edges of the phantom container. The patient images were contoured using PET Edge in MIM (MIM Software Inc., Cleveland, OH). Forty-five features were extracted using 2 preprocessing methods: (1) a fixed-bin-width of 0.5 SUV, as suggested by Leijenaar et al. [166], and (2) rescaling to 64 levels, as suggested by Hatt et al. [167]. Radiomics features were calculated using IBEX, a freely available radiomics tool [137, 138]. The features used are listed in Table 8-2. More information about these features can be found in a publication by Zhang et al. [137]. The settings for each of the features were the same as those listed in Fave et al.'s Supplemental Material [14], except for neighborhood gray tone difference matrix, where we set the neighborhood to 3 owing to the large voxel size in PET images.

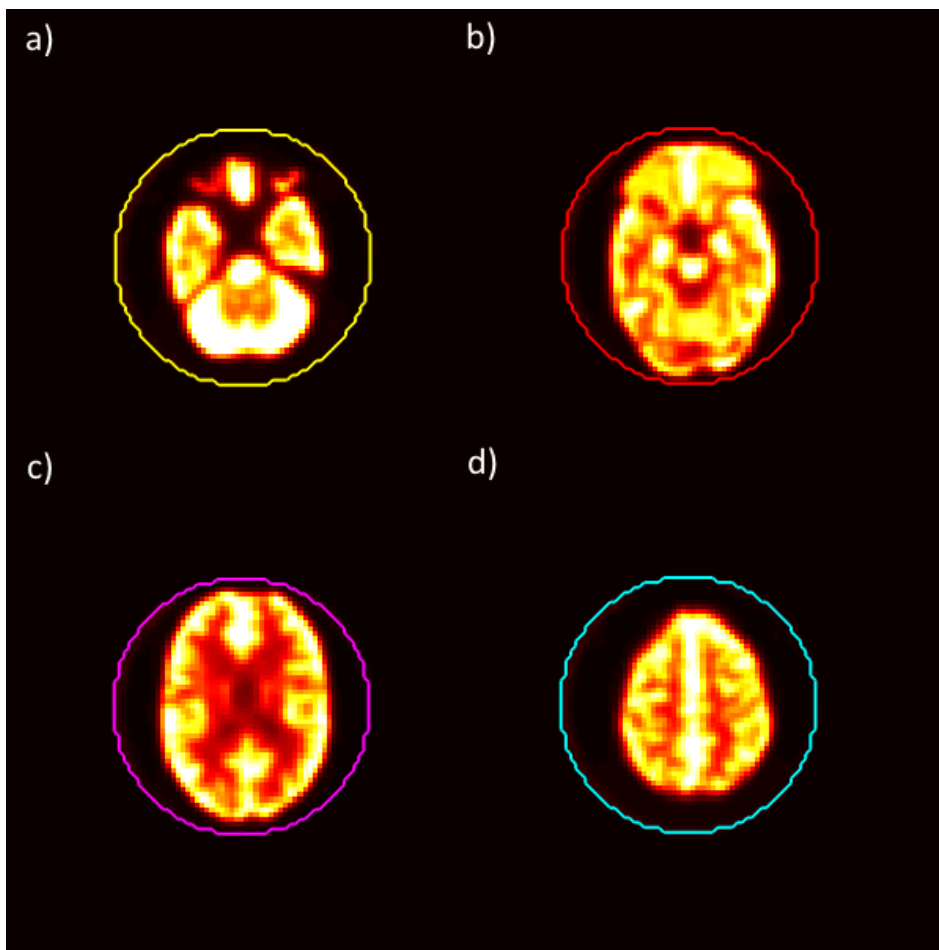


Figure 8-1: Slices of Hoffman Phantom.

Four slices of the Hoffman phantom are shown. Each slice is from a different ROI among the 10 ROIs that were drawn in the phantom. The example slices shown here are from different regions within the phantom: (a) near the bottom of the phantom, (b) between the bottom and the middle of the phantom, (c) between the middle and the top of the phantom, and (d) near the top of the phantom.

Table 8-2. Radiomics Features Used in PET Analysis

Gray Level Co-occurrence Matrix	Gray Level Run Length Matrix	Intensity Histogram	Neighborhood Gray Tone Difference Matrix
Auto Correlation	Gray Level Nonuniformity	Energy	Busyness
Cluster Prominence	High Gray Level Run Emphasis	Entropy	Coarseness
Cluster Shade	Long Run Emphasis	Kurtosis	Complexity
Cluster Tendency	Long Run High Gray Level Emphasis	Skewness	Contrast
Contrast	Long Run Low Gray Level Emphasis	Standard Deviation	Texture Strength
Correlation	Low Gray Level Run Emphasis	Uniformity	
Difference Entropy	Run Length Nonuniformity	Variance	
Dissimilarity	Run Percentage		
Energy	Short Run Emphasis		
Entropy	Short Run High Gray Level Emphasis		
Homogeneity	Short Run Low Gray Level Emphasis		
Homogeneity 2			
Information Measure Correlation 1			
Information Measure Correlation 2			
Inverse Difference Moment Norm			
Inverse Difference Norm			
Inverse Variance			
Max Probability			
Sum Average			
Sum Entropy			
Sum Variance			
Variance			

8.2.4 Statistical Analysis

The intraclass correlation coefficient (ICC) was used to determine whether changes in each parameter affected the measured radiomics features. This was done separately for each adjustable parameter on each scanner, with the ROIs as the subjects and the different parameter values as the raters. The 2-way random effects, consistency, single rater/measurement ICC as described by Shrout and Fleiss [159] was computed in R (version 3.4.3) using the psych package (version 1.7.8) [160]. To determine the level of reliability indicated by the ICC values, the guidelines published by Koo and Li [161] were followed: ICC values lower than 0.5 signified poor reliability, those between 0.5 and 0.75 signified moderate reliability, those between 0.75 and 0.9 signified good reliability, and those greater than 0.9 signified excellent reliability.

Interscanner analysis was performed using the standard-protocol image from each scanner. The standard deviation across the ROIs from the four scanners (GE, Philips, Siemens, and Siemens using continuous bed motion) was compared to the standard deviation from the NSCLC patient cohort. This was done separately for each feature and preprocessing technique. Additionally, the mean value for each feature and preprocessing technique combination from the phantom standard-protocol images was compared to the mean value from the patient images for the same combination. If the phantom mean was not within two standard deviations of the patient mean for a given feature, the feature was not included when calculating the interscanner variation metric.

8.3 Results

For all scanners, most features had good (ICC > 0.75) to excellent (ICC > 0.9) reliability when reasonable parameter choices were used. Here, “reasonable” refers to parameter values that are used in clinics. For example, extremely low or extremely high effective iteration values (iterations × subsets) were excluded, as these are not actually used in clinics. The following paragraphs summarize the results obtained using the reasonable parameters and give the percentage of features in each of the reliability

classifications described in the Statistical Analysis section. The specific ICC values for each feature using all parameter values and the subset of parameter values that were deemed reasonable are presented in Appendix E.

For the GE scanner, when the pixel size was resampled, all features had excellent reliability with both preprocessing types for field of view and matrix size. When only filter cutoff values below 6 mm were included, 96% of features had excellent reliability, 3% of features had good reliability, and 1% of features had moderate reliability (busyness calculated using fixed-bin-width preprocessing). For iterations and subsets, when only effective iterations between 16 and 36 were included, 87% of features had excellent reliability, 12% of features had good reliability, and 1% of features had poor reliability (complexity calculated using 64-level preprocessing). When time per bed position was altered, all features had excellent reliability with both preprocessing types. For the type of reconstruction, when Q.Clear was not included (reconstruction types QCFX-S and QCHD-S), 92% of features had excellent reliability and 8% of features had good reliability. For z smoothing, 89% and 11% of features had excellent and good reliability, respectively.

For the Siemens scanner, when only filter cutoff values below 6 mm were included, 80% of features had excellent reliability, 19% of features had good reliability, and 1% of features had moderate reliability (busyness calculated using fixed-bin-width preprocessing). For matrix size, when the pixel size was resampled, 94% and 6% of features had excellent and good reliability, respectively. For iterations and subsets using TOF, 83% of features had excellent reliability, 12% of features had good reliability, and 4% of features had moderate reliability. For iterations and subsets using non-TOF, when only effective iterations between 16 and 24 were included, 76% of features had excellent reliability, 18% of features had good reliability, and 7% of features had moderate reliability. For the time per bed position, all features had excellent reliability with both preprocessing types. Similar results were found using continuous bed motion.

For the Philips scanner, when only filter cutoff values below 6 mm were included, 71% and 29% of features had excellent and good reliability, respectively. For iterations and subsets, when only effective iterations between 16 and 45 were included, 92% and 8% of features had excellent and good reliability, respectively. For the time per bed position, all features had excellent reliability with both preprocessing types. The distribution of features in each reliability grouping for each imaging protocol parameter and preprocessing technique for the Philips scanner is shown in Figure 8-2. The data used to create this figure, as well as the data for the other scanners, are detailed in Appendix E.

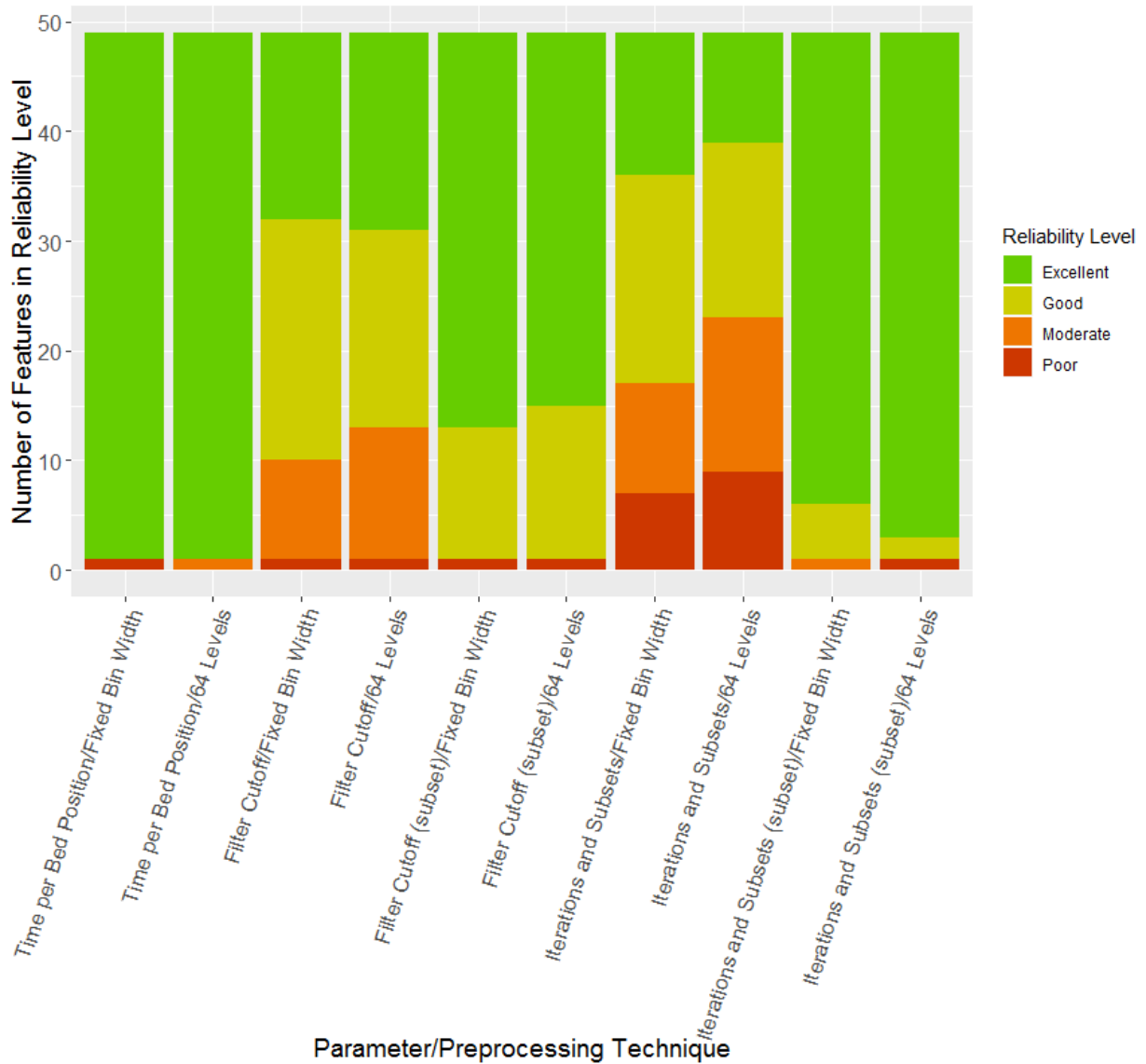


Figure 8-2: Bar Plots of Features by Reliability Level.

For each imaging-protocol parameter using each of the 2 preprocessing techniques (fixed bin width and 64 levels), the number of features in each ICC reliability level is shown: excellent reliability (green) is $ICC > 0.9$, good reliability (yellow) is $0.75 < ICC < 0.9$, moderate reliability (orange) is $0.5 < ICC < 0.75$, and poor reliability (red) is $ICC < 0.5$. When parameters were limited to values seen in clinics, most features had excellent reliability, regardless of preprocessing technique. The subset for filter cutoff contains reconstructions for which the filter cutoff was below 6 mm. The subset for iterations and subsets contains reconstructions for which the effective number of iterations was between 16 and 45.

Across all scanners, the average ICC was typically higher with fixed-bin preprocessing than with 64-level preprocessing. This was the case for all of the imaging parameters on the Siemens (both for the step-and-shoot and the continuous-bed-motion acquisition) and Philips scanners.

For the interscanner analysis, the average ratio of the standard deviation across all features from the standard-protocol phantom scans to the standard deviation from the NSCLC patient scans was 0.73 using fixed-bin-width preprocessing and 1.0 using 64-level preprocessing. With 64-level preprocessing, 7 features on the phantom scans had a mean value more than 2 standard deviations from the patient-scan mean value for that feature. Excluding these features, the mean ratio of the phantom-scan standard deviation to the patient-scan standard deviation was reduced to 0.92. We changed the weight on the GE standard-protocol scan to 35 kg to make the mean SUVs more similar to the mean SUVs on the other scanners. The average ratio of the standard deviation across all features from the standard-protocol phantom scans to the standard deviation from the NSCLC patient scans was 0.76 using fixed-bin-width preprocessing and 1.1 using 64-level preprocessing. With 64-level preprocessing, 8 features on the phantom scans had a mean value more than 2 standard deviations from the patient-scan mean value for that feature. Excluding these features, the mean ratio of the phantom-scan standard deviation to the patient-scan standard deviation was reduced to 0.96. These results demonstrated that scaling the SUVs was not a factor in interscanner variability.

8.4 Discussion

In this study, we investigated the impact on the variability of radiomics feature values of the imaging protocol parameters that could be retrospectively changed on GE, Philips, and Siemens PET scanners. We found that as long as reasonable parameter values were used (i.e., parameter values that are actually used in clinics), almost all features had at least good reliability. These results demonstrate that on a given scanner, radiomics data for patients scanned using different imaging protocols can be combined without adding significant noise.

However, interscanner variability was about equal to interpatient variability. This implies that caution should be used when combining data for patients scanned on equipment from different vendors for radiomics analysis as the observed stratification in a patient cohort may be due to differences in scanners and not true patient differences. A phantom with radiomics feature values similar to those of patients, such as the one used here, should be used to verify which features can be used in the patient analysis. If a phantom cannot be acquired, our full interscanner analysis, which is provided in Appendix E, can serve as a reference on which features are robust enough to be included in the analysis and which features are too variable and should be excluded.

We found that there was less variability in features when using the fixed-bin-width preprocessing method than with the 64-level preprocessing method. Leijenaar et al. [166] also found that using a fixed-bin-width was preferable in their interpatient and inpatient comparisons of two preprocessing techniques (fixed-bin-width and fixed number of levels) in 35 lung cancer patients.

When we included a large range of parameter values, our results agreed with those of previous studies that found that the parameters of filter cutoff, matrix size, and iterations and subsets affect feature values [39-49]. We were also able to show that resampling the image in our radiomics software prior to feature extraction removed the impact of matrix size on feature values. Additionally, the impact of variations in filter cutoffs and iterations and subsets could be removed if only parameter values that are commonly used in clinics were included.

This study has several limitations. First, only one scanner from each vendor was used; therefore, the variability of different models from a given vendor could not be explored. Second, only one acquisition per scanner was used for this study. We previously found the repeatability of a particular acquisition on a scanner to be very high; thus, we do not believe acquisition-level variability affected the results of this study. Third, this study was conducted using a phantom, which allowed for consistency in subject material across scanners, but the phantom is only a representation of patient texture. For this particular study, most features were within two standard deviations of the NSCLC

patient cohort average, showing that the phantom features were a good representation of the features of this patient cohort. However, the activity concentration was higher in our phantom than that seen in typical patient PET scans. This could affect the convergence rate of the reconstruction algorithms. To assess the impact of convergence rates, many scans with different activity levels would have to be acquired and the whole analysis repeated, which is outside the scope of this paper. Another limitation is that, for practical reasons, we only examined a subset of the entire parameter space. For example, the voxel size relative to the filter cutoff values may also have affected the results. For the standard-protocol scan, the voxel size was $0.36 \times 0.36 \times 0.33$ cm on the GE scanner, $0.41 \times 0.41 \times 0.2$ cm on the Siemens scanner, and $0.2 \times 0.2 \times 0.2$ cm on the Philips scanner. The small values of the cutoff value investigated (particularly 1 mm) would only represent part of a voxel and, would therefore, not affect the image. Smaller voxel sizes could be affected more by these filter cutoff values and could result in lower ICC values. Finally, this study used an adult NSCLC patient cohort; different adult patient cohorts may have different interpatient variability levels and different ratios of the standard deviation of the phantom measurements to that of the patients' measurements. Results may be different in pediatric patient cohorts where the average weight is much less than the average weight of the cohort in this study.

8.5 Conclusion

We found that all imaging-protocol parameters had good reliability across feature values when the parameter values were within limits typically used in clinics. However, interscanner variability was about equal to interpatient variability. Therefore, caution must be used when combining patients scanned using equipment from different vendors into single radiomics data sets.

Chapter 9 : CT- and PET-Based Radiomics Survival Modeling of HNSCC Patients

9.1 Introduction

Studies have shown that imaging protocol differences, such as pixel size can increase uncertainties in patient datasets [32, 37, 38, 132, 136]. A recent phantom study showed that inter-scanner variability can be reduced by more than 50% when a controlled imaging protocol is used for CT imaging [136]. For PET images, acquisition and reconstruction parameters have been shown to affect radiomics features; particularly, the number of iterations, matrix size, and smoothing filter have demonstrated variability [42, 46-48, 168].

Based on these uncertainty studies, our hypothesis is that outcome models built with data from patients on controlled imaging protocols should perform better than models built with data from a varied patient cohort since the noise from imaging variability is removed in the former model. We aimed to test this hypothesis in large cohorts of CT and PET head and neck cancer patients.

9.2 Materials and Methods

9.2.1 CT Patients

Patients who were treated with definitive radiotherapy for head and neck squamous cell carcinoma (HNSCC) at least five years ago, had pre-treatment CT images available, did not have a tumor stage of Tx (primary tumor could not be assessed), T0 (no evidence of primary tumor), or Tis (carcinoma *in situ*), and did not have a nodal stage of Nx (regional lymph nodes could not be assessed) were considered eligible. We retrospectively reviewed contrast-enhanced pretreatment CT images and medical records of 652 patients with oropharyngeal HNSCC that were treated between March 2004 and November 2013 with a waiver of informed consent from the Institutional Review Board at The University of Texas MD Anderson Cancer Center. All patients were scanned on GE scanners (GE Healthcare, Chicago, IL). The primary gross tumor volume (GTV) was contoured by two radiation oncologists specific for this study. In addition, 156 HNSCC patients from Aerts et al.'s data set from

MAASTRO were included [26]. Fifty patients were excluded from this data set due to no contoured GTV, other missing data elements, or issues with importing data into our radiomics software.

Patients whose GTV was more than 50% affected by streak artifacts were excluded from our study. Our previous work has shown that this cutoff was useful for including only those patients whose features from GTV not affected by streak artifacts represented features from the whole GTV [156]. Removing all patients with any streak artifact within their GTV would have removed 215 patients. Therefore, this method allows many more patients to be included in the study, as this resulted in the removal of only 32 patients from the study, while not impacting feature values. The remaining 726 patients were divided into training and testing cohorts by medical record number (MRN): those with an odd MRN were placed into the training cohort (377 patients), and those with an even MRN were placed into the testing cohort (349 patients). The patient demographics for each cohort are summarized in Table 9-1.

Our previous work has shown that inter-scanner variability can be significantly reduced when using a controlled protocol [136]. To investigate this impact on the prognostic ability of patient outcome models, we included in these cohorts only patients who had been scanned on a GE scanner with a standard kernel, 1.25-mm image thickness, and 25-cm field of view because the largest subset cohort could be created from the original cohort using these settings. Most of the acquisition parameters have been shown to not impact features, while these reconstruction parameters (kernel, image thickness, and field of view) have been shown to affect features [32, 34, 37, 38, 150]. Thus we focused reconstruction parameters for selecting the subset of patients. These patients were only from MD Anderson as the MAASTRO data was not on a GE scanner.

9.2.2 PET Patients

Patients who were treated with definitive radiotherapy for HNSCC at least four years ago, had pre-treatment PET images available, did not have a tumor stage of Tx (primary tumor could not be

assessed), T0 (no evidence of primary tumor), or Tis (carcinoma *in situ*), and did not have a nodal stage of Nx (regional lymph nodes could not be assessed) were considered eligible. We retrospectively reviewed the images and medical records of 445 patients with oropharyngeal HNSCC that were treated between March 2004 and November 2013 with a waiver of informed consent from the Institutional Review Board at The University of Texas MD Anderson Cancer Center. In addition, we used images, patient survival data, and demographics from the Head-Neck-PET-CT TCIA collection [169, 170]. This collection contained 298 patients, 241 of whom were included; those excluded had lesions with no F18-FDG PET radiotracer uptake or there were issues with importing data into our radiomics software. Each patient's primary GTV was contoured using MIM PET Edge (MIM Software Inc, Cleveland, OH).

The 686 patients were divided into training and testing cohorts by MRN: those with an odd MRN were placed into the training cohort (345 patients), and those with an even MRN were placed into the testing cohort (341 patients). The patient demographics for each cohort are summarized in Table 9-1.

To investigate the effect of reducing inter-scanner variability on the predictive performance of patient outcome models, we included in these cohorts only patients who had been scanned on a GE scanner with two iterations and 20 or 21 subsets; these reconstruction settings were chosen to enable the largest subset cohort to be created from the original cohort. Additionally, in our unpublished work we have found that iterations and subsets can cause the largest discrepancies in radiomics features from the reconstruction parameters that can be changed. However, inter-vendor variances can be large, thus restricting this subset to only patients imaged on a GE scanner is the main driving force in reducing the uncertainty for this study.

Table 9-1. Patient Demographics

	CT Patients		PET Patients	
	Training Cohort	Testing Cohort	Training Cohort	Testing Cohort
Number of patients	377	349	345	341
Number of events	97	75	76	51
Age (years)*	59 (21-87)	57 (30-80)	60 (34-87)	58 (35-90)
HPV status				
Positive	224	189	207	206
Negative/unknown	153	160	138	135
Tumor stage				
T1	71	78	52	63
T2	143	142	131	142
T3	88	72	111	75
T4	75	57	51	61
Nodal stage				
N0	47	40	47	38
N1	34	34	39	40
N2	286	260	248	245
N3	10	15	11	18

AJCC stage				
I-II	20	20	18	21
III	48	45	57	52
IV	309	284	270	268
Primary Gross Tumor Volume (cm ³)*	9 (0.3-326)	8 (0.3-150)	9 (0.8-81)	9 (0.4-123)

* median; range in parentheses

9.2.3 Feature Extraction

The radiomics features were calculated using IBEX, an open-source radiomics tool [137, 138]. Tables of the extracted features are provided in the Supplemental Material. The settings for each feature were the same as those listed in Fave et al.'s Supplemental Material [14]. All of the features were calculated by using four different preprocessing techniques for the CT images: (1) thresholding (lower limit -100 HU, no upper limit), (2) thresholding and a Butterworth smoothing filter (order of 2, cut-off of 125), (3) thresholding and 8-bit depth resampling, and (4) thresholding, 8-bit depth resampling, and Butterworth smoothing. Different features have been shown to be most prognostic with different preprocessing techniques, which is why this assortment of preprocessing techniques was chosen [143]. For the PET images, all of these features were preprocessed using two methods: (1) a fixed bin width of 0.5 SUV, as suggested by Leijenaar et al. [166], and (2) rescaling to 64 levels, as suggested by Hatt et al. [167]. The volume of each GTV was also extracted.

9.2.4 Model Building

The modeling process used here is based on that used for several of our previous, successful radiomics studies [14, 15, 22]. The overall survival was defined as the time interval from the end of definitive radiotherapy to death, and was censored at the last follow-up for patients who were alive. The end point of overall survival was selected for this study because the number of events were higher than events using locoregional control or freedom from distant metastases as an end point. The model was built by using the training data and then receiver operator curve statistics were obtained by using the trained model on the testing data. The radiomics features and volume of the training data were scaled by subtracting the mean and dividing by the standard deviation for each since Lasso penalizes larger values more (R version 3.5.1). Tumor volume and HPV status were the only clinical variables used in order to focus on the effect of the radiomics features.

To begin building the model, we first used univariate Cox proportional hazards models to select

the one preprocessing technique for each feature that had the most significant association with overall survival. Clinical variables were used in forward selection, keeping only those that reduced the Akaike information criteria (AIC) by more than 2. Next, the selected clinical variable(s) were held constant in a univariate Cox proportional hazards model with the prescreened radiomics features to further reduce the dimensionality of the data (R survival package version 2.42-6). The features that had a p-value less than 0.01 were kept. One thousand bootstrap iterations of Lasso regression, using the selected radiomics features and clinical variables, were conducted (R glmnet package version 2.0-16). For these 1000 iterations, the Lasso was fit by using the minimum lambda determined from a 10-fold cross-validation with a maximum of 1000 iterations. The covariates selected in more than 50% of the 1000 bootstrap iterations were kept. Due to the minimum lambda under penalizing the regression, a final forward selection was performed. Those covariates that reduced the AIC by more than 2 were selected. A final Cox model was fit by using these covariates and the non-scaled training data.

The area under the curve (AUC) of the final Cox model when predicting overall survival in the testing data was calculated at 3 years (R survivalROC package version 1.0.3). Patients were assigned to the “High Risk” group if their prediction score was higher than the median; otherwise they were assigned to the “Low Risk” group. The survival probability curve of each group was estimated by the Kaplan-Meier method. The separation between these groups was evaluated by the log-rank test and determined to be significant if the p-value was less than 0.05 (R survival package). Models were built separately for the whole patient cohort and the subset of patients with the same imaging protocol.

We also examined the HPV positive and negative/unknown patients separately because HPV status is a strong known predictor of overall survival. Additionally, most of the patients in our original patient cohort had oropharyngeal cancer, therefore, we analyzed the data using only these patients as well. For these subgroups, the whole modeling process was repeated, including modeling with only those patients with the same imaging protocol to allow for comparisons.

9.3 Results

9.3.1 CT Patients

When using the whole patient cohort, volume and HPV status were selected from the forward selection of the clinical variables. Twelve radiomics features had a p-value < 0.01 when tumor volume and HPV status were held within the Cox proportional hazards model. Five covariates were selected from the bootstrap Lasso. The final selected model contained the following four covariates: tumor volume, HPV status, gray level nonuniformity calculated using thresholding and bit depth resampling, and inverse difference norm calculated using thresholding. The AUC of this model on the testing data was 0.72. The High Risk and Low Risk groups were statistically separated ($p=5 \times 10^{-4}$). Survival plots are shown in Figure 9-1. However, when a Cox model with these covariates was fit on the testing data, volume, gray level nonuniformity, and inverse difference norm were just under the significance threshold ($p=0.027$, $p=0.024$, and $p=0.017$, respectively), and HPV status was not significant ($p=0.18$). Volume alone or volume and HPV status fit in a Cox model on the training data and evaluated on the testing data provided an AUC of 0.73. Adding any radiomics features to this reduced the AUC.

The CT imaging protocol is also known to affect radiomics features measured from CT images [32, 37, 38, 132, 136]. Therefore, to reduce the noise in the data sets, only those patients scanned on a GE scanner with the same imaging protocol were included. This reduced the training data to 260 patients and the testing data to 251 patients. The final model using this data included two covariates and had an AUC of 0.55 on the testing data. However, neither covariate was significant ($p=0.90$, $p=0.79$) in the testing data, so this attempt to control for imaging parameters was not effective. The High Risk and Low Risk groups were not statistically separated, and the survival curves for these risk groups are shown in Figure 9-1 alongside the survival curves using all of the patients.

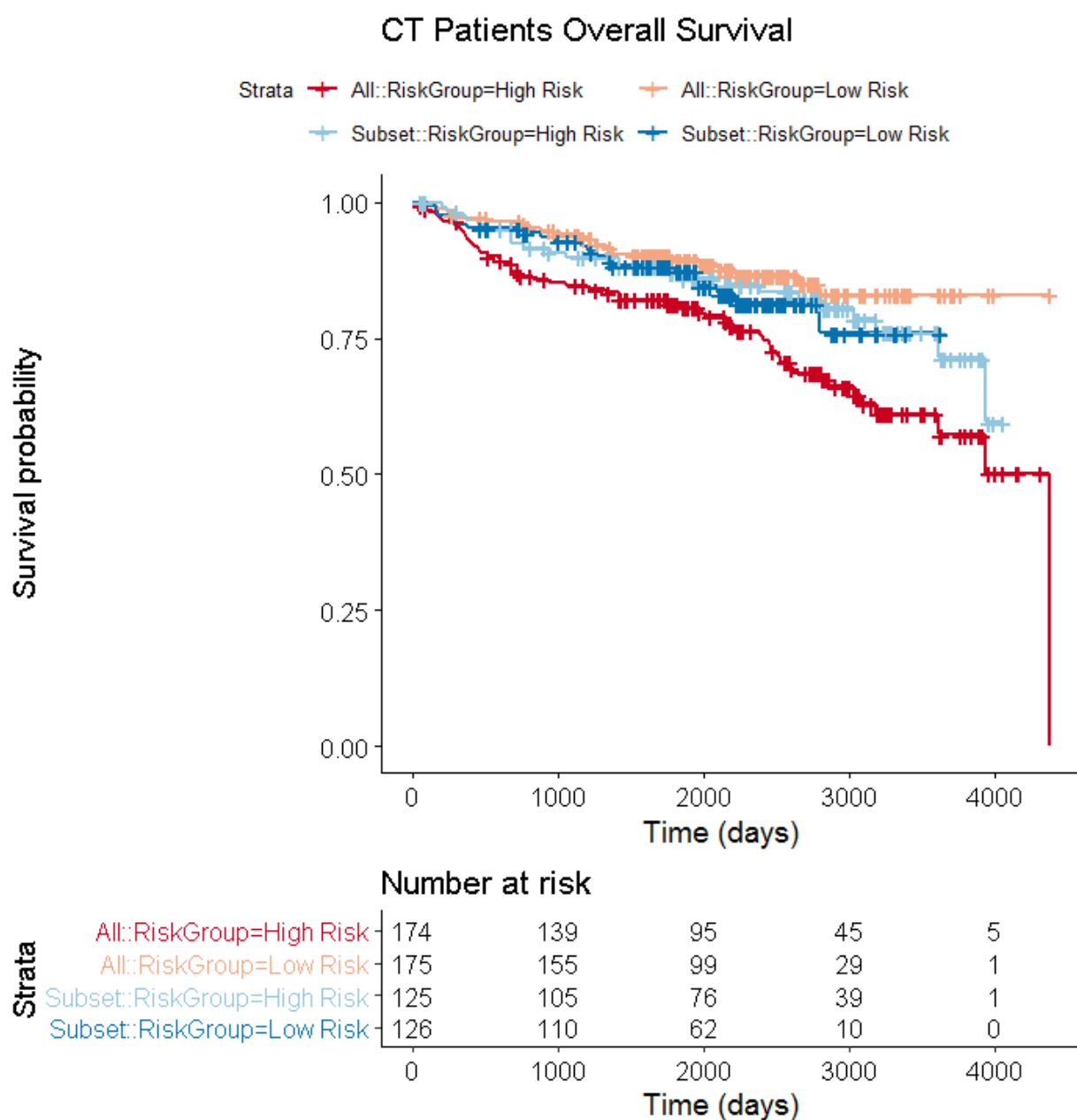


Figure 9-1: Patient survival curves for CT radiomics models.

Patient survival curves using CT patient data for the cohort using all patients and the subset of patients that had the same imaging protocol. For the cohort using all patients, the testing data were from 349 patients who were assigned to High Risk or Low Risk groups according to prediction scores from the Cox model fit using the training data and the four covariates: volume, HPV status, gray level nonuniformity calculated using thresholding and bit depth resampling, and inverse difference norm calculated using thresholding. The separation between the curves was statistically significant ($p=5 \times 10^{-4}$). These patient

curves are called “All” and are in red and orange. For the subset of patients with the same imaging protocol, the testing data were from 251 patients who were assigned to High Risk or Low Risk groups according to prediction scores from the Cox model fit using the training data and the two covariates: HPV status and cluster tendency calculated using thresholding, smoothing, and bit depth resampling. The separation between the curves was not statistically significant. These patient curves are called “Subset” and are in blue and light blue.

When analyzing the CT data, inclusion of data from Aerts et al. [26] substantially affected the results. Although the MD Anderson data set was large, no radiomics feature was produced from the modeling process that was also significant in the testing data. However, inclusion of Aerts et al. [26] data produced two radiomics features that were also significant in the testing data and an AUC above 0.7, as discussed at the beginning of the results presented here.

Examining subgroups of only HPV positive, HPV negative/unknown, or oropharyngeal cancer patients did not improve these results. The information on the covariates selected and the AUC for these patient cohorts can be found in Appendix F.

9.3.2 PET Patients

When using the whole patient cohort, HPV status was selected from the forward selection of the clinical variables. Four radiomics features had a p-value < 0.01 when HPV status was held within the Cox proportional hazards model. Three covariates were selected from the bootstrap Lasso. The final selected model contained two covariates: HPV status and coarseness calculated using 64 gray levels. The AUC of this model on the testing data was 0.59. However, neither of the covariates was significant ($p=0.69$, $p=0.16$) when the Cox model was fit using the testing data or when selecting only one covariate. The High Risk and Low Risk groups were not statistically separated, as shown by the survival plots for these patients in Figure 9-2 where the curves overlap.

The subsets and iterations in a PET imaging protocol are known to affect radiomics features measured from PET images [42, 46-48, 168]. Therefore, including only patients scanned on a GE scanner with 20 or 21 subsets and two iterations reduced the training data to 144 patients and the testing data to 168 patients. These patients were imaged on Discovery ST, Discovery STE, or Discovery RX PET scanners which are all non-time of flight and did not model point spread function. The final model included no covariates, even when relaxing the p-value for passing the additional prescreening univariate Cox analysis, so this attempt to control for imaging parameters was not effective.

PET Patients Overall Survival

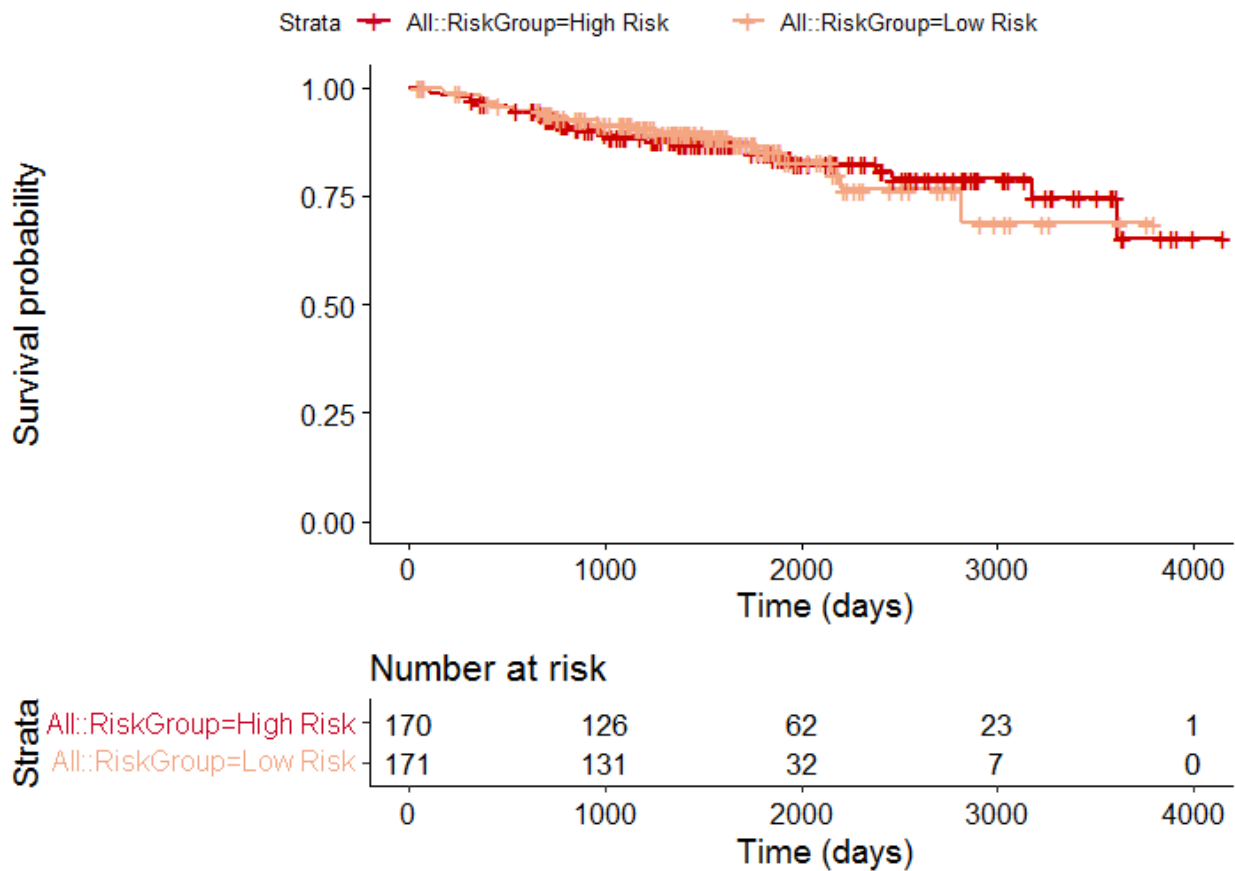


Figure 9-2: Patient survival curves for PET radiomics models.

Patient survival curves using PET patient data. The testing data were from 341 patients who were assigned to High Risk or Low Risk groups according to prediction scores from the Cox model fit using the training data and two covariates: HPV status and coarseness calculated with use of using 64 gray levels. The High Risk and Low Risk groups were not statistically separated as shown by the overlap of the survival curves. These patient curves are called “All” and are in red and orange. For the subset of patients with the same imaging protocol, no covariates were selected, therefore, the patients could not be separated into High Risk and Low Risk and no curves are displayed for the subset patient group.

Examining subgroups of only HPV positive, HPV negative/unknown, or oropharyngeal cancer patients did not improve these results. The information on the covariates selected and the AUC for these patient cohorts is in Appendix F.

9.4 Discussion

In this study, we investigated radiomics features for HNSCC patients by using CT and PET images. Both studies included more than 600 patients. A recent study published by Orhac et al. found that since the first published PET radiomics study in 2009, almost 80% of studies have included fewer than 100 patients [168]. Similarly, CT head and neck cancer studies often include about 200 patients. Our study included more than three times this amount for both CT and PET analyses.

While some features selected were significant in both the training and testing cohorts, none of our analyses in CT or PET studies could find a reliable radiomics feature that correlated with overall survival that was better than tumor volume. Our negative results are in contrast to other radiomics studies of head and neck cancer patients. Other studies have found radiomics features that correlated with overall survival, locoregional control, and freedom from distant metastases [17, 18, 26, 31, 171, 172]. We investigated these other outcomes as well and found similar results to the overall survival results presented here. We chose to focus on overall survival since there were more events which typically results in better model building. However, most of our patients had oropharyngeal cancer, whereas most other radiomics studies included patients with general head and neck cancers that included sites such as the larynx. Foy et al. also showed that there are differences in implementation of the various radiomics software tools [173]. All of these differences could contribute to the differences in results found in other studies compared to our study.

We attempted many manipulations of the data, including making the patient cohorts all one HPV status (e.g., positive), which removed the issue of HPV status having different rates of survival, and radiomics features were not consistently correlated with survival. Different splitting techniques of the

training and testing cohorts yielded similar, negative, results. The data between the training and testing was similar in patient number, HPV status composition, event rate, and other patient demographics.

The positive results from the inclusion of the Aerts et al. data is consistent with the positive results found by Aerts et al. [26] in their original study with these data. Patient demographics (e.g., age, stage) were similar between the Aerts et al. [26] data set and the MD Anderson data set; however, there was a difference in event rate. This, in addition to the substantially different radiomics results between the two patient cohorts, indicated that there may have been differences in the patient population that we cannot understand through this study, such as differences in the patient population as it relates to the overall health care system. This suggests that radiomics signatures may not always be transferrable due to unseen differences in patient populations.

These discrepancies were not observed with the PET data. The MD Anderson and TCIA data sets were similar in patient demographics and event rate. These data sets produced no AUC above 0.6, even when sources of noise, such as imaging protocol or HPV status, were removed. Some of these models resulted in no covariates selected, which meant that even the two included clinical variables in the first stages of the modeling were not good predictors of survival. Since PET scans are not part of the standard of care for HNSCC patients, the patient cohort in this study that underwent PET scans may not be representative of HNSCC patients in general. This could explain why the traditional strong clinical correlates of survival, tumor volume and HPV status, were not selected or significant.

There are several limitations to this study. First, there are known clinical factors that affect survival that were not included in the analysis, such as smoking pack-years. The focus of this study was to demonstrate improvement in patient outcome models when imaging protocols are controlled, not to build the best possible outcome model that would include these clinical factors. Also, in common with other radiomics studies, only the primary GTV was analyzed. In patients with HNSCC, often nodes are involved, and some nodes may be irradiated due to suspected tumor involvement without definitive confirmation on images. It is difficult to determine how to best include these data in a conventional

radiomics study such as this one. Deep learning approaches have shown promising results as a different technique to radiomics studies and may handle these challenges better [174, 175].

While the results of this particular study are negative, they highlight the areas that radiomics research should go towards in head and neck cancer patients. Our large CT study showed that the noise due to different imaging protocols can be overshadowed by noise due to differences between patient cohorts. This needs to be considered and investigated when applying radiomics signatures to patient groups from different regions with potentially different characteristics. Additionally, for PET, the lack of any texture signature correlation with overall survival outweighed the noise due to different imaging protocols. This again identifies an avenue for future studies as alternative approaches are needed, for example, deep learning or development of PET-specific features. Lastly, we showed that harmonizing imaging protocols does reduce some uncertainties in radiomics features. Reducing this source of uncertainty should make it easier to investigate other sources of uncertainty (such as differences in patient cohorts) that impact the results of radiomics studies. If these additional sources of uncertainty can also be reduced, then this harmonization of imaging protocols could result in more precise radiomics studies.

9.5 Conclusion

This is one of the largest radiomics studies in head and neck cancer patients and one of the largest PET radiomics studies in general. CT and PET-based radiomics features failed to improve survival models for head and neck cancer patients. Controlling the imaging protocol to minimize image uncertainties did not improve the radiomics models. The inconsistent CT findings here demonstrate that radiomics signatures for head and neck cancer patients may not be transferable, even when patient cohorts appear to be very similar. Head and neck cancer patient images may not have enough PET texture to be used in conventional radiomics studies.

Chapter 10 : Discussion

In this dissertation we examined uncertainties in MRI, CT, and PET that could add noise to quantitative metrics used in patient analysis and then determined if controlling these uncertainties could improve prediction accuracy in models for salivary gland function and overall survival. For each modality we identified uncertainties and were able to control for several of these and then apply patient prediction models. However, for each aim we were unable to demonstrate that there was an improvement in modeling when controlling for the uncertainties due to the imaging. The individual studies are discussed in detail in the discussion section for each respective chapter above.

For Aim 1 (MRI), the porcine phantom and synthetic images, showed that both Velocity and the in-house demons-based image registration system performed well, with all RMSEs below 3 mm. Therefore, the image registration uncertainty was quantified and we attempted to use this information in a longitudinal study that utilized deformable image registration. A cohort of 15 HNSCC patients imaged with DCE-MRI pre-, mid-, and post-treatment were used to quantify inter-algorithm reliability. Algorithms were able to determine high values from low values on DROs, but workflow differences may obscure the ability to discern values across algorithms in patients. Specifically, trends among algorithms from the same institution (institution supplied both Tofts-Kermode and extended Tofts algorithms) were consistent, but not across institutions. Therefore, translatability of DCE-MRI across algorithms is not currently feasible. Due to this and non-physiological values output from algorithms, voxel based tracking for prediction of salivary gland function was not pursued further. Thus, the Aim 1 hypothesis that “DCE-MRI parameters at pre- and mid-treatment time points are associated with normal tissue outcomes” was not proven.

For Aim 2 (CT), we identified several potential noise sources that could impact radiomics studies: artifacts, imaging protocols, and inter-scanner variability. We showed that streak artifacts affect radiomics feature values, suggesting that regions containing such artifacts should not be included in radiomics data sets. We demonstrated that a simple technique, removing the slices with the artifact,

can be used to remove up to 50% of the original GTV from the ROI while retaining similar feature values. Additionally, while the presence of bone within the image can affect some feature values, the effect is typically smaller than the spread in values in the patient population and can, therefore, be ignored. We then showed that a controlled scan can be helpful for reducing uncertainties in prospective studies, as there was statistically significantly less variability in the controlled protocol scans than in the local protocol scans. The controlled protocol reduced the total variability by over 50% compared with both local chest and local head protocol scans. It is theoretically possible to correct for the manufacturer and the individual scanner. If this were done perfectly, the imaging variability could be reduced by an additional 7-8% compared with the reduction due to implementing a controlled protocol. We also demonstrated that tube voltage did not impact most features when measured in high-textured phantom cartridges which are more representative of patient tumors. This, in conjunction with other studies [32, 34, 35, 37, 38, 150], showed that the majority of the benefit of the controlled protocols can be achieved using the reconstruction parameters. Therefore, a radiomics reconstruction can be established that reduces noise of radiomics features for patient studies, while adding no extra dose to the patient for an additional CT scan.

We then tested this reduction of noise of the radiomics features for patient studies by modeling overall survival with 726 HNSCC patients with CT images and then repeating the modeling with a subset of the patients with the same imaging protocol. While some features selected were significant in both the training and testing cohorts, none of our analyses could find a reliable radiomics feature that correlated with overall survival that was better than tumor volume. Positive results were found when Aerts et al. [26] data set was included. Patient demographics (e.g., age, stage) were similar between this data set and the MD Anderson data set; however, there was a difference in event rate. This, in addition to the substantially different radiomics results between the two patient cohorts, indicated that there may have been differences in the patient population that we cannot understand through this study, such as differences in the patient population as it relates to the overall health care

system. This suggests that radiomics signatures may not always be transferrable due to unseen differences in patient populations. Therefore, the noise due to underlying differences in patient populations from different institutions was found to be larger than the noise due to differences in imaging protocols. Thus, the Aim 2 hypothesis that “reducing the variability in CT-based radiomics features due to imaging protocols improves prediction accuracy when these features are used in HNSCC patient outcome models” was not proven.

For Aim 3 (PET), we identified imaging protocols as a potential noise source. We found that as long as reasonable parameter values were used (i.e. parameter values that are actually used in clinics), almost all features had good reliability. This shows that on a given scanner, patients scanned using different imaging protocols can be combined without adding significant noise to the patient cohort. However, inter-scanner variability was about equal to inter-patient variability. This implies that patients scanned on different vendors should be combined with caution. Then, we tested this in 686 HNSCC patients with PET images. Cox proportional hazards models were created for overall survival using all patients and a subset of patients that had the same imaging protocol on the same vendor. No significant stratification of patients into Low and High Risk was achieved for any patient cohort. The MD Anderson and TCIA data sets were similar in patient demographics and event rate. These data sets produced no AUC above 0.6, even when sources of noise, such as imaging protocol or HPV status, were removed. Some of these models resulted in no covariates selected, which meant that even the two included clinical variables in the first stages of the modeling were not good predictors of survival. Since PET scans are not part of the standard of care for HNSCC patients, the patient cohort in this study that underwent PET scans may not be representative of HNSCC patients in general. This could explain why the traditional strong clinical correlates of survival, tumor volume and HPV status, were not selected or significant. As no patient cohort demonstrated a correlation of radiomics features with survival, PET radiomics studies such as this may not be suitable for these patients. Thus, the Aim 3 hypothesis that

“reducing the variability in PET-based radiomics features due to imaging protocols improves prediction accuracy when these features are used in HNSCC patient outcome models” was not proven.

Overall, we investigated noise sources in each imaging modality that could contribute to uncertainties in patient analysis. For each modality we were able to quantify the uncertainty and determine a method to minimize its impact on the patient analysis. However, for each modality larger sources of noise were identified that caused each aim’s hypothesis to not be proven.

Future Applications

Potential future directions for each study were discussed in the respective chapter for that study. Additional potential future studies are discussed here.

For DCE-MRI, besides development of algorithms that are more robust to noise, there are several potential inter-institution and inter-algorithm studies that could be conducted based on the results from this work. First, similar studies comparing the output from different algorithms could be evaluated for different anatomical sites or different DCE-MRI metrics. A study in breast DCE-MRI was able to find some agreement between algorithms [94]. Therefore, investigating other sites provides valuable information into the issues discovered in this dissertation work. There could be large discrepancies in output for most anatomical sites or head and neck could be a difficult site for reliable results from DCE-MRI due to the many air cavities within the imaging space. If head and neck was found to be a more challenging site than others and this caused the inter-algorithm variability observed in this work to be much higher than for other sites, research into development of different sequences for quantitative analysis would be a productive avenue. Otherwise, development of consensus algorithms or standardization of algorithms by organizations such as QIBA would be a productive research path. In order to determine which path, however, additional inter-algorithm studies into other anatomical sites are needed.

Additionally, semi-quantitative metrics, such as area under the curve, could be more robust across algorithms and should be investigated. In this study we only investigated two quantitative metrics. Semi-quantitative metrics are more robust and thus, may have less inter-algorithm variability. These metrics have been found to correlate with outcomes in DCE-MRI studies before demonstrating that they are also useful [176-179]. Determination of the inter-algorithm variability of these metrics could lead to multi-institutional DCE-MRI research that can currently be conducted until such time that quantitative metrics are consistent across algorithms.

For the radiomics studies, there are several potential future applications of this research. First, the modeling structure defined here was developed with close collaboration with a biostatistician and is therefore a strong statistical approach. This modeling structure can be used for studying outcomes in other patient cohorts. NSCLC has shown better prediction accuracy in the literature than HNSCC and is one possible avenue. Using a NSCLC cohort may enable demonstration of the impact of minimizing imaging uncertainties in patient studies.

Another potential future application is the use of larger patient cohorts from more institutions and more head and neck sites. The current work was mostly composed of patients from MD Anderson. There was a demonstrable difference between patient cohorts from the two institutions in the CT analysis. Having data from more institutions from additional places in the United States and Europe can help elucidate if these are differences due to each institution or if they are regional differences. This can be expanded to other areas as well, such as Canada. The more institutions involved the easier it will be to determine if there are subregions that produce the same results, such as if all southern states in the United States had similar results. HNSCC encompasses a heterogeneous group of malignancies with HPV associated SCC involving only certain anatomical sites [180]. With the decline of tobacco use in the United States and other developed countries, there has been a decrease in oral cavity and laryngeal SCC incidence, but an increase in oropharyngeal SCC incidence with the increase in HPV incidence [180]. In Europe the overall mortality rates for HNSCC are higher with an annual incidence of 43 per 100,000

persons, where in the United States this is 15 per 100,000 persons [181]. This difference could be due different proportions HNSCC anatomical sites. Even within Europe there was a 14% difference in 5-year survival of head and neck cancer in 2004 across five countries: France, Italy, Spain, Switzerland, Belgium, and Portugal [182]. This difference was larger than the 6% difference in 1992 [182], indicating that discrepancies between areas may be growing and it may be possible to distinguish these using a large study.

Beyond elucidating the differences between institutions, a large patient cohort could allow evaluation of the different head and neck sites. The patients in the study here mostly had oropharyngeal cancer. The oropharynx may be a particularly difficult site for radiomics studies. Using a larger cohort with more diverse head and neck cancer sites could allow for studies to determine if there are differences when predicting on patients with different anatomical head and neck cancer sites, such as larynx compared to oropharynx.

Further, the conventional radiomics analysis that was conducted here may not be the best option. The studies here used features that were created for satellite images and applied them to medical images as many radiomics studies have also done. Feature generation specific for medical images is a potential research option that would allow for classical machine learning and statistical approaches, such as those used in this study. In this case the feature space and weighting of each feature is defined by the user. Creating new features could be useful as the hand-crafted methodology allows for interpretability of the modeling. Additionally, deep learning, where the feature space is not defined by the user, is a potential future research option. Deep learning has shown potential for radiomics applications [183-186]. This approach allows for inputting the whole 3D volume which may work better for head and neck patients as there is often large nodal involvement. The current work did not include analysis of the nodal involvement as it is difficult to include in conventional radiomics studies. Additionally, in head and neck treatments whole nodal levels are irradiated due to suspicion of

involvement which is difficult to include. Deep learning may allow for this information to be included due to the non-defined feature space that is used to create the model.

Lastly, the identified controlled protocol for CT scans could be established as an additional reconstruction for patients. This would allow for accrual of consistent patient images that could be assessed in several years after outcome data has been recorded for these patients. Establishing this protocol would not result in immediate studies, but provides valuable data for future studies that eliminates the imaging protocol variabilities. While reduction of this variability did not improve performance of survival models in this work, reduction of this variability improves any study. Therefore, a move toward establishing this reconstruction at different institutions allows for high-quality data in the future.

Conclusion

We were able to quantify and control for the impact of several identified imaging uncertainties that could add noise to the analysis of quantitative imaging metrics in patient cohorts. However, prediction accuracy in head and neck patient cohorts was not improved by this noise reduction due to other noise sources. Therefore, the central hypothesis that DCE-MRI parameter changes during treatment are associated with salivary gland toxicity and pre-treatment CT and PET-based radiomics features are predictive of patient outcome in HNSCC was not shown within this work.

Appendix A: Supplemental Material for Chapter 4

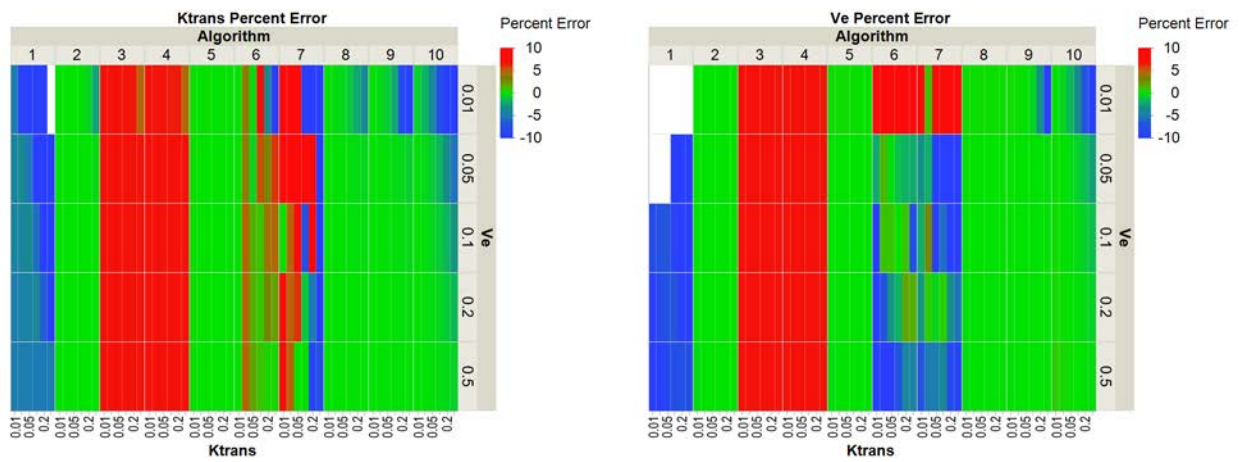


Figure A-1: Heat Maps of No Noise DRO Error.

Heat maps of the percentage error for K^{trans} (left) and v_e (right) in the DRO without noise.

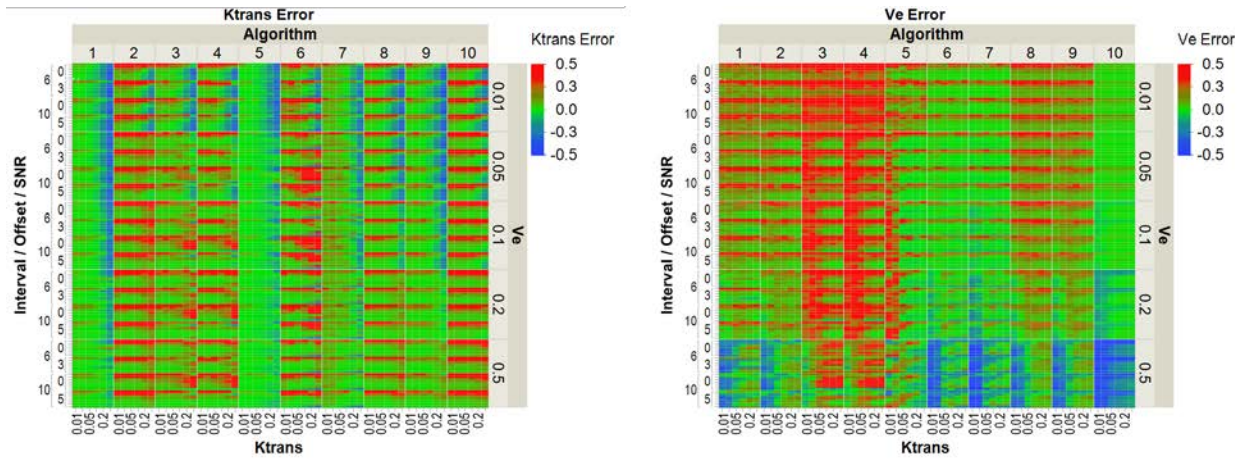


Figure A-2: Heat Maps of Noisy DRO Error.

Heat maps of the error for K^{trans} (left) and v_e (right) in the 28 DROs with noise. Maximum error is defined as 0.5, and minimum error is defined as -0.5. Any differences between the measured and simulated values greater than 0.5 are mapped to 0.5 and any differences less than -0.5 are mapped to -0.5. (See the inset in Fig. 2 for all K^{trans} and SNR values.)

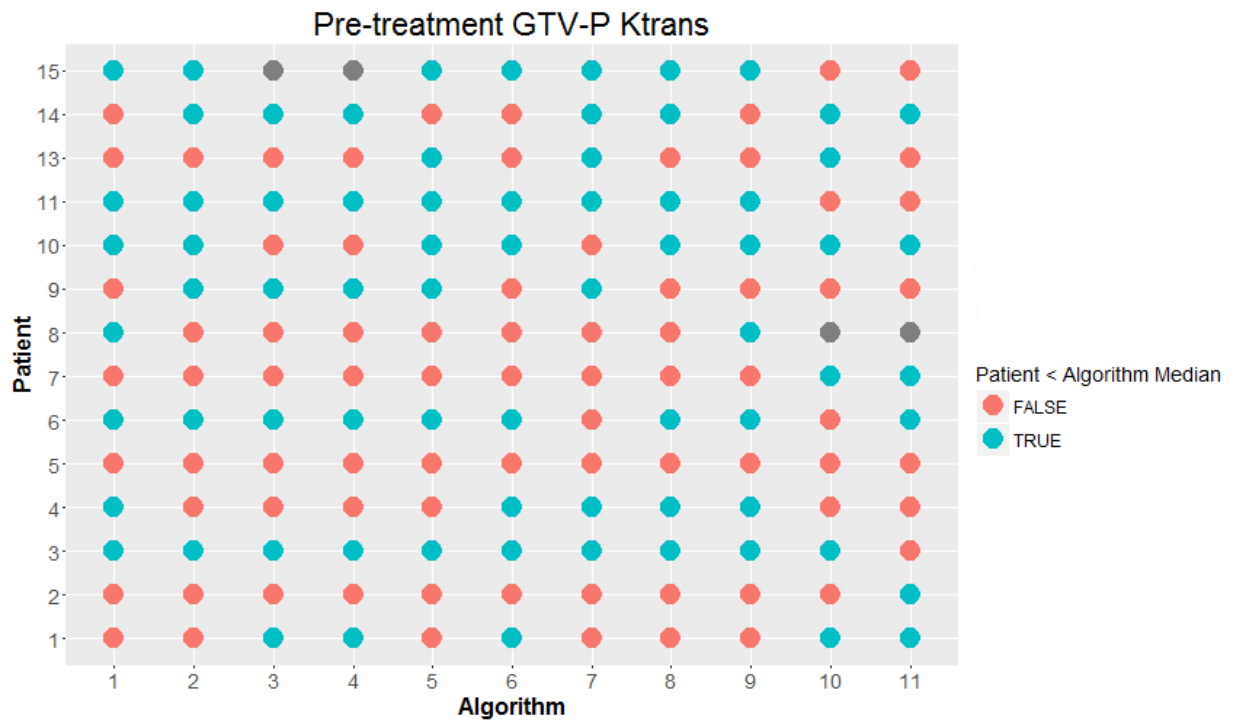


Figure A-3: Sorting of Patients Above and Below Median.

Demonstration of the Krippendorff's alpha test for pretreatment GTV-P. Red indicates that a patient's pretreatment GTV-P K^{trans} value is above the median pretreatment GTV-P K^{trans} value for that algorithm. Blue indicates that a patient's pretreatment GTV-P K^{trans} value is below the median pretreatment GTV-P K^{trans} value for that algorithm. Gray indicates that the patient's pretreatment GTV-P K^{trans} value was outside the bounds of the threshold. Overall, algorithms do not agree in classifying if a patient's pretreatment GTV-P K^{trans} value is above or below the median for all patients.

Description of MDA Model

Pre-contrast T1 maps were generated from the variable flip angle patient data. T1 values were calculated for each voxel in MATLAB (R2013a; MathWorks, Natick, MA) by performing a nonlinear curve fit (using “lsqcurvefit”) between signal intensity and the steady state signal equation for fast spoiled gradient echo sequences. Voxels with nonphysiological T1 values for soft tissue ($T1 < 0.3s$ or $T1 > 3.3s$) were flagged and excluded from further analysis. Dynamic gadolinium concentration was calculated for the population AIF and each voxel of dynamic data which exceeded a minimum signal intensity threshold, assuming a relaxivity of 3.3/mM/s for Gd-DTPA and a baseline T1 value of 1600ms for blood at 3T. A hematocrit value of 41% was assumed in these analyses.

All dynamic Gd concentration curves were trimmed to align the beginning of enhancement in AIF and tissue concentration curves and to ensure a consistent number of dynamic frames. Voxels that yielded negative or complex Gd concentration were flagged and excluded from further analysis. DCE-MRI vascular parameters were calculated by fitting dynamic data to the Tofts and extended Tofts models.

Description of MGH Model

The method for the DRO images and patient images were similar. The `mri_ms_fitparms` command from Freesurfer (https://surfer.nmr.mgh.harvard.edu/fswiki/mri_ms_fitparms) was used to create T1 maps from the provided variable flip angle files. These T1 maps then had a 2D Gaussian blur applied on each axial slice, or on just the one slice in the case of the DROs. For the patients, a population AIF was used from the AIF spreadsheet provided. For the DROs, AIFs were averaged from the provided AIF ROIs. All images except for the noiseless DRO were subject to a 2D Gaussian filter applied to axial slices. Signal intensity was converted into Gd concentration using the provided parameters for repetition time, flip angle and relaxivity. Hematocrit was assumed at 45%. T1 parameters were determined voxel-by-voxel from the T1 map. K^{trans} and v_e values were fit using the two-parameter Tofts model. Fitting was achieved via the Nelder-Mead Simplex algorithm in Matlab, with a cost-function determined by

subtracting the AUC between expected signal intensity and observed signal intensity. AUCs were determined by piecewise linear integration across the entire timespan of the scan. Voxels that fit to unreasonable K^{trans} or v_e values were set to -.01.

R code for Permutation Test

```
ord <- function(x){  
  n.row <- length(x)  
  res <- NULL  
  for (i in 1:(n.row-1)){  
    for (j in (i+1):n.row){  
      res <- c(res, ifelse(x[i]<x[j], 0, 1))  
    }  
  }  
  return(sum(res))  
}
```

#for k-trans which differs down columns

```
perm.test <- function(mat, n.sim=1000, seed=1){  
  set.seed(seed)  
  res <- numeric(n.sim)  
  for (i in 1:n.sim){  
    res[i] <- sum(apply(mat, 2, function(x){ord(sample(x))}))  
  }  
  res.obs <- sum(apply(mat, 2, function(x){ord(x)}))  
  p.value <- mean(res < res.obs) + mean(res == res.obs)/2  
}
```

```
p.value <- matrix(data=NA, nrow=noise, ncol=inst)  
for (i in 1:inst){  
  for (j in 1:noise){  
    data.sub <- data1[data1$institution==i&data1$noise==j,]  
    datasub.mat <- matrix(data.sub$measured.ktrans, nrow=length(table(data.sub$ktrans)))  
    p.value[j,i] <- perm.test(datasub.mat)
```



```

    } #institution is to identify algorithm 1 thru 11, noise is a value of 1 to 28 to identify the different noisy
    DROs
  }

```

```

#for ve which is compared across rows

```

```

perm.testrows <- function(mat, n.sim=1000, seed=1){
  set.seed(seed)
  res <- numeric(n.sim)
  for (i in 1:n.sim){
    res[i] <- sum(apply(mat, 1, function(x){ord(sample(x))}))
  }
  res.obs <- sum(apply(mat, 1, function(x){ord(x)}))
  p.value <- mean(res < res.obs) + mean(res == res.obs)/2
}

```

```

p.valueve <- matrix(data=NA, nrow=noise, ncol=inst)
for (i in 1:inst){
  for (j in 1:noise){
    data.sub <- data1[data1$institution==i&data1$noise==j,]
    datasub.mat <- matrix(data.sub$measured.ve, nrow=length(table(data.sub$ktrans)))
    p.valueve[j,i] <- perm.testrows(datasub.mat)
  }
}

```


Appendix B: Supplemental Material for Chapter 5

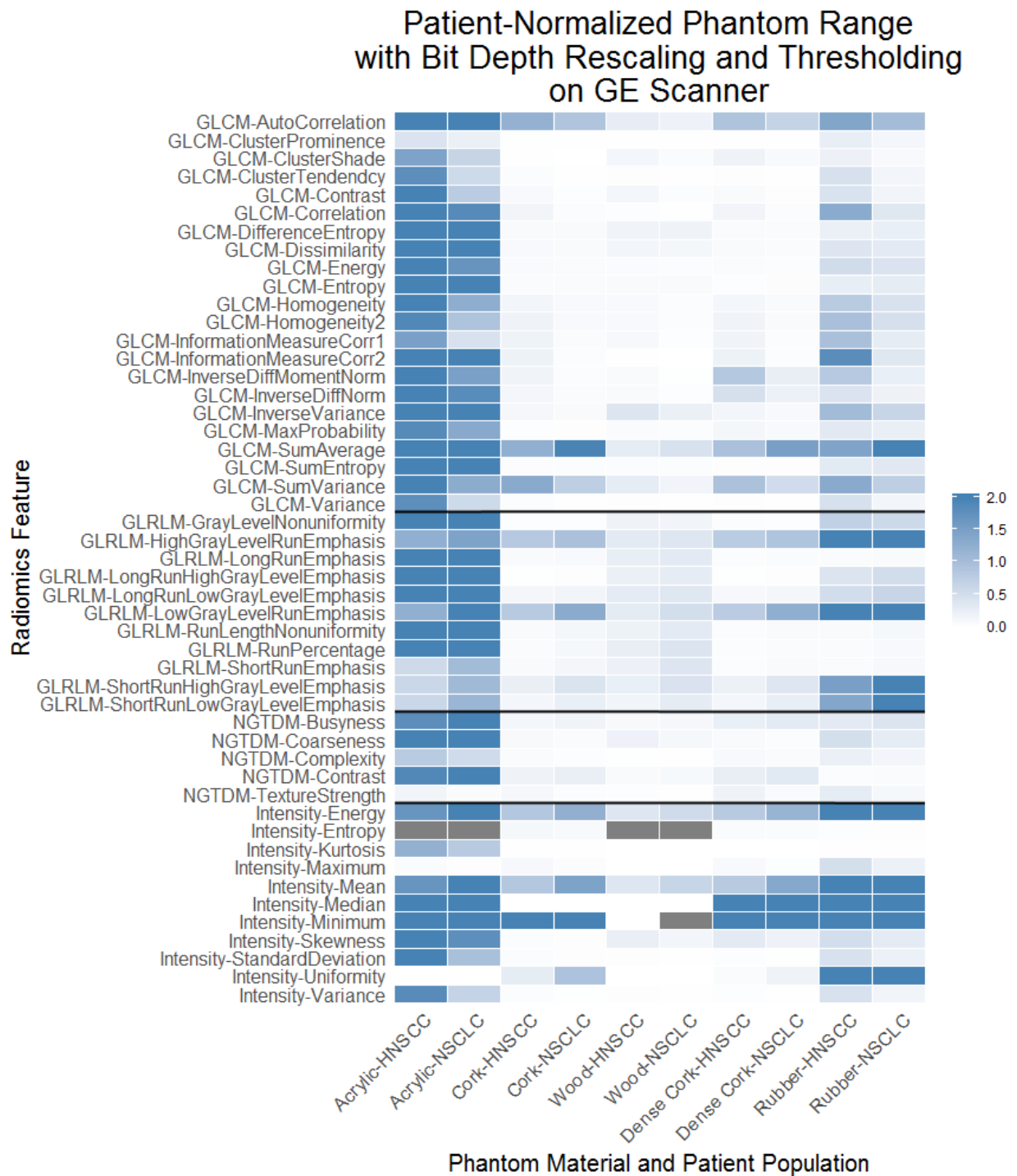


Figure B-1: Heat map of Patient-Normalized Phantom Range Values with Bit Depth Rescaling and Thresholding on GE.

The patient-normalized phantom range values are shown for each feature on the basis of each material and patient population. The values are between 0 and 2; any values above 2 were mapped to the maximum color. Gray implies that the standard deviation among the patients was 0, therefore, the

denominator of the patient-normalized phantom range was 0. The materials along the x-axis are listed in order from least to most texture on the basis of the measured standard deviation of the CT numbers of the material. The patient-normalized phantom range that was calculated using the two patient cohorts appears together for each material. Each feature along the y-axis is identified first by the acronym for the feature group that it is part of. Each feature group is shown together with black lines separating the feature groups within the heat map. Textured materials have lower patient-normalized phantom range values. In addition, the gray level run length matrix features can be seen to have the highest values in comparison to the other feature groups. GLCM: gray level co-occurrence matrix; GLRLM: gray level run length matrix; NGTDM: neighborhood gray tone difference matrix; HNSCC: head and neck squamous cell carcinoma; NSCLC: non-small cell lung cancer.

Patient-Normalized Phantom Range with Smoothing and Thresholding on GE Scanner

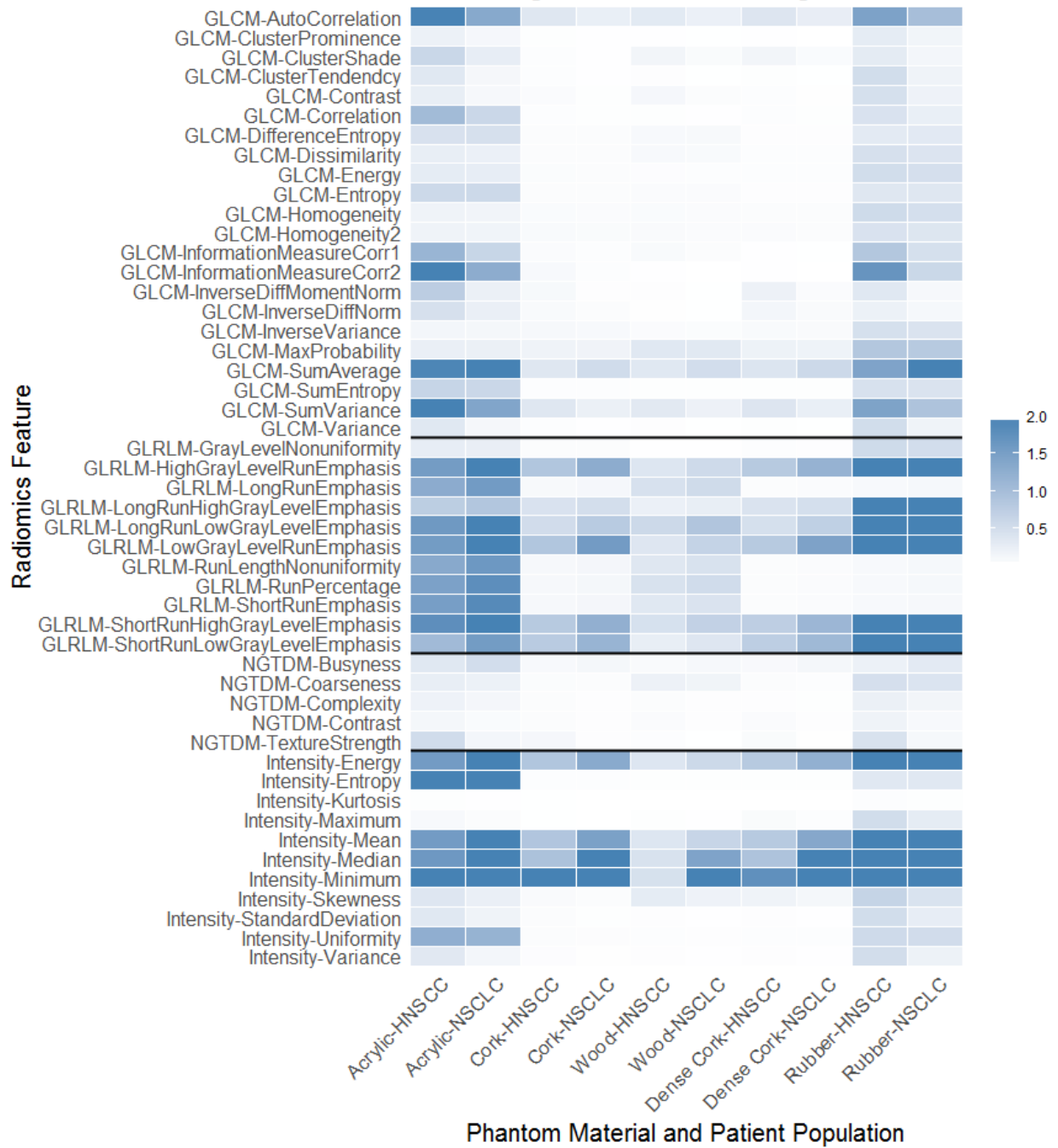


Figure B-2: Heat map of Patient-Normalized Phantom Range Values with Smoothing and Thresholding on GE.

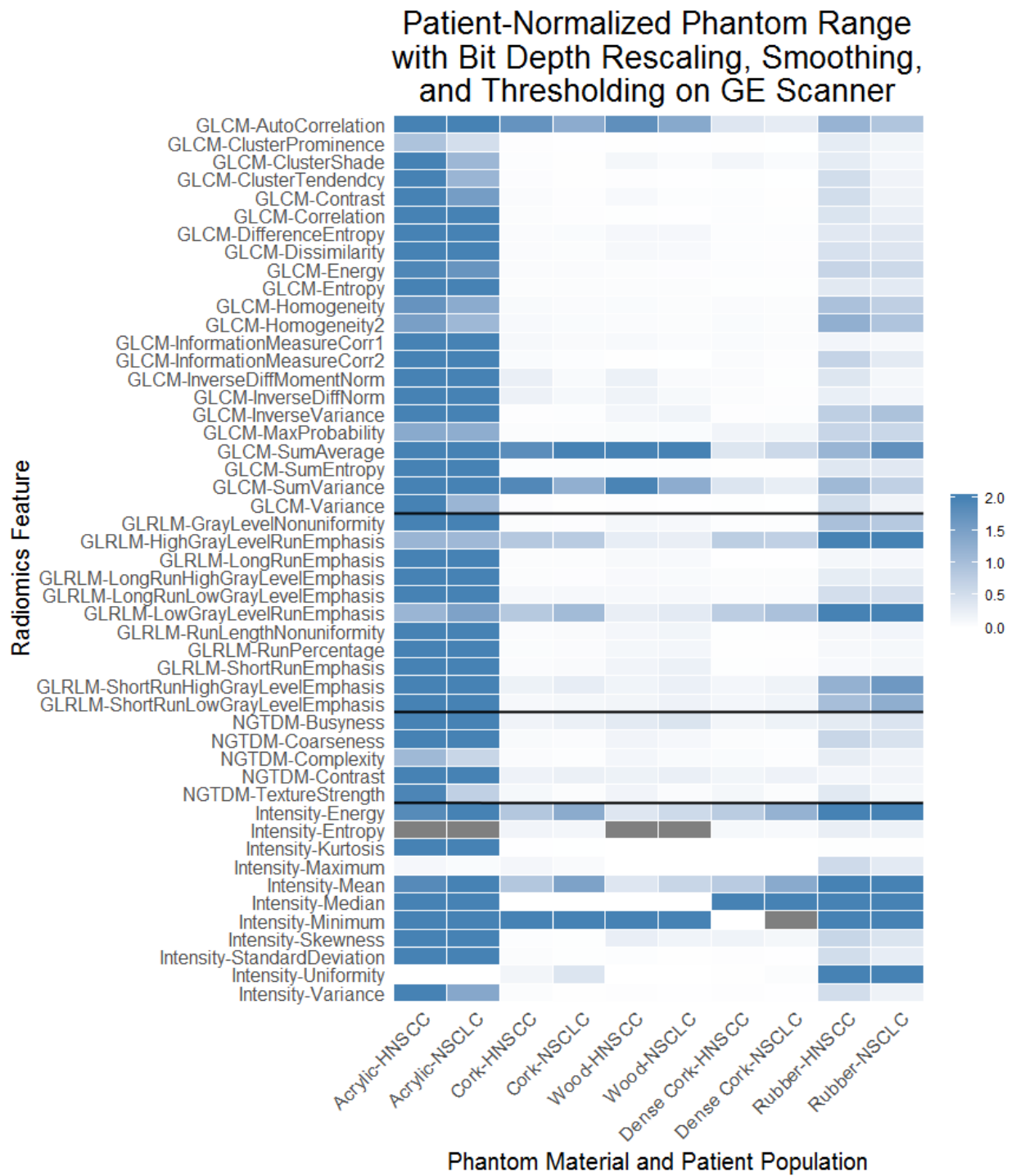


Figure B-3: Heat map of Patient-Normalized Phantom Range Values with Bit Depth Rescaling, Smoothing, and Thresholding on GE.

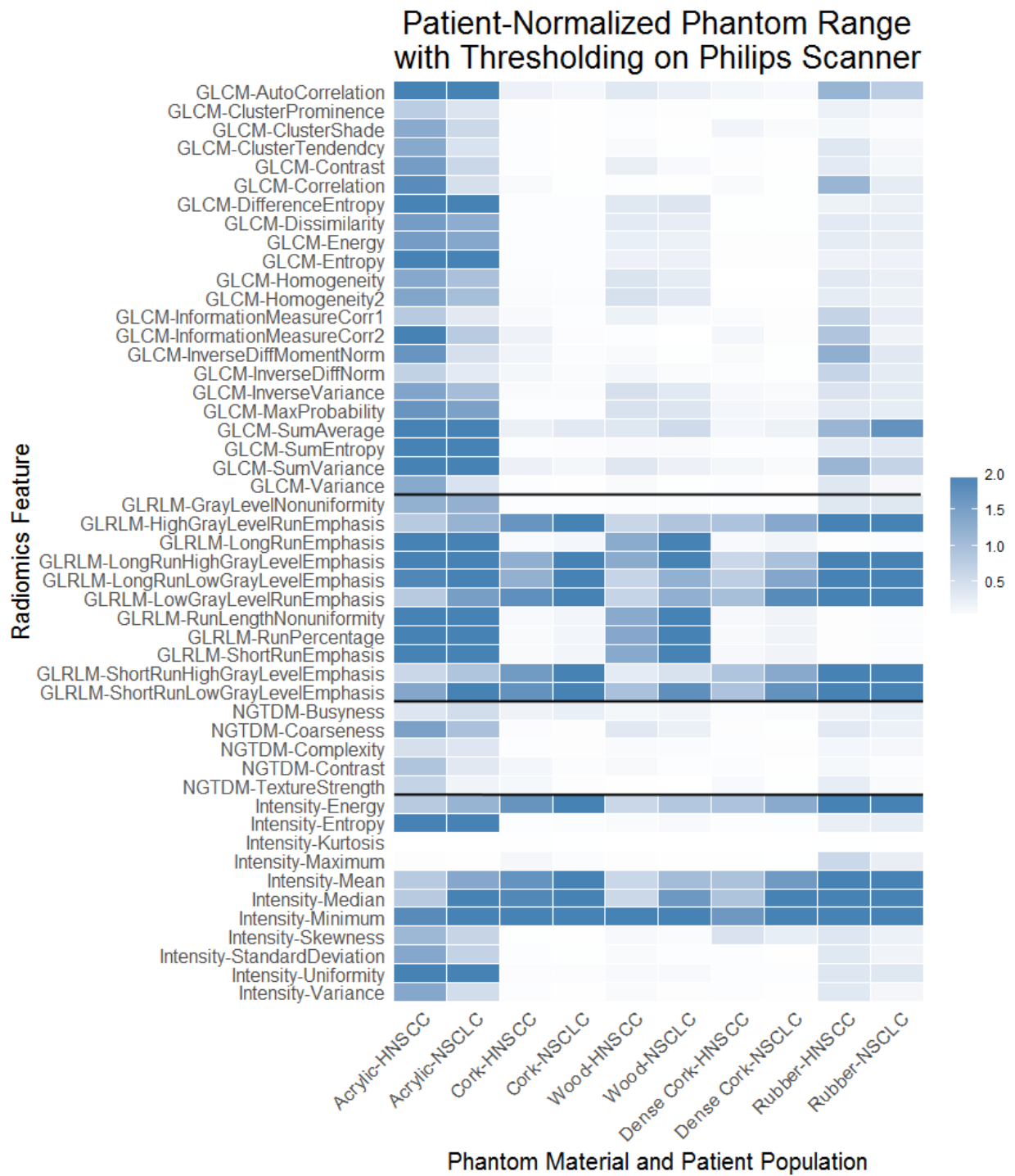


Figure B-4: Heat map of Patient-Normalized Phantom Range Values with Thresholding on Philips.

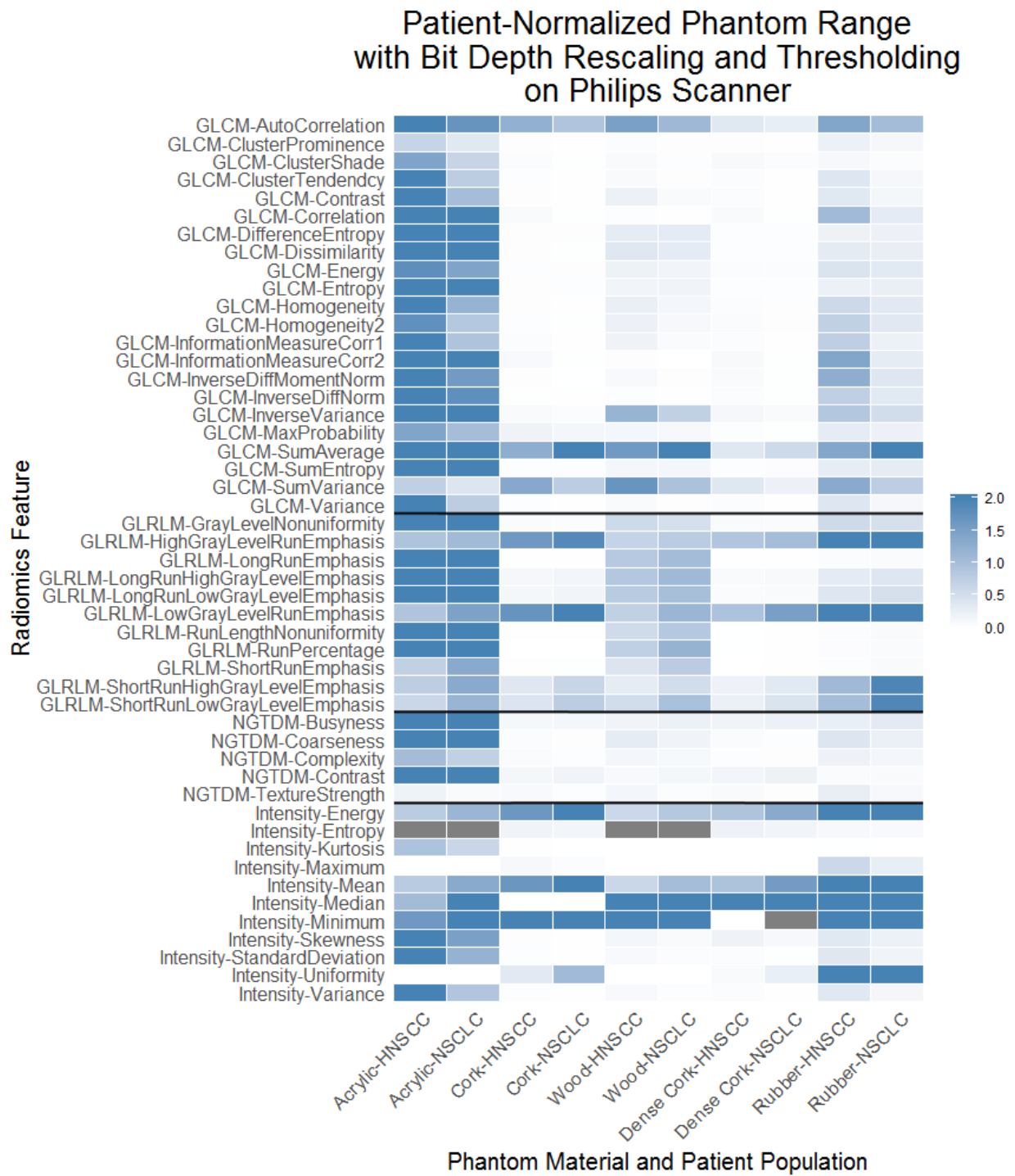


Figure B-5: Heat map of Patient-Normalized Phantom Range Values with Bit Depth Rescaling and Thresholding on Philips.

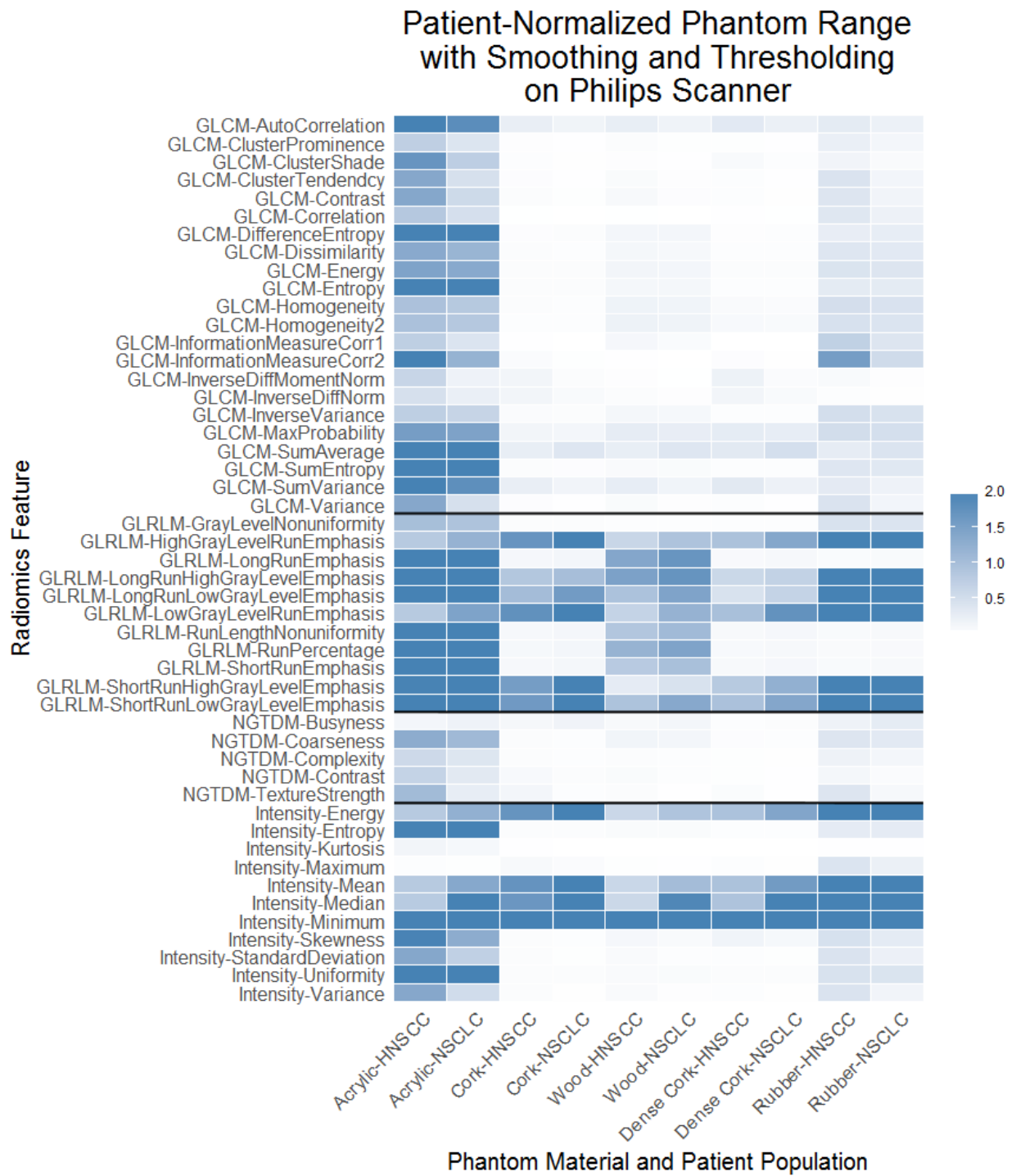


Figure B-6: Heat map of Patient-Normalized Phantom Range Values with Smoothing and Thresholding on Philips.

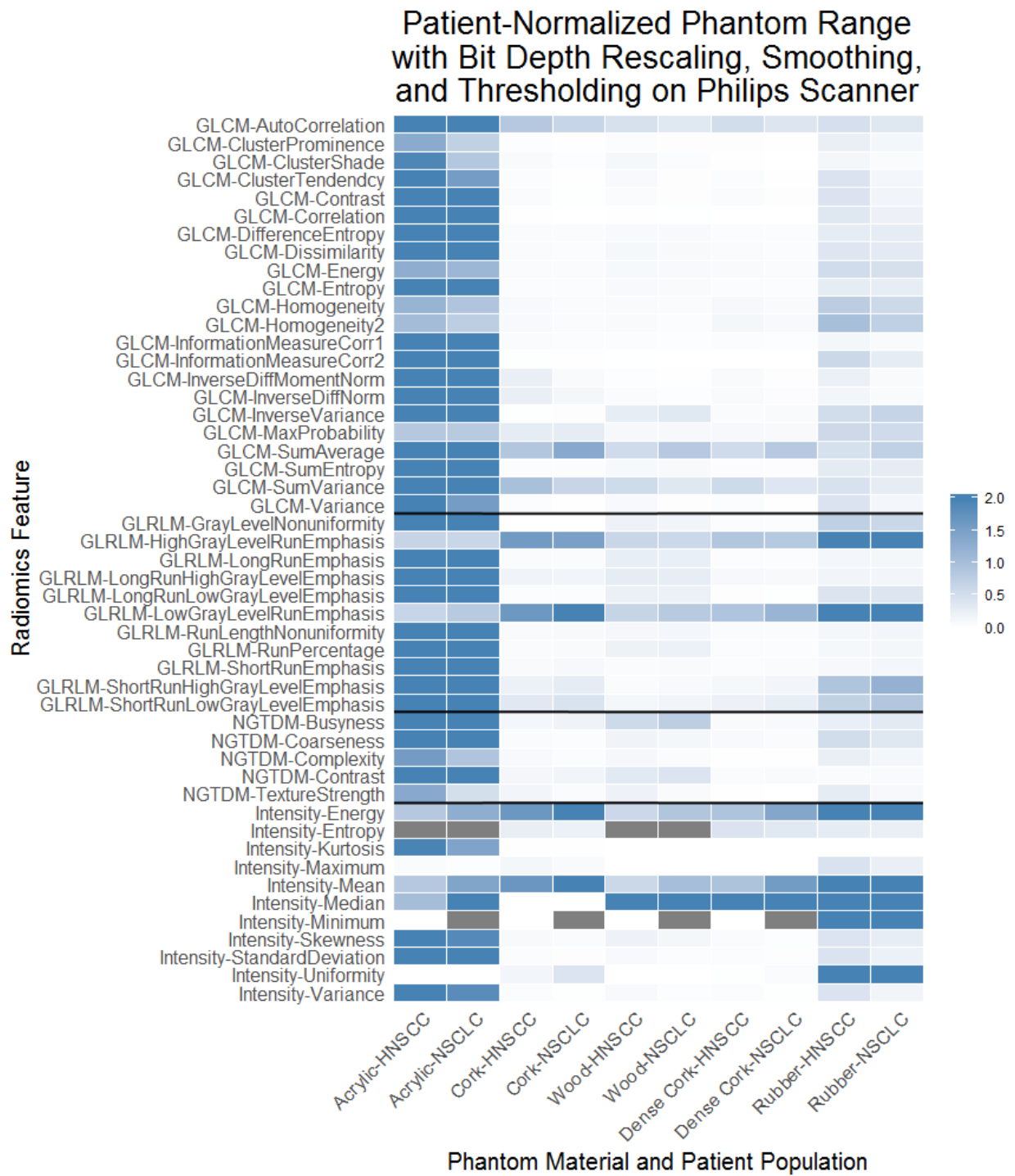


Figure B-7: Heat map of Patient-Normalized Phantom Range Values with Bit Depth Rescaling, Smoothing, and Thresholding on Philips.

Table B-1. Acrylic Spearman Rho and P Values

Preprocessing	Category	Feature	Acrylic			
			GE		Philips	
			Rho	P Value	Rho	P Value
Thresholding	GLCM	Auto Correlation	0.80	0.17	-0.50	0.48
Thresholding	GLCM	ClusterProminence	0.80	0.17	-0.50	0.48
Thresholding	GLCM	Cluster Shade	-0.40	0.49	-1.00	0.16
Thresholding	GLCM	Cluster Tendendcy	0.80	0.17	-0.50	0.48
Thresholding	GLCM	Contrast	0.80	0.17	-0.50	0.48
Thresholding	GLCM	Correlation	0.40	0.49	-1.00	0.16
Thresholding	GLCM	Difference Entropy	0.80	0.17	-0.50	0.48
Thresholding	GLCM	Dissimilarity	0.80	0.17	-0.50	0.48
Thresholding	GLCM	Energy	-0.80	0.17	0.50	0.48
Thresholding	GLCM	Entropy	0.80	0.17	-0.50	0.48
Thresholding	GLCM	Homogeneity	-0.80	0.17	0.50	0.48
Thresholding	GLCM	Homogeneity2	-0.80	0.17	0.50	0.48
		Information Measure Correlation				
Thresholding	GLCM	1	-0.40	0.49	1.00	0.16
		Information Measure Correlation				
Thresholding	GLCM	2	0.40	0.49	-1.00	0.16
		Inverse Difference				
Thresholding	GLCM	Moment Norm	0.80	0.17	0.50	0.48
		Inverse Difference				
Thresholding	GLCM	Norm	0.80	0.17	0.50	0.48
Thresholding	GLCM	Inverse Variance	-0.80	0.17	0.50	0.48
Thresholding	GLCM	Max Probability	-0.80	0.17	0.50	0.48
Thresholding	GLCM	Sum Average	0.80	0.17	-0.50	0.48
Thresholding	GLCM	Sum Entropy	0.80	0.17	-0.50	0.48
Thresholding	GLCM	Sum Variance	0.80	0.17	-0.50	0.48
Thresholding	GLCM	Variance	0.80	0.17	-0.50	0.48
		Gray Level				
Thresholding	GLRLM	Nonuniformity	-0.80	0.17	0.50	0.48
		High Gray Level Run				
Thresholding	GLRLM	Emphasis	1.00	0.08	1.00	0.16
Thresholding	GLRLM	Long Run Emphasis	-0.80	0.17	0.50	0.48
		Long Run High Gray				
Thresholding	GLRLM	Level Emphasis	0.80	0.17	0.50	0.48
		Long Run Low Gray				
Thresholding	GLRLM	Level Emphasis	-1.00	0.08	0.50	0.48
		Low Gray Level Run				
Thresholding	GLRLM	Emphasis	-1.00	0.08	-1.00	0.16
		Run Length				
Thresholding	GLRLM	Nonuniformity	0.80	0.17	-0.50	0.48
Thresholding	GLRLM	Run Percentage	0.80	0.17	-0.50	0.48

Thresholding	GLRLM	Short Run Emphasis	0.80	0.17	-0.50	0.48
Thresholding	GLRLM	Short Run High Gray				
Thresholding	GLRLM	Level Emphasis	1.00	0.08	0.50	0.48
Thresholding	GLRLM	Short Run Low Gray				
Thresholding	GLRLM	Level Emphasis	-1.00	0.08	-0.50	0.48
Thresholding	NGTDM	Busyness	-0.80	0.17	-0.50	0.48
Thresholding	NGTDM	Coarseness	-0.80	0.17	0.50	0.48
Thresholding	NGTDM	Complexity	0.80	0.17	-0.50	0.48
Thresholding	NGTDM	Contrast	-0.40	0.49	-0.50	0.48
Thresholding	NGTDM	TextureStrength	1.00	0.08	-1.00	0.16
Thresholding	IH	Energy	1.00	0.08	1.00	0.16
Thresholding	IH	Entropy	-0.80	0.17	-0.50	0.48
Thresholding	IH	Max	1.00	0.08	1.00	0.16
Thresholding	IH	Mean	1.00	0.08	1.00	0.16
Thresholding	IH	Median	1.00	0.08	1.00	0.16
Thresholding	IH	Min	1.00	0.08	1.00	0.16
Thresholding	IH	Standard Deviation	0.80	0.17	-0.50	0.48
Thresholding	IH	Uniformity	0.80	0.17	0.50	0.48
Thresholding	IH	Kurtosis	-0.40	0.49	1.00	0.16
Thresholding	IH	Skewness	-0.80	0.17	-1.00	0.16
Thresholding	IH	Variance	0.80	0.17	-0.50	0.48
Thresholding and Bit Depth						
Resampling	GLCM	Auto Correlation	-0.40	0.49	-0.50	0.48
Thresholding and Bit Depth						
Resampling	GLCM	ClusterProminence	0.00	1.00	-0.50	0.48
Thresholding and Bit Depth						
Resampling	GLCM	Cluster Shade	-0.80	0.17	-0.50	0.48
Thresholding and Bit Depth						
Resampling	GLCM	Cluster Tendendcy	0.00	1.00	-0.50	0.48
Thresholding and Bit Depth						
Resampling	GLCM	Contrast	0.00	1.00	-0.50	0.48
Thresholding and Bit Depth						
Resampling	GLCM	Correlation	-0.40	0.49	-1.00	0.16
Thresholding and Bit Depth						
Resampling	GLCM	Difference Entropy	0.00	1.00	-0.50	0.48
Thresholding and Bit Depth						
Resampling	GLCM	Dissimilarity	0.00	1.00	-0.50	0.48
Thresholding and Bit Depth						
Resampling	GLCM	Energy	0.00	1.00	0.50	0.48

Thresholding and Bit Depth Resampling	GLCM	Entropy	0.00	1.00	-0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Homogeneity	0.00	1.00	0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Homogeneity2	0.00	1.00	0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Information Measure Correlation 1	0.40	0.49	1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Information Measure Correlation 2	0.00	1.00	-1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Inverse Difference Moment Norm	0.00	1.00	0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Inverse Difference Norm	0.00	1.00	0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Inverse Variance	0.00	1.00	-0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Max Probability	0.00	1.00	0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Sum Average	-0.40	0.49	-0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Sum Entropy	0.00	1.00	-0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Sum Variance	-0.20	0.73	1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Variance	0.00	1.00	-0.50	0.48
Thresholding and Bit Depth Resampling	GLRLM	Gray Level Nonuniformity	0.00	1.00	-0.50	0.48
Thresholding and Bit Depth Resampling	GLRLM	High Gray Level Run Emphasis	1.00	0.08	1.00	0.16
Thresholding and Bit Depth Resampling	GLRLM	Long Run Emphasis	0.00	1.00	0.50	0.48

Thresholding and Bit Depth Resampling	GLRLM	Long Run High Gray Level Emphasis	0.40	0.49	0.50	0.48
Thresholding and Bit Depth Resampling	GLRLM	Long Run Low Gray Level Emphasis	0.00	1.00	0.50	0.48
Thresholding and Bit Depth Resampling	GLRLM	Low Gray Level Run Emphasis	-1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	GLRLM	Run Length Nonuniformity	0.00	1.00	-0.50	0.48
Thresholding and Bit Depth Resampling	GLRLM	Run Percentage	0.00	1.00	-0.50	0.48
Thresholding and Bit Depth Resampling	GLRLM	Short Run Emphasis	0.80	0.17	-0.50	0.48
Thresholding and Bit Depth Resampling	GLRLM	Short Run High Gray Level Emphasis	1.00	0.08	0.50	0.48
Thresholding and Bit Depth Resampling	GLRLM	Short Run Low Gray Level Emphasis	0.40	0.49	-0.50	0.48
Thresholding and Bit Depth Resampling	NGTDM	Busyness	-0.80	0.17	-0.50	0.48
Thresholding and Bit Depth Resampling	NGTDM	Coarseness	0.80	0.17	0.50	0.48
Thresholding and Bit Depth Resampling	NGTDM	Complexity	-0.80	0.17	-0.50	0.48
Thresholding and Bit Depth Resampling	NGTDM	Contrast	-0.40	0.49	-0.50	0.48
Thresholding and Bit Depth Resampling	NGTDM	TextureStrength	0.80	0.17	0.50	0.48
Thresholding and Bit Depth Resampling	IH	Energy	1.00	0.08	1.00	0.16
Thresholding and Bit Depth Resampling	IH	Entropy	#N/A	#N/A	#N/A	#N/A
Thresholding and Bit Depth Resampling	IH	Max	0.89	0.12	#N/A	#N/A

Thresholding and Bit Depth Resampling	IH	Mean	1.00	0.08	1.00	0.16
Thresholding and Bit Depth Resampling	IH	Median	0.95	0.10	0.87	0.22
Thresholding and Bit Depth Resampling	IH	Min	0.95	0.10	0.87	0.22
Thresholding and Bit Depth Resampling	IH	Standard Deviation	0.00	1.00	-0.50	0.48
Thresholding and Bit Depth Resampling	IH	Uniformity	#N/A	#N/A	#N/A	#N/A
Thresholding and Bit Depth Resampling	IH	Kurtosis	0.00	1.00	0.50	0.48
Thresholding and Bit Depth Resampling	IH	Skewness	-0.80	0.17	0.50	0.48
Thresholding and Bit Depth Resampling	IH	Variance	0.00	1.00	-0.50	0.48
Thresholding and Smoothing	GLCM	Auto Correlation	0.40	0.49	-0.50	0.48
Thresholding and Smoothing	GLCM	ClusterProminence	1.00	0.08	-0.50	0.48
Thresholding and Smoothing	GLCM	Cluster Shade	-0.80	0.17	-1.00	0.16
Thresholding and Smoothing	GLCM	Cluster Tendendcy	1.00	0.08	-0.50	0.48
Thresholding and Smoothing	GLCM	Contrast	0.80	0.17	-0.50	0.48
Thresholding and Smoothing	GLCM	Correlation	0.20	0.73	-1.00	0.16
Thresholding and Smoothing	GLCM	Difference Entropy	0.80	0.17	-0.50	0.48
Thresholding and Smoothing	GLCM	Dissimilarity	0.80	0.17	-0.50	0.48

Thresholding and Smoothing	GLCM	Energy	-1.00	0.08	0.50	0.48
Thresholding and Smoothing	GLCM	Entropy	1.00	0.08	-0.50	0.48
Thresholding and Smoothing	GLCM	Homogeneity	-0.80	0.17	0.50	0.48
Thresholding and Smoothing	GLCM	Homogeneity2	-0.80	0.17	0.50	0.48
Thresholding and Smoothing	GLCM	Information Measure Correlation 1	-0.20	0.73	0.50	0.48
Thresholding and Smoothing	GLCM	Information Measure Correlation 2	0.20	0.73	-1.00	0.16
Thresholding and Smoothing	GLCM	Inverse Difference Moment Norm	0.80	0.17	1.00	0.16
Thresholding and Smoothing	GLCM	Inverse Difference Norm	0.80	0.17	1.00	0.16
Thresholding and Smoothing	GLCM	Inverse Variance	-0.80	0.17	0.50	0.48
Thresholding and Smoothing	GLCM	Max Probability	-0.80	0.17	0.50	0.48
Thresholding and Smoothing	GLCM	Sum Average	0.40	0.49	-0.50	0.48
Thresholding and Smoothing	GLCM	Sum Entropy	1.00	0.08	-0.50	0.48
Thresholding and Smoothing	GLCM	Sum Variance	0.40	0.49	-0.50	0.48
Thresholding and Smoothing	GLCM	Variance	1.00	0.08	-0.50	0.48
Thresholding and Smoothing	GLRLM	Gray Level Nonuniformity	-1.00	0.08	0.50	0.48
Thresholding and Smoothing	GLRLM	High Gray Level Run Emphasis	1.00	0.08	1.00	0.16

Thresholding and Smoothing	GLRLM	Long Run Emphasis	-0.80	0.17	0.50	0.48
Thresholding and Smoothing	GLRLM	Long Run High Gray Level Emphasis	0.20	0.73	0.50	0.48
Thresholding and Smoothing	GLRLM	Long Run Low Gray Level Emphasis	-1.00	0.08	0.50	0.48
Thresholding and Smoothing	GLRLM	Low Gray Level Run Emphasis	-1.00	0.08	-1.00	0.16
Thresholding and Smoothing	GLRLM	Run Length Nonuniformity	0.80	0.17	-0.50	0.48
Thresholding and Smoothing	GLRLM	Run Percentage	0.80	0.17	-0.50	0.48
Thresholding and Smoothing	GLRLM	Short Run Emphasis	0.80	0.17	-0.50	0.48
Thresholding and Smoothing	GLRLM	Short Run High Gray Level Emphasis	1.00	0.08	-0.50	0.48
Thresholding and Smoothing	GLRLM	Short Run Low Gray Level Emphasis	-1.00	0.08	-0.50	0.48
Thresholding and Smoothing	NGTDM	Busyness	-1.00	0.08	0.50	0.48
Thresholding and Smoothing	NGTDM	Coarseness	-0.80	0.17	0.50	0.48
Thresholding and Smoothing	NGTDM	Complexity	1.00	0.08	-0.50	0.48
Thresholding and Smoothing	NGTDM	Contrast	-0.80	0.17	-0.50	0.48
Thresholding and Smoothing	NGTDM	TextureStrength	1.00	0.08	-0.50	0.48
Thresholding and Smoothing	IH	Energy	1.00	0.08	1.00	0.16
Thresholding and Smoothing	IH	Entropy	-0.80	0.17	-0.50	0.48

Thresholding and Smoothing	IH	Max	1.00	0.08	1.00	0.16
Thresholding and Smoothing	IH	Mean	1.00	0.08	1.00	0.16
Thresholding and Smoothing	IH	Median	1.00	0.08	1.00	0.16
Thresholding and Smoothing	IH	Min	1.00	0.08	1.00	0.16
Thresholding and Smoothing	IH	Standard Deviation	1.00	0.08	-0.50	0.48
Thresholding and Smoothing	IH	Uniformity	0.80	0.17	0.50	0.48
Thresholding and Smoothing	IH	Kurtosis	0.80	0.17	1.00	0.16
Thresholding and Smoothing	IH	Skewness	-0.40	0.49	-1.00	0.16
Thresholding and Smoothing	IH	Variance	1.00	0.08	-0.50	0.48
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Auto Correlation	0.40	0.49	1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	ClusterProminence	0.00	1.00	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Cluster Shade	-0.80	0.17	-0.50	0.48
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Cluster Tendendcy	0.00	1.00	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Contrast	0.00	1.00	-1.00	0.16
Thresholding, Smoothing,	GLCM	Correlation	0.00	1.00	-1.00	0.16

Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Sum Average	0.40	0.49	1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Sum Entropy	0.00	1.00	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Sum Variance	0.40	0.49	1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Variance	0.00	1.00	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	Gray Level Nonuniformity	0.00	1.00	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	High Gray Level Run Emphasis	1.00	0.08	1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	Long Run Emphasis	0.40	0.49	1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	Long Run High Gray Level Emphasis	0.40	0.49	1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	Long Run Low Gray Level Emphasis	0.00	1.00	1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	Low Gray Level Run Emphasis	-1.00	0.08	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	Run Length Nonuniformity	-0.40	0.49	-0.50	0.48
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	Run Percentage	0.00	1.00	-1.00	0.16
Thresholding, Smoothing,	GLRLM	Short Run Emphasis	0.80	0.17	-1.00	0.16

Thresholding, Smoothing, and Bit Depth Resampling	IH	Min	0.89	0.12	#N/A	#N/A
Thresholding, Smoothing, and Bit Depth Resampling	IH	Standard Deviation	0.00	1.00	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	IH	Uniformity	#N/A	#N/A	#N/A	#N/A
Thresholding, Smoothing, and Bit Depth Resampling	IH	Kurtosis	0.00	1.00	1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	IH	Skewness	-0.80	0.17	0.50	0.48
Thresholding, Smoothing, and Bit Depth Resampling	IH	Variance	0.00	1.00	-1.00	0.16

Table B-2. Cork Spearman Rho and P Values

Preprocessing	Category	Feature	Cork			
			GE		Philips	
			Rho	P Value	Rho	P Value
Thresholding	GLCM	Auto Correlation	-0.20	0.73	0.50	0.48
Thresholding	GLCM	ClusterProminence	0.80	0.17	-0.50	0.48
Thresholding	GLCM	Cluster Shade	0.60	0.30	-0.50	0.48
Thresholding	GLCM	Cluster Tendendcy	0.80	0.17	-1.00	0.16
Thresholding	GLCM	Contrast	0.20	0.73	-0.50	0.48
Thresholding	GLCM	Correlation	-0.20	0.73	-1.00	0.16
Thresholding	GLCM	Difference Entropy	0.20	0.73	-0.50	0.48
Thresholding	GLCM	Dissimilarity	0.20	0.73	-0.50	0.48
Thresholding	GLCM	Energy	-0.80	0.17	0.50	0.48
Thresholding	GLCM	Entropy	0.80	0.17	-0.50	0.48
Thresholding	GLCM	Homogeneity	-0.40	0.49	0.50	0.48
Thresholding	GLCM	Homogeneity2	-0.20	0.73	0.50	0.48
		Information Measure Correlation				
Thresholding	GLCM	1	0.20	0.73	1.00	0.16
		Information Measure Correlation				
Thresholding	GLCM	2	-0.20	0.73	-1.00	0.16
		Inverse Difference				
Thresholding	GLCM	Moment Norm	-0.20	0.73	0.50	0.48
		Inverse Difference				
Thresholding	GLCM	Norm	-0.20	0.73	0.50	0.48
Thresholding	GLCM	Inverse Variance	-0.80	0.17	1.00	0.16
Thresholding	GLCM	Max Probability	-0.80	0.17	0.00	1.00
Thresholding	GLCM	Sum Average	-0.20	0.73	0.50	0.48
Thresholding	GLCM	Sum Entropy	0.80	0.17	-1.00	0.16
Thresholding	GLCM	Sum Variance	-0.20	0.73	0.50	0.48
Thresholding	GLCM	Variance	0.80	0.17	-1.00	0.16
		Gray Level				
Thresholding	GLRLM	Nonuniformity	-0.80	0.17	1.00	0.16
		High Gray Level Run				
Thresholding	GLRLM	Emphasis	1.00	0.08	1.00	0.16
Thresholding	GLRLM	Long Run Emphasis	0.20	0.73	0.50	0.48
		Long Run High Gray				
Thresholding	GLRLM	Level Emphasis	1.00	0.08	1.00	0.16
		Long Run Low Gray				
Thresholding	GLRLM	Level Emphasis	-1.00	0.08	-1.00	0.16
		Low Gray Level Run				
Thresholding	GLRLM	Emphasis	-1.00	0.08	-1.00	0.16
		Run Length				
Thresholding	GLRLM	Nonuniformity	0.40	0.49	-0.50	0.48
Thresholding	GLRLM	Run Percentage	0.11	0.86	-0.50	0.48

Thresholding	GLRLM	Short Run Emphasis	0.40	0.49	-0.50	0.48
Thresholding	GLRLM	Short Run High Gray				
Thresholding	GLRLM	Level Emphasis	1.00	0.08	1.00	0.16
Thresholding	GLRLM	Short Run Low Gray				
Thresholding	GLRLM	Level Emphasis	-1.00	0.08	-1.00	0.16
Thresholding	NGTDM	Busyness	-0.20	0.73	-0.50	0.48
Thresholding	NGTDM	Coarseness	-0.20	0.73	1.00	0.16
Thresholding	NGTDM	Complexity	0.40	0.49	0.50	0.48
Thresholding	NGTDM	Contrast	0.40	0.49	-0.50	0.48
Thresholding	NGTDM	TextureStrength	0.20	0.73	0.50	0.48
Thresholding	IH	Energy	1.00	0.08	1.00	0.16
Thresholding	IH	Entropy	0.80	0.17	-0.50	0.48
Thresholding	IH	Max	0.40	0.49	0.50	0.48
Thresholding	IH	Mean	1.00	0.08	1.00	0.16
Thresholding	IH	Median	0.95	0.10	1.00	0.16
Thresholding	IH	Min	1.00	0.08	1.00	0.16
Thresholding	IH	Standard Deviation	0.80	0.17	-0.50	0.48
Thresholding	IH	Uniformity	-0.80	0.17	0.50	0.48
Thresholding	IH	Kurtosis	-0.80	0.17	0.50	0.48
Thresholding	IH	Skewness	-0.40	0.49	0.50	0.48
Thresholding	IH	Variance	0.80	0.17	-0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Auto Correlation	-0.60	0.30	-0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	ClusterProminence	0.80	0.17	-0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Cluster Shade	0.40	0.49	-0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Cluster Tendendcy	1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Contrast	0.40	0.49	-0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Correlation	-0.20	0.73	-1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Difference Entropy	0.20	0.73	-0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Dissimilarity	0.80	0.17	-0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Energy	-1.00	0.08	1.00	0.16

Thresholding and Bit Depth Resampling	GLCM	Entropy	1.00	0.08	-0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Homogeneity	-1.00	0.08	1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Homogeneity2	-1.00	0.08	1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Information Measure Correlation 1	0.40	0.49	1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Information Measure Correlation 2	-0.40	0.49	-1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Inverse Difference Moment Norm	-0.60	0.30	0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Inverse Difference Norm	-0.60	0.30	0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Inverse Variance	-0.20	0.73	0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Max Probability	-0.80	0.17	1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Sum Average	-0.60	0.30	-0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Sum Entropy	1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Sum Variance	-0.60	0.30	-0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Variance	1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	GLRLM	Gray Level Nonuniformity	-0.80	0.17	0.50	0.48
Thresholding and Bit Depth Resampling	GLRLM	High Gray Level Run Emphasis	1.00	0.08	1.00	0.16
Thresholding and Bit Depth Resampling	GLRLM	Long Run Emphasis	-0.80	0.17	-0.50	0.48

Thresholding and Bit Depth Resampling	GLRLM	Long Run High Gray Level Emphasis	-0.20	0.73	1.00	0.16
Thresholding and Bit Depth Resampling	GLRLM	Long Run Low Gray Level Emphasis	-1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	GLRLM	Low Gray Level Run Emphasis	-1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	GLRLM	Run Length Nonuniformity	0.40	0.49	-0.50	0.48
Thresholding and Bit Depth Resampling	GLRLM	Run Percentage	0.80	0.17	-1.00	0.16
Thresholding and Bit Depth Resampling	GLRLM	Short Run Emphasis	0.40	0.49	-0.50	0.48
Thresholding and Bit Depth Resampling	GLRLM	Short Run High Gray Level Emphasis	1.00	0.08	1.00	0.16
Thresholding and Bit Depth Resampling	GLRLM	Short Run Low Gray Level Emphasis	-1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	NGTDM	Busyness	0.80	0.17	0.50	0.48
Thresholding and Bit Depth Resampling	NGTDM	Coarseness	-0.40	0.49	1.00	0.16
Thresholding and Bit Depth Resampling	NGTDM	Complexity	-0.20	0.73	-1.00	0.16
Thresholding and Bit Depth Resampling	NGTDM	Contrast	0.80	0.17	0.50	0.48
Thresholding and Bit Depth Resampling	NGTDM	TextureStrength	-0.40	0.49	-1.00	0.16
Thresholding and Bit Depth Resampling	IH	Energy	1.00	0.08	1.00	0.16
Thresholding and Bit Depth Resampling	IH	Entropy	1.00	0.08	1.00	0.16
Thresholding and Bit Depth Resampling	IH	Max	0.77	0.18	0.87	0.22

Thresholding and Bit Depth Resampling	IH	Mean	1.00	0.08	1.00	0.16
Thresholding and Bit Depth Resampling	IH	Median	#N/A	#N/A	#N/A	#N/A
Thresholding and Bit Depth Resampling	IH	Min	0.89	0.12	0.87	0.22
Thresholding and Bit Depth Resampling	IH	Standard Deviation	1.00	0.08	-0.50	0.48
Thresholding and Bit Depth Resampling	IH	Uniformity	-1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	IH	Kurtosis	-0.80	0.17	0.50	0.48
Thresholding and Bit Depth Resampling	IH	Skewness	-0.40	0.49	-0.50	0.48
Thresholding and Bit Depth Resampling	IH	Variance	1.00	0.08	-0.50	0.48
Thresholding and Smoothing	GLCM	Auto Correlation	0.40	0.49	0.50	0.48
Thresholding and Smoothing	GLCM	ClusterProminence	0.80	0.17	-0.50	0.48
Thresholding and Smoothing	GLCM	Cluster Shade	0.80	0.17	-0.50	0.48
Thresholding and Smoothing	GLCM	Cluster Tendendcy	1.00	0.08	-1.00	0.16
Thresholding and Smoothing	GLCM	Contrast	0.40	0.49	-0.50	0.48
Thresholding and Smoothing	GLCM	Correlation	-0.40	0.49	-0.50	0.48
Thresholding and Smoothing	GLCM	Difference Entropy	0.40	0.49	-1.00	0.16
Thresholding and Smoothing	GLCM	Dissimilarity	0.40	0.49	-1.00	0.16

Thresholding and Smoothing	GLCM	Energy	-0.40	0.49	1.00	0.16
Thresholding and Smoothing	GLCM	Entropy	0.80	0.17	-1.00	0.16
Thresholding and Smoothing	GLCM	Homogeneity	-0.80	0.17	0.50	0.48
Thresholding and Smoothing	GLCM	Homogeneity2	-0.80	0.17	0.50	0.48
Thresholding and Smoothing	GLCM	Information Measure Correlation 1	0.20	0.73	0.50	0.48
Thresholding and Smoothing	GLCM	Information Measure Correlation 2	-0.20	0.73	-0.50	0.48
Thresholding and Smoothing	GLCM	Inverse Difference Moment Norm	-0.80	0.17	0.50	0.48
Thresholding and Smoothing	GLCM	Inverse Difference Norm	-0.80	0.17	0.50	0.48
Thresholding and Smoothing	GLCM	Inverse Variance	-0.20	0.73	1.00	0.16
Thresholding and Smoothing	GLCM	Max Probability	0.74	0.20	0.50	0.48
Thresholding and Smoothing	GLCM	Sum Average	0.40	0.49	0.50	0.48
Thresholding and Smoothing	GLCM	Sum Entropy	1.00	0.08	-1.00	0.16
Thresholding and Smoothing	GLCM	Sum Variance	0.40	0.49	0.50	0.48
Thresholding and Smoothing	GLCM	Variance	1.00	0.08	-1.00	0.16
Thresholding and Smoothing	GLRLM	Gray Level Nonuniformity	-0.40	0.49	0.50	0.48
Thresholding and Smoothing	GLRLM	High Gray Level Run Emphasis	1.00	0.08	1.00	0.16

Thresholding and Smoothing	GLRLM	Long Run Emphasis	-0.40	0.49	-1.00	0.16
Thresholding and Smoothing	GLRLM	Long Run High Gray Level Emphasis	1.00	0.08	1.00	0.16
Thresholding and Smoothing	GLRLM	Long Run Low Gray Level Emphasis	-1.00	0.08	-1.00	0.16
Thresholding and Smoothing	GLRLM	Low Gray Level Run Emphasis	-1.00	0.08	-1.00	0.16
Thresholding and Smoothing	GLRLM	Run Length Nonuniformity	0.40	0.49	1.00	0.16
Thresholding and Smoothing	GLRLM	Run Percentage	0.40	0.49	1.00	0.16
Thresholding and Smoothing	GLRLM	Short Run Emphasis	0.40	0.49	1.00	0.16
Thresholding and Smoothing	GLRLM	Short Run High Gray Level Emphasis	1.00	0.08	1.00	0.16
Thresholding and Smoothing	GLRLM	Short Run Low Gray Level Emphasis	-1.00	0.08	-1.00	0.16
Thresholding and Smoothing	NGTDM	Busyness	-0.80	0.17	-0.50	0.48
Thresholding and Smoothing	NGTDM	Coarseness	-0.40	0.49	1.00	0.16
Thresholding and Smoothing	NGTDM	Complexity	0.80	0.17	-0.50	0.48
Thresholding and Smoothing	NGTDM	Contrast	-0.60	0.30	-0.50	0.48
Thresholding and Smoothing	NGTDM	TextureStrength	0.80	0.17	0.50	0.48
Thresholding and Smoothing	IH	Energy	1.00	0.08	1.00	0.16
Thresholding and Smoothing	IH	Entropy	1.00	0.08	-1.00	0.16

Thresholding and Smoothing	IH	Max	0.00	1.00	0.87	0.22
Thresholding and Smoothing	IH	Mean	1.00	0.08	1.00	0.16
Thresholding and Smoothing	IH	Median	1.00	0.08	1.00	0.16
Thresholding and Smoothing	IH	Min	0.60	0.30	0.87	0.22
Thresholding and Smoothing	IH	Standard Deviation	0.80	0.17	-1.00	0.16
Thresholding and Smoothing	IH	Uniformity	-0.80	0.17	1.00	0.16
Thresholding and Smoothing	IH	Kurtosis	0.60	0.30	0.50	0.48
Thresholding and Smoothing	IH	Skewness	-0.20	0.73	0.50	0.48
Thresholding and Smoothing	IH	Variance	0.80	0.17	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Auto Correlation	0.40	0.49	1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	ClusterProminence	0.80	0.17	-0.50	0.48
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Cluster Shade	0.40	0.49	-0.50	0.48
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Cluster Tendendcy	1.00	0.08	-0.50	0.48
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Contrast	0.40	0.49	-0.50	0.48
Thresholding, Smoothing,	GLCM	Correlation	-0.40	0.49	-0.50	0.48

Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Sum Average	0.40	0.49	1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Sum Entropy	1.00	0.08	-0.50	0.48
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Sum Variance	0.40	0.49	1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Variance	1.00	0.08	-0.50	0.48
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	Gray Level Nonuniformity	-0.40	0.49	0.50	0.48
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	High Gray Level Run Emphasis	1.00	0.08	1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	Long Run Emphasis	-0.40	0.49	1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	Long Run High Gray Level Emphasis	0.60	0.30	1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	Long Run Low Gray Level Emphasis	-0.80	0.17	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	Low Gray Level Run Emphasis	-1.00	0.08	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	Run Length Nonuniformity	1.00	0.08	-0.50	0.48
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	Run Percentage	0.80	0.17	-0.50	0.48
Thresholding, Smoothing,	GLRLM	Short Run Emphasis	1.00	0.08	-0.50	0.48

Thresholding, Smoothing, and Bit Depth Resampling	IH	Min	0.26	0.65	#N/A	#N/A
Thresholding, Smoothing, and Bit Depth Resampling	IH	Standard Deviation	1.00	0.08	-0.50	0.48
Thresholding, Smoothing, and Bit Depth Resampling	IH	Uniformity	-1.00	0.08	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	IH	Kurtosis	0.40	0.49	0.50	0.48
Thresholding, Smoothing, and Bit Depth Resampling	IH	Skewness	-0.80	0.17	0.50	0.48
Thresholding, Smoothing, and Bit Depth Resampling	IH	Variance	1.00	0.08	-0.50	0.48

Table B-3. Dense Cork Spearman Rho and P Values

Preprocessing	Category	Feature	Dense Cork			
			GE		Philips	
			Rho	P Value	Rho	P Value
Thresholding	GLCM	Auto Correlation	0.40	0.49	1.00	0.16
Thresholding	GLCM	ClusterProminence	0.40	0.49	-1.00	0.16
Thresholding	GLCM	Cluster Shade	-1.00	0.08	-1.00	0.16
Thresholding	GLCM	Cluster Tendendcy	0.20	0.73	-1.00	0.16
Thresholding	GLCM	Contrast	0.20	0.73	-1.00	0.16
Thresholding	GLCM	Correlation	-0.20	0.73	-1.00	0.16
Thresholding	GLCM	Difference Entropy	0.20	0.73	-1.00	0.16
Thresholding	GLCM	Dissimilarity	0.20	0.73	-1.00	0.16
Thresholding	GLCM	Energy	-0.40	0.49	1.00	0.16
Thresholding	GLCM	Entropy	0.20	0.73	-1.00	0.16
Thresholding	GLCM	Homogeneity	-1.00	0.08	0.50	0.48
Thresholding	GLCM	Homogeneity2	-1.00	0.08	-0.50	0.48
		Information Measure Correlation				
Thresholding	GLCM	1	0.20	0.73	1.00	0.16
		Information Measure Correlation				
Thresholding	GLCM	2	-0.20	0.73	-1.00	0.16
		Inverse Difference				
Thresholding	GLCM	Moment Norm	-0.40	0.49	-0.50	0.48
		Inverse Difference				
Thresholding	GLCM	Norm	-0.40	0.49	-0.50	0.48
Thresholding	GLCM	Inverse Variance	-0.40	0.49	0.50	0.48
Thresholding	GLCM	Max Probability	-0.32	0.58	-0.50	0.48
Thresholding	GLCM	Sum Average	0.40	0.49	1.00	0.16
Thresholding	GLCM	Sum Entropy	0.20	0.73	-1.00	0.16
Thresholding	GLCM	Sum Variance	0.40	0.49	1.00	0.16
Thresholding	GLCM	Variance	0.20	0.73	-1.00	0.16
		Gray Level				
Thresholding	GLRLM	Nonuniformity	-0.20	0.73	1.00	0.16
		High Gray Level Run				
Thresholding	GLRLM	Emphasis	1.00	0.08	1.00	0.16
Thresholding	GLRLM	Long Run Emphasis	-0.40	0.49	-1.00	0.16
		Long Run High Gray				
Thresholding	GLRLM	Level Emphasis	1.00	0.08	1.00	0.16
		Long Run Low Gray				
Thresholding	GLRLM	Level Emphasis	-1.00	0.08	-1.00	0.16
		Low Gray Level Run				
Thresholding	GLRLM	Emphasis	-1.00	0.08	-1.00	0.16
		Run Length				
Thresholding	GLRLM	Nonuniformity	0.80	0.17	1.00	0.16
Thresholding	GLRLM	Run Percentage	0.40	0.49	1.00	0.16

Thresholding	GLRLM	Short Run Emphasis	0.80	0.17	1.00	0.16
Thresholding	GLRLM	Short Run High Gray				
Thresholding	GLRLM	Level Emphasis	1.00	0.08	1.00	0.16
Thresholding	GLRLM	Short Run Low Gray				
Thresholding	GLRLM	Level Emphasis	-1.00	0.08	-1.00	0.16
Thresholding	NGTDM	Busyness	-0.80	0.17	1.00	0.16
Thresholding	NGTDM	Coarseness	-0.20	0.73	1.00	0.16
Thresholding	NGTDM	Complexity	0.40	0.49	-1.00	0.16
Thresholding	NGTDM	Contrast	0.40	0.49	1.00	0.16
Thresholding	NGTDM	TextureStrength	0.80	0.17	-1.00	0.16
Thresholding	IH	Energy	1.00	0.08	1.00	0.16
Thresholding	IH	Entropy	0.40	0.49	-1.00	0.16
Thresholding	IH	Max	0.40	0.49	-0.50	0.48
Thresholding	IH	Mean	1.00	0.08	1.00	0.16
Thresholding	IH	Median	1.00	0.08	1.00	0.16
Thresholding	IH	Min	0.60	0.30	1.00	0.16
Thresholding	IH	Standard Deviation	0.20	0.73	-1.00	0.16
Thresholding	IH	Uniformity	-0.40	0.49	1.00	0.16
Thresholding	IH	Kurtosis	-0.80	0.17	-0.50	0.48
Thresholding	IH	Skewness	-1.00	0.08	-1.00	0.16
Thresholding	IH	Variance	0.20	0.73	-1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Auto Correlation	0.40	0.49	1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	ClusterProminence	1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Cluster Shade	-1.00	0.08	-0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Cluster Tendendcy	0.40	0.49	-1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Contrast	0.20	0.73	-1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Correlation	-0.20	0.73	-0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Difference Entropy	0.20	0.73	-1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Dissimilarity	0.20	0.73	-1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Energy	-0.20	0.73	1.00	0.16

Thresholding and Bit Depth Resampling	GLCM	Entropy	0.20	0.73	-1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Homogeneity	-0.20	0.73	1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Homogeneity2	-0.20	0.73	1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Information Measure Correlation 1	0.20	0.73	0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Information Measure Correlation 2	-0.20	0.73	-0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Inverse Difference Moment Norm	-0.40	0.49	1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Inverse Difference Norm	-0.40	0.49	1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Inverse Variance	-0.40	0.49	1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Max Probability	-0.80	0.17	0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Sum Average	0.40	0.49	1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Sum Entropy	0.20	0.73	-1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Sum Variance	0.40	0.49	1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Variance	0.40	0.49	-1.00	0.16
Thresholding and Bit Depth Resampling	GLRLM	Gray Level Nonuniformity	-0.20	0.73	1.00	0.16
Thresholding and Bit Depth Resampling	GLRLM	High Gray Level Run Emphasis	1.00	0.08	1.00	0.16
Thresholding and Bit Depth Resampling	GLRLM	Long Run Emphasis	-1.00	0.08	1.00	0.16

Thresholding and Bit Depth Resampling	GLRLM	Long Run High Gray Level Emphasis	0.80	0.17	1.00	0.16
Thresholding and Bit Depth Resampling	GLRLM	Long Run Low Gray Level Emphasis	-1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	GLRLM	Low Gray Level Run Emphasis	-1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	GLRLM	Run Length Nonuniformity	1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	GLRLM	Run Percentage	1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	GLRLM	Short Run Emphasis	1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	GLRLM	Short Run High Gray Level Emphasis	1.00	0.08	1.00	0.16
Thresholding and Bit Depth Resampling	GLRLM	Short Run Low Gray Level Emphasis	-1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	NGTDM	Busyness	0.40	0.49	-1.00	0.16
Thresholding and Bit Depth Resampling	NGTDM	Coarseness	-0.20	0.73	1.00	0.16
Thresholding and Bit Depth Resampling	NGTDM	Complexity	0.20	0.73	0.50	0.48
Thresholding and Bit Depth Resampling	NGTDM	Contrast	0.40	0.49	-1.00	0.16
Thresholding and Bit Depth Resampling	NGTDM	TextureStrength	-0.40	0.49	1.00	0.16
Thresholding and Bit Depth Resampling	IH	Energy	1.00	0.08	1.00	0.16
Thresholding and Bit Depth Resampling	IH	Entropy	-1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	IH	Max	0.00	1.00	#N/A	#N/A

Thresholding and Bit Depth Resampling	IH	Mean	1.00	0.08	1.00	0.16
Thresholding and Bit Depth Resampling	IH	Median	0.89	0.12	0.87	0.22
Thresholding and Bit Depth Resampling	IH	Min	0.26	0.65	#N/A	#N/A
Thresholding and Bit Depth Resampling	IH	Standard Deviation	0.20	0.73	-1.00	0.16
Thresholding and Bit Depth Resampling	IH	Uniformity	1.00	0.08	1.00	0.16
Thresholding and Bit Depth Resampling	IH	Kurtosis	-0.20	0.73	0.50	0.48
Thresholding and Bit Depth Resampling	IH	Skewness	-1.00	0.08	-0.50	0.48
Thresholding and Bit Depth Resampling	IH	Variance	0.20	0.73	-1.00	0.16
Thresholding and Smoothing	GLCM	Auto Correlation	0.80	0.17	-1.00	0.16
Thresholding and Smoothing	GLCM	ClusterProminence	1.00	0.08	-1.00	0.16
Thresholding and Smoothing	GLCM	Cluster Shade	-1.00	0.08	-0.50	0.48
Thresholding and Smoothing	GLCM	Cluster Tendendcy	1.00	0.08	-1.00	0.16
Thresholding and Smoothing	GLCM	Contrast	0.40	0.49	-1.00	0.16
Thresholding and Smoothing	GLCM	Correlation	-0.20	0.73	-1.00	0.16
Thresholding and Smoothing	GLCM	Difference Entropy	0.20	0.73	-1.00	0.16
Thresholding and Smoothing	GLCM	Dissimilarity	0.20	0.73	-1.00	0.16

Thresholding and Smoothing	GLCM	Energy	-0.80	0.17	1.00	0.16
Thresholding and Smoothing	GLCM	Entropy	0.80	0.17	-1.00	0.16
Thresholding and Smoothing	GLCM	Homogeneity	-0.40	0.49	1.00	0.16
Thresholding and Smoothing	GLCM	Homogeneity2	-0.40	0.49	0.50	0.48
Thresholding and Smoothing	GLCM	Information Measure Correlation 1	1.00	0.08	0.50	0.48
Thresholding and Smoothing	GLCM	Information Measure Correlation 2	-0.80	0.17	-1.00	0.16
Thresholding and Smoothing	GLCM	Inverse Difference Moment Norm	0.80	0.17	-1.00	0.16
Thresholding and Smoothing	GLCM	Inverse Difference Norm	0.80	0.17	-1.00	0.16
Thresholding and Smoothing	GLCM	Inverse Variance	-0.40	0.49	1.00	0.16
Thresholding and Smoothing	GLCM	Max Probability	-0.80	0.17	-0.50	0.48
Thresholding and Smoothing	GLCM	Sum Average	0.80	0.17	-1.00	0.16
Thresholding and Smoothing	GLCM	Sum Entropy	1.00	0.08	-1.00	0.16
Thresholding and Smoothing	GLCM	Sum Variance	0.80	0.17	-1.00	0.16
Thresholding and Smoothing	GLCM	Variance	1.00	0.08	-1.00	0.16
Thresholding and Smoothing	GLRLM	Gray Level Nonuniformity	-0.20	0.73	1.00	0.16
Thresholding and Smoothing	GLRLM	High Gray Level Run Emphasis	1.00	0.08	1.00	0.16

Thresholding and Smoothing	GLRLM	Long Run Emphasis	-0.80	0.17	0.50	0.48
Thresholding and Smoothing	GLRLM	Long Run High Gray Level Emphasis	1.00	0.08	1.00	0.16
Thresholding and Smoothing	GLRLM	Long Run Low Gray Level Emphasis	-1.00	0.08	-1.00	0.16
Thresholding and Smoothing	GLRLM	Low Gray Level Run Emphasis	-1.00	0.08	-1.00	0.16
Thresholding and Smoothing	GLRLM	Run Length Nonuniformity	0.40	0.49	-0.50	0.48
Thresholding and Smoothing	GLRLM	Run Percentage	0.80	0.17	-0.50	0.48
Thresholding and Smoothing	GLRLM	Short Run Emphasis	0.40	0.49	-0.50	0.48
Thresholding and Smoothing	GLRLM	Short Run High Gray Level Emphasis	1.00	0.08	1.00	0.16
Thresholding and Smoothing	GLRLM	Short Run Low Gray Level Emphasis	-1.00	0.08	-1.00	0.16
Thresholding and Smoothing	NGTDM	Busyness	-0.60	0.30	-0.50	0.48
Thresholding and Smoothing	NGTDM	Coarseness	-0.20	0.73	1.00	0.16
Thresholding and Smoothing	NGTDM	Complexity	0.80	0.17	-0.50	0.48
Thresholding and Smoothing	NGTDM	Contrast	-0.60	0.30	-0.50	0.48
Thresholding and Smoothing	NGTDM	TextureStrength	0.60	0.30	-0.50	0.48
Thresholding and Smoothing	IH	Energy	1.00	0.08	1.00	0.16
Thresholding and Smoothing	IH	Entropy	1.00	0.08	-1.00	0.16

Thresholding and Smoothing	IH	Max	0.95	0.10	-0.50	0.48
Thresholding and Smoothing	IH	Mean	1.00	0.08	1.00	0.16
Thresholding and Smoothing	IH	Median	1.00	0.08	1.00	0.16
Thresholding and Smoothing	IH	Min	0.32	0.58	1.00	0.16
Thresholding and Smoothing	IH	Standard Deviation	1.00	0.08	-1.00	0.16
Thresholding and Smoothing	IH	Uniformity	-0.80	0.17	1.00	0.16
Thresholding and Smoothing	IH	Kurtosis	0.40	0.49	1.00	0.16
Thresholding and Smoothing	IH	Skewness	-1.00	0.08	-1.00	0.16
Thresholding and Smoothing	IH	Variance	1.00	0.08	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Auto Correlation	1.00	0.08	1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	ClusterProminence	1.00	0.08	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Cluster Shade	-1.00	0.08	-0.50	0.48
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Cluster Tendendcy	1.00	0.08	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Contrast	0.40	0.49	-1.00	0.16
Thresholding, Smoothing,	GLCM	Correlation	-0.40	0.49	-0.50	0.48

Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Sum Average	1.00	0.08	1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Sum Entropy	1.00	0.08	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Sum Variance	1.00	0.08	1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Variance	1.00	0.08	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	Gray Level Nonuniformity	-0.40	0.49	0.50	0.48
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	High Gray Level Run Emphasis	1.00	0.08	1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	Long Run Emphasis	-0.80	0.17	1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	Long Run High Gray Level Emphasis	1.00	0.08	1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	Long Run Low Gray Level Emphasis	-1.00	0.08	-0.50	0.48
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	Low Gray Level Run Emphasis	-1.00	0.08	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	Run Length Nonuniformity	0.20	0.73	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	Run Percentage	0.40	0.49	-1.00	0.16
Thresholding, Smoothing,	GLRLM	Short Run Emphasis	0.20	0.73	-1.00	0.16

Thresholding, Smoothing, and Bit Depth Resampling	IH	Min	#N/A	#N/A	#N/A	#N/A
Thresholding, Smoothing, and Bit Depth Resampling	IH	Standard Deviation	1.00	0.08	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	IH	Uniformity	1.00	0.08	1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	IH	Kurtosis	0.20	0.73	0.50	0.48
Thresholding, Smoothing, and Bit Depth Resampling	IH	Skewness	-1.00	0.08	-0.50	0.48
Thresholding, Smoothing, and Bit Depth Resampling	IH	Variance	1.00	0.08	-1.00	0.16

Table B-4. Rubber Spearman Rho and P Values

Preprocessing	Category	Feature	Rubber			
			GE		Philips	
			Rho	P Value	Rho	P Value
Thresholding	GLCM	Auto Correlation	-1.00	0.08	-1.00	0.16
Thresholding	GLCM	ClusterProminence	-1.00	0.08	-1.00	0.16
Thresholding	GLCM	Cluster Shade	-1.00	0.08	-1.00	0.16
Thresholding	GLCM	Cluster Tendendcy	-1.00	0.08	-1.00	0.16
Thresholding	GLCM	Contrast	-1.00	0.08	-1.00	0.16
Thresholding	GLCM	Correlation	-1.00	0.08	-1.00	0.16
Thresholding	GLCM	Difference Entropy	-1.00	0.08	-1.00	0.16
Thresholding	GLCM	Dissimilarity	-1.00	0.08	-1.00	0.16
Thresholding	GLCM	Energy	1.00	0.08	1.00	0.16
Thresholding	GLCM	Entropy	-1.00	0.08	-1.00	0.16
Thresholding	GLCM	Homogeneity	1.00	0.08	1.00	0.16
Thresholding	GLCM	Homogeneity2	1.00	0.08	1.00	0.16
		Information				
Thresholding	GLCM	Measure Correlation				
		1	1.00	0.08	1.00	0.16
		Information				
		Measure Correlation				
Thresholding	GLCM	2	-1.00	0.08	-1.00	0.16
		Inverse Difference				
Thresholding	GLCM	Moment Norm	-0.80	0.17	-1.00	0.16
		Inverse Difference				
Thresholding	GLCM	Norm	-0.80	0.17	-1.00	0.16
Thresholding	GLCM	Inverse Variance	1.00	0.08	1.00	0.16
Thresholding	GLCM	Max Probability	0.63	0.27	1.00	0.16
Thresholding	GLCM	Sum Average	-1.00	0.08	-1.00	0.16
Thresholding	GLCM	Sum Entropy	-1.00	0.08	-1.00	0.16
Thresholding	GLCM	Sum Variance	-1.00	0.08	-1.00	0.16
Thresholding	GLCM	Variance	-1.00	0.08	-1.00	0.16
		Gray Level				
Thresholding	GLRLM	Nonuniformity	1.00	0.08	1.00	0.16
		High Gray Level Run				
Thresholding	GLRLM	Emphasis	-1.00	0.08	-1.00	0.16
Thresholding	GLRLM	Long Run Emphasis	-0.60	0.30	-1.00	0.16
		Long Run High Gray				
Thresholding	GLRLM	Level Emphasis	-1.00	0.08	-1.00	0.16
		Long Run Low Gray				
Thresholding	GLRLM	Level Emphasis	1.00	0.08	1.00	0.16
		Low Gray Level Run				
Thresholding	GLRLM	Emphasis	1.00	0.08	1.00	0.16
		Run Length				
Thresholding	GLRLM	Nonuniformity	0.80	0.17	0.50	0.48
Thresholding	GLRLM	Run Percentage	0.74	0.20	0.50	0.48

Thresholding	GLRLM	Short Run Emphasis	0.80	0.17	0.50	0.48
Thresholding	GLRLM	Short Run High Gray				
Thresholding	GLRLM	Level Emphasis	-1.00	0.08	-1.00	0.16
Thresholding	GLRLM	Short Run Low Gray				
Thresholding	GLRLM	Level Emphasis	1.00	0.08	1.00	0.16
Thresholding	NGTDM	Busyness	1.00	0.08	1.00	0.16
Thresholding	NGTDM	Coarseness	1.00	0.08	1.00	0.16
Thresholding	NGTDM	Complexity	-1.00	0.08	-1.00	0.16
Thresholding	NGTDM	Contrast	-1.00	0.08	-1.00	0.16
Thresholding	NGTDM	TextureStrength	-1.00	0.08	-1.00	0.16
Thresholding	IH	Energy	-1.00	0.08	-1.00	0.16
Thresholding	IH	Entropy	-1.00	0.08	-1.00	0.16
Thresholding	IH	Max	-1.00	0.08	-1.00	0.16
Thresholding	IH	Mean	-1.00	0.08	-1.00	0.16
Thresholding	IH	Median	-1.00	0.08	-1.00	0.16
Thresholding	IH	Min	-0.20	0.73	-0.50	0.48
Thresholding	IH	Standard Deviation	-1.00	0.08	-1.00	0.16
Thresholding	IH	Uniformity	1.00	0.08	1.00	0.16
Thresholding	IH	Kurtosis	1.00	0.08	1.00	0.16
Thresholding	IH	Skewness	-1.00	0.08	-1.00	0.16
Thresholding	IH	Variance	-1.00	0.08	-1.00	0.16
Thresholding and Bit Depth						
Resampling	GLCM	Auto Correlation	-0.80	0.17	-0.50	0.48
Thresholding and Bit Depth						
Resampling	GLCM	ClusterProminence	-1.00	0.08	-1.00	0.16
Thresholding and Bit Depth						
Resampling	GLCM	Cluster Shade	-1.00	0.08	-1.00	0.16
Thresholding and Bit Depth						
Resampling	GLCM	Cluster Tendendcy	-1.00	0.08	-1.00	0.16
Thresholding and Bit Depth						
Resampling	GLCM	Contrast	-1.00	0.08	-1.00	0.16
Thresholding and Bit Depth						
Resampling	GLCM	Correlation	-1.00	0.08	-1.00	0.16
Thresholding and Bit Depth						
Resampling	GLCM	Difference Entropy	-1.00	0.08	-1.00	0.16
Thresholding and Bit Depth						
Resampling	GLCM	Dissimilarity	-1.00	0.08	-1.00	0.16
Thresholding and Bit Depth						
Resampling	GLCM	Energy	1.00	0.08	1.00	0.16

Thresholding and Bit Depth Resampling	GLCM	Entropy	-1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Homogeneity	1.00	0.08	1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Homogeneity2	1.00	0.08	1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Information Measure Correlation 1	1.00	0.08	1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Information Measure Correlation 2	-1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Inverse Difference Moment Norm	-0.80	0.17	-1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Inverse Difference Norm	-0.80	0.17	-1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Inverse Variance	1.00	0.08	1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Max Probability	0.40	0.49	1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Sum Average	-0.80	0.17	-0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Sum Entropy	-1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Sum Variance	-0.80	0.17	-0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Variance	-1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	GLRLM	Gray Level Nonuniformity	1.00	0.08	1.00	0.16
Thresholding and Bit Depth Resampling	GLRLM	High Gray Level Run Emphasis	-1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	GLRLM	Long Run Emphasis	1.00	0.08	1.00	0.16

Thresholding and Bit Depth Resampling	GLRLM	Long Run High Gray Level Emphasis	-1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	GLRLM	Long Run Low Gray Level Emphasis	1.00	0.08	1.00	0.16
Thresholding and Bit Depth Resampling	GLRLM	Low Gray Level Run Emphasis	1.00	0.08	1.00	0.16
Thresholding and Bit Depth Resampling	GLRLM	Run Length Nonuniformity	-1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	GLRLM	Run Percentage	-1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	GLRLM	Short Run Emphasis	-1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	GLRLM	Short Run High Gray Level Emphasis	-1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	GLRLM	Short Run Low Gray Level Emphasis	1.00	0.08	1.00	0.16
Thresholding and Bit Depth Resampling	NGTDM	Busyness	1.00	0.08	1.00	0.16
Thresholding and Bit Depth Resampling	NGTDM	Coarseness	1.00	0.08	1.00	0.16
Thresholding and Bit Depth Resampling	NGTDM	Complexity	-1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	NGTDM	Contrast	-0.80	0.17	-0.50	0.48
Thresholding and Bit Depth Resampling	NGTDM	TextureStrength	-1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	IH	Energy	-1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	IH	Entropy	0.80	0.17	1.00	0.16
Thresholding and Bit Depth Resampling	IH	Max	-1.00	0.08	-1.00	0.16

Thresholding and Bit Depth Resampling	IH	Mean	-1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	IH	Median	-1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	IH	Min	-0.20	0.73	0.00	1.00
Thresholding and Bit Depth Resampling	IH	Standard Deviation	-1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	IH	Uniformity	-1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	IH	Kurtosis	1.00	0.08	0.50	0.48
Thresholding and Bit Depth Resampling	IH	Skewness	-1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	IH	Variance	-1.00	0.08	-1.00	0.16
Thresholding and Smoothing	GLCM	Auto Correlation	-0.80	0.17	-1.00	0.16
Thresholding and Smoothing	GLCM	ClusterProminence	-1.00	0.08	-1.00	0.16
Thresholding and Smoothing	GLCM	Cluster Shade	-1.00	0.08	-1.00	0.16
Thresholding and Smoothing	GLCM	Cluster Tendendcy	-1.00	0.08	-1.00	0.16
Thresholding and Smoothing	GLCM	Contrast	-1.00	0.08	-1.00	0.16
Thresholding and Smoothing	GLCM	Correlation	-1.00	0.08	-1.00	0.16
Thresholding and Smoothing	GLCM	Difference Entropy	-1.00	0.08	-1.00	0.16
Thresholding and Smoothing	GLCM	Dissimilarity	-1.00	0.08	-1.00	0.16

Thresholding and Smoothing	GLCM	Energy	1.00	0.08	1.00	0.16
Thresholding and Smoothing	GLCM	Entropy	-1.00	0.08	-1.00	0.16
Thresholding and Smoothing	GLCM	Homogeneity	1.00	0.08	1.00	0.16
Thresholding and Smoothing	GLCM	Homogeneity2	1.00	0.08	1.00	0.16
Thresholding and Smoothing	GLCM	Information Measure Correlation 1	1.00	0.08	1.00	0.16
Thresholding and Smoothing	GLCM	Information Measure Correlation 2	-1.00	0.08	-1.00	0.16
Thresholding and Smoothing	GLCM	Inverse Difference Moment Norm	-0.80	0.17	-1.00	0.16
Thresholding and Smoothing	GLCM	Inverse Difference Norm	-0.80	0.17	-1.00	0.16
Thresholding and Smoothing	GLCM	Inverse Variance	1.00	0.08	1.00	0.16
Thresholding and Smoothing	GLCM	Max Probability	0.80	0.17	1.00	0.16
Thresholding and Smoothing	GLCM	Sum Average	-0.80	0.17	-1.00	0.16
Thresholding and Smoothing	GLCM	Sum Entropy	-1.00	0.08	-1.00	0.16
Thresholding and Smoothing	GLCM	Sum Variance	-0.80	0.17	-1.00	0.16
Thresholding and Smoothing	GLCM	Variance	-1.00	0.08	-1.00	0.16
Thresholding and Smoothing	GLRLM	Gray Level Nonuniformity	1.00	0.08	1.00	0.16
Thresholding and Smoothing	GLRLM	High Gray Level Run Emphasis	-1.00	0.08	-1.00	0.16

Thresholding and Smoothing	GLRLM	Long Run Emphasis	0.80	0.17	0.50	0.48
Thresholding and Smoothing	GLRLM	Long Run High Gray Level Emphasis	-1.00	0.08	-1.00	0.16
Thresholding and Smoothing	GLRLM	Long Run Low Gray Level Emphasis	1.00	0.08	1.00	0.16
Thresholding and Smoothing	GLRLM	Low Gray Level Run Emphasis	1.00	0.08	1.00	0.16
Thresholding and Smoothing	GLRLM	Run Length Nonuniformity	-0.80	0.17	-0.50	0.48
Thresholding and Smoothing	GLRLM	Run Percentage	-0.80	0.17	-0.50	0.48
Thresholding and Smoothing	GLRLM	Short Run Emphasis	-0.80	0.17	-0.50	0.48
Thresholding and Smoothing	GLRLM	Short Run High Gray Level Emphasis	-1.00	0.08	-1.00	0.16
Thresholding and Smoothing	GLRLM	Short Run Low Gray Level Emphasis	1.00	0.08	1.00	0.16
Thresholding and Smoothing	NGTDM	Busyness	1.00	0.08	0.50	0.48
Thresholding and Smoothing	NGTDM	Coarseness	1.00	0.08	1.00	0.16
Thresholding and Smoothing	NGTDM	Complexity	-1.00	0.08	-1.00	0.16
Thresholding and Smoothing	NGTDM	Contrast	-0.80	0.17	-1.00	0.16
Thresholding and Smoothing	NGTDM	TextureStrength	-1.00	0.08	-1.00	0.16
Thresholding and Smoothing	IH	Energy	-1.00	0.08	-1.00	0.16
Thresholding and Smoothing	IH	Entropy	-1.00	0.08	-1.00	0.16

Thresholding and Smoothing	IH	Max	-1.00	0.08	-1.00	0.16
Thresholding and Smoothing	IH	Mean	-1.00	0.08	-1.00	0.16
Thresholding and Smoothing	IH	Median	-1.00	0.08	-1.00	0.16
Thresholding and Smoothing	IH	Min	-0.80	0.17	-1.00	0.16
Thresholding and Smoothing	IH	Standard Deviation	-1.00	0.08	-1.00	0.16
Thresholding and Smoothing	IH	Uniformity	1.00	0.08	1.00	0.16
Thresholding and Smoothing	IH	Kurtosis	1.00	0.08	1.00	0.16
Thresholding and Smoothing	IH	Skewness	-1.00	0.08	-1.00	0.16
Thresholding and Smoothing	IH	Variance	-1.00	0.08	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Auto Correlation	-1.00	0.08	-0.50	0.48
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	ClusterProminence	-1.00	0.08	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Cluster Shade	-1.00	0.08	-0.50	0.48
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Cluster Tendendcy	-1.00	0.08	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Contrast	-1.00	0.08	-1.00	0.16
Thresholding, Smoothing,	GLCM	Correlation	-1.00	0.08	-1.00	0.16

Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Sum Average	-1.00	0.08	-0.50	0.48
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Sum Entropy	-1.00	0.08	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Sum Variance	-1.00	0.08	-0.50	0.48
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Variance	-1.00	0.08	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	Gray Level Nonuniformity	1.00	0.08	1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	High Gray Level Run Emphasis	-1.00	0.08	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	Long Run Emphasis	1.00	0.08	1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	Long Run High Gray Level Emphasis	-1.00	0.08	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	Long Run Low Gray Level Emphasis	1.00	0.08	1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	Low Gray Level Run Emphasis	1.00	0.08	1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	Run Length Nonuniformity	-1.00	0.08	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	Run Percentage	-1.00	0.08	-1.00	0.16
Thresholding, Smoothing,	GLRLM	Short Run Emphasis	-1.00	0.08	-1.00	0.16

Thresholding, Smoothing, and Bit Depth Resampling	IH	Min	-0.95	0.10	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	IH	Standard Deviation	-1.00	0.08	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	IH	Uniformity	-1.00	0.08	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	IH	Kurtosis	1.00	0.08	0.50	0.48
Thresholding, Smoothing, and Bit Depth Resampling	IH	Skewness	-1.00	0.08	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	IH	Variance	-1.00	0.08	-1.00	0.16

Table B-5. Wood Spearman Rho and P Values

Preprocessing	Category	Feature	Wood			
			GE		Philips	
			Rho	P Value	Rho	P Value
Thresholding	GLCM	Auto Correlation	0.80	0.17	-1.00	0.16
Thresholding	GLCM	ClusterProminence	1.00	0.08	-1.00	0.16
Thresholding	GLCM	Cluster Shade	-0.40	0.49	0.50	0.48
Thresholding	GLCM	Cluster Tendendcy	1.00	0.08	-1.00	0.16
Thresholding	GLCM	Contrast	0.40	0.49	-0.50	0.48
Thresholding	GLCM	Correlation	-0.40	0.49	0.50	0.48
Thresholding	GLCM	Difference Entropy	0.80	0.17	-0.50	0.48
Thresholding	GLCM	Dissimilarity	0.80	0.17	-0.50	0.48
Thresholding	GLCM	Energy	-1.00	0.08	0.50	0.48
Thresholding	GLCM	Entropy	0.80	0.17	-0.50	0.48
Thresholding	GLCM	Homogeneity	-0.80	0.17	0.50	0.48
Thresholding	GLCM	Homogeneity2	-1.00	0.08	0.50	0.48
		Information				
Thresholding	GLCM	Measure Correlation				
		1	0.80	0.17	-0.50	0.48
		Information				
		Measure Correlation				
Thresholding	GLCM	2	-0.80	0.17	0.50	0.48
		Inverse Difference				
Thresholding	GLCM	Moment Norm	0.20	0.73	0.50	0.48
		Inverse Difference				
Thresholding	GLCM	Norm	0.00	1.00	0.50	0.48
Thresholding	GLCM	Inverse Variance	-1.00	0.08	0.50	0.48
Thresholding	GLCM	Max Probability	-0.40	0.49	1.00	0.16
Thresholding	GLCM	Sum Average	0.80	0.17	-1.00	0.16
Thresholding	GLCM	Sum Entropy	0.20	0.73	-1.00	0.16
Thresholding	GLCM	Sum Variance	0.80	0.17	-1.00	0.16
Thresholding	GLCM	Variance	1.00	0.08	-1.00	0.16
		Gray Level				
Thresholding	GLRLM	Nonuniformity	-0.40	0.49	1.00	0.16
		High Gray Level Run				
Thresholding	GLRLM	Emphasis	1.00	0.08	1.00	0.16
Thresholding	GLRLM	Long Run Emphasis	-0.80	0.17	0.50	0.48
		Long Run High Gray				
Thresholding	GLRLM	Level Emphasis	0.20	0.73	1.00	0.16
		Long Run Low Gray				
Thresholding	GLRLM	Level Emphasis	-0.80	0.17	0.50	0.48
		Low Gray Level Run				
Thresholding	GLRLM	Emphasis	-1.00	0.08	-1.00	0.16
		Run Length				
Thresholding	GLRLM	Nonuniformity	0.80	0.17	-0.50	0.48
Thresholding	GLRLM	Run Percentage	0.80	0.17	-0.50	0.48

Thresholding	GLRLM	Short Run Emphasis	0.80	0.17	-0.50	0.48
Thresholding	GLRLM	Short Run High Gray				
Thresholding	GLRLM	Level Emphasis	1.00	0.08	1.00	0.16
Thresholding	GLRLM	Short Run Low Gray				
Thresholding	GLRLM	Level Emphasis	-0.80	0.17	-1.00	0.16
Thresholding	NGTDM	Busyness	0.80	0.17	-1.00	0.16
Thresholding	NGTDM	Coarseness	-0.40	0.49	0.50	0.48
Thresholding	NGTDM	Complexity	0.40	0.49	-0.50	0.48
Thresholding	NGTDM	Contrast	0.80	0.17	-0.50	0.48
Thresholding	NGTDM	TextureStrength	-0.80	0.17	-0.50	0.48
Thresholding	IH	Energy	1.00	0.08	1.00	0.16
Thresholding	IH	Entropy	0.20	0.73	-1.00	0.16
Thresholding	IH	Max	0.63	0.27	1.00	0.16
Thresholding	IH	Mean	1.00	0.08	1.00	0.16
Thresholding	IH	Median	1.00	0.08	1.00	0.16
Thresholding	IH	Min	-0.63	0.27	1.00	0.16
Thresholding	IH	Standard Deviation	1.00	0.08	-1.00	0.16
Thresholding	IH	Uniformity	-0.40	0.49	1.00	0.16
Thresholding	IH	Kurtosis	-0.40	0.49	-0.50	0.48
Thresholding	IH	Skewness	-0.40	0.49	-0.50	0.48
Thresholding	IH	Variance	1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Auto Correlation	1.00	0.08	-0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	ClusterProminence	0.80	0.17	-1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Cluster Shade	-0.40	0.49	1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Cluster Tendendcy	1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Contrast	0.40	0.49	-0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Correlation	-0.40	0.49	0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Difference Entropy	0.40	0.49	-0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Dissimilarity	0.80	0.17	-0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Energy	-0.80	0.17	0.50	0.48

Thresholding and Bit Depth Resampling	GLCM	Entropy	0.40	0.49	-0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Homogeneity	-0.80	0.17	0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Homogeneity2	-0.80	0.17	0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Information Measure Correlation 1	0.40	0.49	-0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Information Measure Correlation 2	-0.40	0.49	0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Inverse Difference Moment Norm	-0.40	0.49	-1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Inverse Difference Norm	-0.80	0.17	-0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Inverse Variance	1.00	0.08	-0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Max Probability	-0.40	0.49	0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Sum Average	1.00	0.08	-0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Sum Entropy	0.80	0.17	-1.00	0.16
Thresholding and Bit Depth Resampling	GLCM	Sum Variance	1.00	0.08	-0.50	0.48
Thresholding and Bit Depth Resampling	GLCM	Variance	1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	GLRLM	Gray Level Nonuniformity	0.80	0.17	-0.50	0.48
Thresholding and Bit Depth Resampling	GLRLM	High Gray Level Run Emphasis	1.00	0.08	1.00	0.16
Thresholding and Bit Depth Resampling	GLRLM	Long Run Emphasis	-1.00	0.08	0.50	0.48

Thresholding and Bit Depth Resampling	GLRLM	Long Run High Gray Level Emphasis	-1.00	0.08	0.50	0.48
Thresholding and Bit Depth Resampling	GLRLM	Long Run Low Gray Level Emphasis	-1.00	0.08	0.50	0.48
Thresholding and Bit Depth Resampling	GLRLM	Low Gray Level Run Emphasis	-1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	GLRLM	Run Length Nonuniformity	0.80	0.17	-0.50	0.48
Thresholding and Bit Depth Resampling	GLRLM	Run Percentage	1.00	0.08	-0.50	0.48
Thresholding and Bit Depth Resampling	GLRLM	Short Run Emphasis	0.40	0.49	-0.50	0.48
Thresholding and Bit Depth Resampling	GLRLM	Short Run High Gray Level Emphasis	0.40	0.49	-0.50	0.48
Thresholding and Bit Depth Resampling	GLRLM	Short Run Low Gray Level Emphasis	0.40	0.49	-0.50	0.48
Thresholding and Bit Depth Resampling	NGTDM	Busyness	0.40	0.49	0.50	0.48
Thresholding and Bit Depth Resampling	NGTDM	Coarseness	-0.40	0.49	0.50	0.48
Thresholding and Bit Depth Resampling	NGTDM	Complexity	0.40	0.49	-0.50	0.48
Thresholding and Bit Depth Resampling	NGTDM	Contrast	0.80	0.17	0.50	0.48
Thresholding and Bit Depth Resampling	NGTDM	TextureStrength	-0.40	0.49	-1.00	0.16
Thresholding and Bit Depth Resampling	IH	Energy	1.00	0.08	1.00	0.16
Thresholding and Bit Depth Resampling	IH	Entropy	#N/A	#N/A	#N/A	#N/A
Thresholding and Bit Depth Resampling	IH	Max	#N/A	#N/A	#N/A	#N/A

Thresholding and Bit Depth Resampling	IH	Mean	1.00	0.08	1.00	0.16
Thresholding and Bit Depth Resampling	IH	Median	#N/A	#N/A	0.87	0.22
Thresholding and Bit Depth Resampling	IH	Min	#N/A	#N/A	0.87	0.22
Thresholding and Bit Depth Resampling	IH	Standard Deviation	1.00	0.08	-1.00	0.16
Thresholding and Bit Depth Resampling	IH	Uniformity	#N/A	#N/A	#N/A	#N/A
Thresholding and Bit Depth Resampling	IH	Kurtosis	0.00	1.00	0.50	0.48
Thresholding and Bit Depth Resampling	IH	Skewness	-0.40	0.49	0.50	0.48
Thresholding and Bit Depth Resampling	IH	Variance	1.00	0.08	-1.00	0.16
Thresholding and Smoothing	GLCM	Auto Correlation	0.80	0.17	-1.00	0.16
Thresholding and Smoothing	GLCM	ClusterProminence	1.00	0.08	-1.00	0.16
Thresholding and Smoothing	GLCM	Cluster Shade	-0.40	0.49	0.50	0.48
Thresholding and Smoothing	GLCM	Cluster Tendendcy	1.00	0.08	-1.00	0.16
Thresholding and Smoothing	GLCM	Contrast	0.40	0.49	-0.50	0.48
Thresholding and Smoothing	GLCM	Correlation	-0.40	0.49	0.50	0.48
Thresholding and Smoothing	GLCM	Difference Entropy	0.40	0.49	-1.00	0.16
Thresholding and Smoothing	GLCM	Dissimilarity	0.40	0.49	-0.50	0.48

Thresholding and Smoothing	GLCM	Energy	-0.80	0.17	1.00	0.16
Thresholding and Smoothing	GLCM	Entropy	0.40	0.49	-1.00	0.16
Thresholding and Smoothing	GLCM	Homogeneity	-1.00	0.08	1.00	0.16
Thresholding and Smoothing	GLCM	Homogeneity2	-1.00	0.08	1.00	0.16
Thresholding and Smoothing	GLCM	Information Measure Correlation 1	0.40	0.49	-0.50	0.48
Thresholding and Smoothing	GLCM	Information Measure Correlation 2	-0.40	0.49	0.50	0.48
Thresholding and Smoothing	GLCM	Inverse Difference Moment Norm	-0.20	0.73	0.50	0.48
Thresholding and Smoothing	GLCM	Inverse Difference Norm	-0.80	0.17	0.50	0.48
Thresholding and Smoothing	GLCM	Inverse Variance	-0.80	0.17	1.00	0.16
Thresholding and Smoothing	GLCM	Max Probability	0.80	0.17	-1.00	0.16
Thresholding and Smoothing	GLCM	Sum Average	0.80	0.17	-1.00	0.16
Thresholding and Smoothing	GLCM	Sum Entropy	0.20	0.73	-1.00	0.16
Thresholding and Smoothing	GLCM	Sum Variance	0.80	0.17	-1.00	0.16
Thresholding and Smoothing	GLCM	Variance	1.00	0.08	-1.00	0.16
Thresholding and Smoothing	GLRLM	Gray Level Nonuniformity	0.40	0.49	1.00	0.16
Thresholding and Smoothing	GLRLM	High Gray Level Run Emphasis	1.00	0.08	1.00	0.16

Thresholding and Smoothing	GLRLM	Long Run Emphasis	-1.00	0.08	0.50	0.48
Thresholding and Smoothing	GLRLM	Long Run High Gray Level Emphasis	-0.80	0.17	0.50	0.48
Thresholding and Smoothing	GLRLM	Long Run Low Gray Level Emphasis	-1.00	0.08	0.50	0.48
Thresholding and Smoothing	GLRLM	Low Gray Level Run Emphasis	-1.00	0.08	-1.00	0.16
Thresholding and Smoothing	GLRLM	Run Length Nonuniformity	1.00	0.08	-0.50	0.48
Thresholding and Smoothing	GLRLM	Run Percentage	1.00	0.08	-0.50	0.48
Thresholding and Smoothing	GLRLM	Short Run Emphasis	1.00	0.08	-1.00	0.16
Thresholding and Smoothing	GLRLM	Short Run High Gray Level Emphasis	1.00	0.08	1.00	0.16
Thresholding and Smoothing	GLRLM	Short Run Low Gray Level Emphasis	-0.80	0.17	-1.00	0.16
Thresholding and Smoothing	NGTDM	Busyness	0.40	0.49	-0.50	0.48
Thresholding and Smoothing	NGTDM	Coarseness	-0.40	0.49	0.50	0.48
Thresholding and Smoothing	NGTDM	Complexity	0.40	0.49	-1.00	0.16
Thresholding and Smoothing	NGTDM	Contrast	0.80	0.17	-0.50	0.48
Thresholding and Smoothing	NGTDM	TextureStrength	-0.40	0.49	-0.50	0.48
Thresholding and Smoothing	IH	Energy	1.00	0.08	1.00	0.16
Smoothing	IH	Entropy	0.20	0.73	-1.00	0.16

Thresholding and Smoothing	IH	Max	0.63	0.27	1.00	0.16
Thresholding and Smoothing	IH	Mean	1.00	0.08	1.00	0.16
Thresholding and Smoothing	IH	Median	1.00	0.08	1.00	0.16
Thresholding and Smoothing	IH	Min	0.26	0.65	1.00	0.16
Thresholding and Smoothing	IH	Standard Deviation	1.00	0.08	-1.00	0.16
Thresholding and Smoothing	IH	Uniformity	-0.40	0.49	1.00	0.16
Thresholding and Smoothing	IH	Kurtosis	-0.40	0.49	-0.50	0.48
Thresholding and Smoothing	IH	Skewness	-0.40	0.49	-0.50	0.48
Thresholding and Smoothing	IH	Variance	1.00	0.08	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Auto Correlation	0.40	0.49	1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	ClusterProminence	0.80	0.17	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Cluster Shade	-0.40	0.49	1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Cluster Tendendcy	0.80	0.17	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Contrast	0.40	0.49	-1.00	0.16
Thresholding, Smoothing,	GLCM	Correlation	-0.40	0.49	-1.00	0.16

Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Sum Average	0.40	0.49	1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Sum Entropy	1.00	0.08	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Sum Variance	0.40	0.49	1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Variance	0.80	0.17	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	Gray Level Nonuniformity	0.80	0.17	-0.50	0.48
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	High Gray Level Run Emphasis	1.00	0.08	1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	Long Run Emphasis	-0.80	0.17	1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	Long Run High Gray Level Emphasis	-0.80	0.17	1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	Long Run Low Gray Level Emphasis	-1.00	0.08	1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	Low Gray Level Run Emphasis	-1.00	0.08	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	Run Length Nonuniformity	0.40	0.49	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	GLRLM	Run Percentage	0.80	0.17	-1.00	0.16
Thresholding, Smoothing,	GLRLM	Short Run Emphasis	0.40	0.49	-0.50	0.48

Thresholding, Smoothing, and Bit Depth Resampling	IH	Min	0.26	0.65	#N/A	#N/A
Thresholding, Smoothing, and Bit Depth Resampling	IH	Standard Deviation	0.80	0.17	-1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	IH	Uniformity	#N/A	#N/A	#N/A	#N/A
Thresholding, Smoothing, and Bit Depth Resampling	IH	Kurtosis	0.20	0.73	1.00	0.16
Thresholding, Smoothing, and Bit Depth Resampling	IH	Skewness	-0.40	0.49	0.50	0.48
Thresholding, Smoothing, and Bit Depth Resampling	IH	Variance	0.80	0.17	-1.00	0.16

Appendix C: Supplemental Material for Chapter 6

Table C-1. Features that were not robust across the volume range for each slice removal and preprocessing technique

ROI Volume Range	Preprocessing technique	Sequential Slice Removal	Random Slice Removal
75%-100%	Thresholding	n/a	n/a
	Thresholding and 8-bit depth resampling	<ul style="list-style-type: none"> • Contrast 	n/a
	Thresholding and Butterworth smoothing	n/a	n/a
	Thresholding, smoothing, and 8-bit depth resampling	n/a	n/a
50%-100%	Thresholding	<ul style="list-style-type: none"> • Information measure correlation 1 • Busyness • Texture strength • Skewness 	<ul style="list-style-type: none"> • Texture strength • Skewness
	Thresholding and 8-bit depth resampling	<ul style="list-style-type: none"> • Busyness • Contrast • Texture strength • Skewness 	<ul style="list-style-type: none"> • Texture strength • Skewness
	Thresholding and Butterworth smoothing	<ul style="list-style-type: none"> • Information measure correlation 1 • Information measure correlation 2 • Busyness • Texture strength 	<ul style="list-style-type: none"> • Information measure correlation 1 • Texture strength
	Thresholding, smoothing, and 8-bit depth resampling	<ul style="list-style-type: none"> • Busyness • Contrast • Texture strength 	<ul style="list-style-type: none"> • Texture strength
25%-100%	Thresholding	<ul style="list-style-type: none"> • Auto correlation • Cluster tendency • Contrast (GLCM) • Correlation • Information measure correlation 1 	<ul style="list-style-type: none"> • Correlation • Information measure correlation 1 • Information measure correlation 2 • Max probability

		<ul style="list-style-type: none"> • Information measure correlation 2 • Max probability • Sum average • Sum variance • Variance (GLCM) • Long run high gray level emphasis • Short run low gray level emphasis • Busyness • Contrast • Texture strength • Median • Minimum • Uniformity • Skewness • Variance 	<ul style="list-style-type: none"> • Busyness • Contrast • Texture strength • Minimum • Skewness
	Thresholding and 8-bit depth resampling	<ul style="list-style-type: none"> • Auto correlation • Cluster tendency • Contrast (GLCM) • Correlation • Information measure correlation 1 • Information measure correlation 2 • Sum average • Sum variance • Variance (GLCM) • Gray level nonuniformity • Busyness • Contrast • Texture strength • Entropy • Median • Skewness 	<ul style="list-style-type: none"> • Correlation • Busyness • Contrast • Texture strength • Skewness
	Thresholding and Butterworth smoothing	<ul style="list-style-type: none"> • Auto correlation • Cluster shade • Cluster tendency • Contrast (GLCM) • Correlation • Information measure correlation 1 	<ul style="list-style-type: none"> • Correlation • Information measure correlation 1 • Information measure correlation 2 • Max probability • Busyness • Contrast

		<ul style="list-style-type: none"> • Information measure correlation 2 • Max probability • Sum average • Sum variance • Variance (GLCM) • Gray level nonuniformity • Long run high gray level emphasis • Short run low gray level emphasis • Busyness • Contrast • Texture strength • Median • Minimum • Uniformity • Skewness • Variance 	<ul style="list-style-type: none"> • Texture strength • Minimum • Skewness
	Thresholding, smoothing, and 8-bit depth resampling	<ul style="list-style-type: none"> • Auto correlation • Cluster shade • Cluster tendency • Contrast (GLCM) • Correlation • Information measure correlation 1 • Information measure correlation 2 • Max probability • Sum average • Sum variance • Variance (GLCM) • Gray level nonuniformity • Busyness • Contrast • Texture strength • Entropy • Median • Minimum • Skewness • Variance 	<ul style="list-style-type: none"> • Correlation • Busyness • Contrast • Texture strength • Skewness
0%-100%	Thresholding	<p>All features fail except:</p> <ul style="list-style-type: none"> • Complexity 	<p>All features fail except:</p> <ul style="list-style-type: none"> • Complexity

			<ul style="list-style-type: none"> • Maximum
	Thresholding and 8-bit depth resampling	All features fail except: <ul style="list-style-type: none"> • Complexity 	All features fail except: <ul style="list-style-type: none"> • Long run emphasis • Long run low gray level emphasis • Complexity • Maximum • Uniformity
	Thresholding and Butterworth smoothing	All features fail except: <ul style="list-style-type: none"> • Complexity 	All features fail except: <ul style="list-style-type: none"> • Complexity • Maximum
	Thresholding, smoothing, and 8-bit depth resampling	All features fail except: <ul style="list-style-type: none"> • Complexity 	All features fail except: <ul style="list-style-type: none"> • Complexity • Maximum • Uniformity

Appendix D: Supplemental Material Chapter 7



Figure D-1: Pearson Correlation of Features with Image Thickness with Only Robust Cartridges.

Absolute value of the Pearson correlation rho for the correlation between feature value and image thickness for each region of interest (ROI). Each ROI is a different shape. Each category of feature is a different color. The correlation varies between and within features depending on the ROI. COM: gray level co-occurrence matrix, GLCM: gray level co-occurrence (used when there are features with the

same name in different categories to differentiate them), GLRLM: gray level run length matrix, NGTDM: neighborhood gray tone difference matrix.

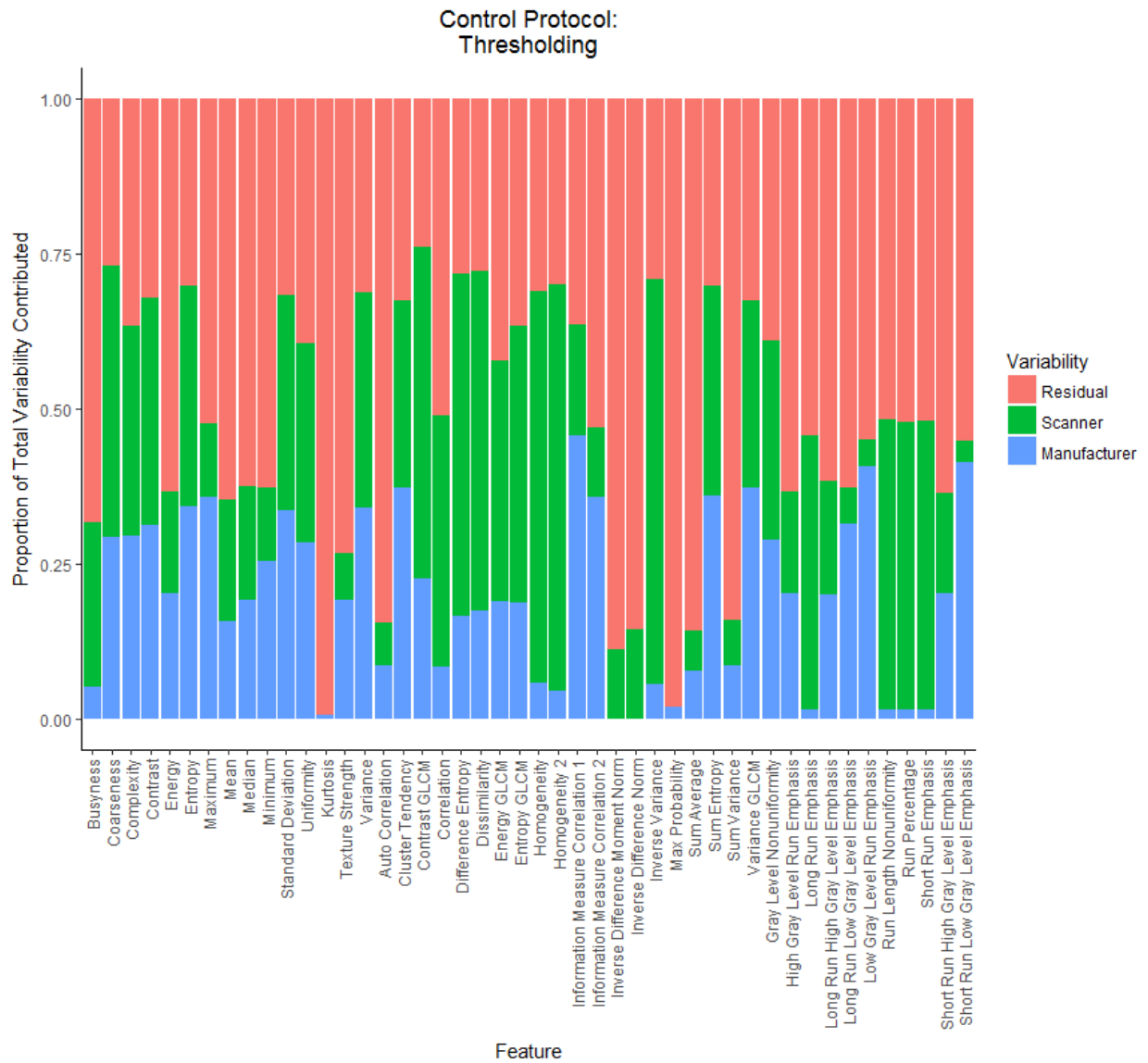


Figure D-2: Bar Plots of Variability in Controlled Protocol with Thresholding.

Bar plots of the relative contribution of the scanner wise variability, manufacturer wise variability, and residual variability for each feature using thresholding calculated on the controlled protocol.

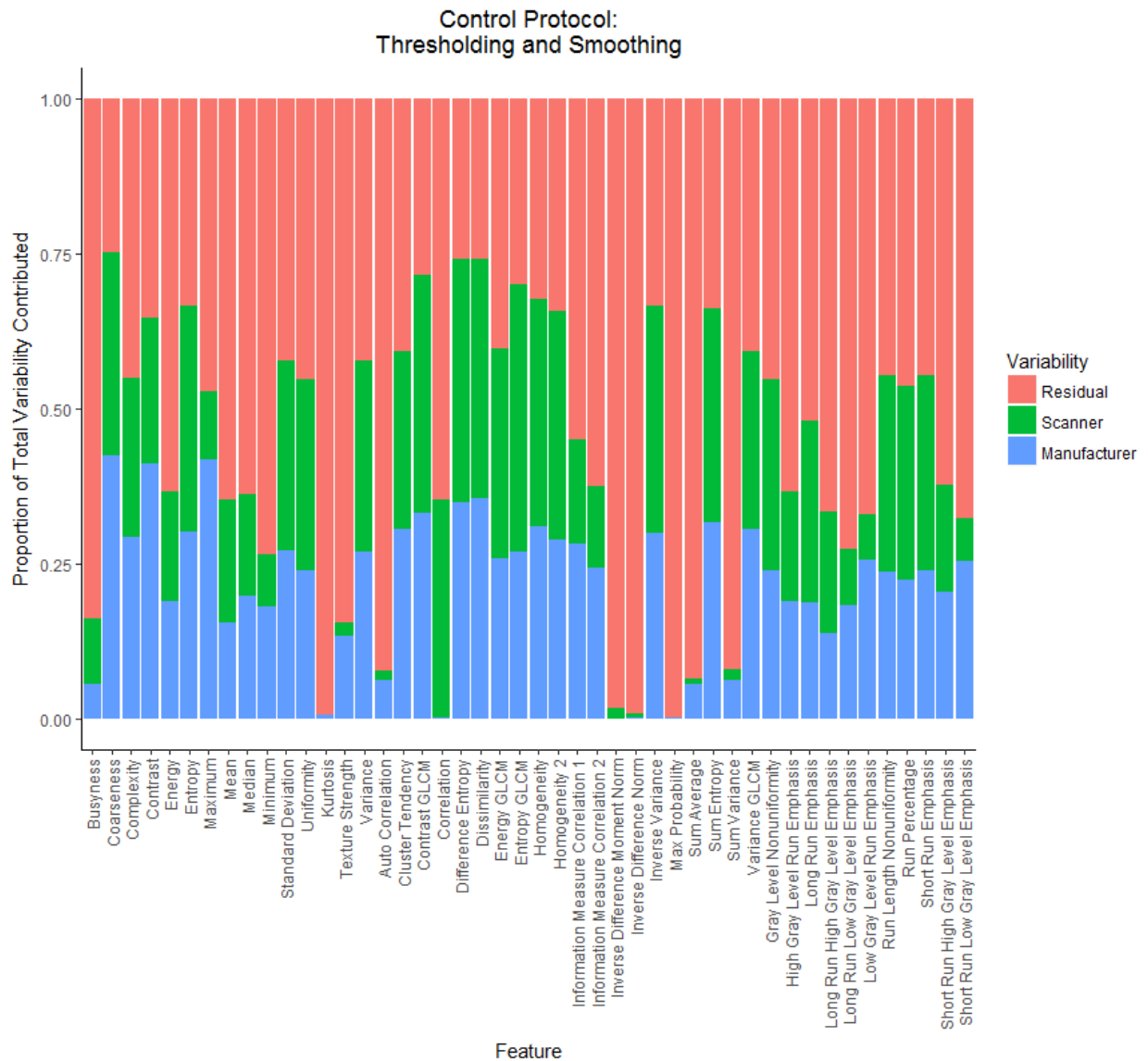


Figure D-3: Bar Plots of Variability in Controlled Protocol with Thresholding and Smoothing.

Bar plots of the relative contribution of the scanner wise variability, manufacturer wise variability, and residual variability for each feature using thresholding and smoothing calculated on the controlled protocol.

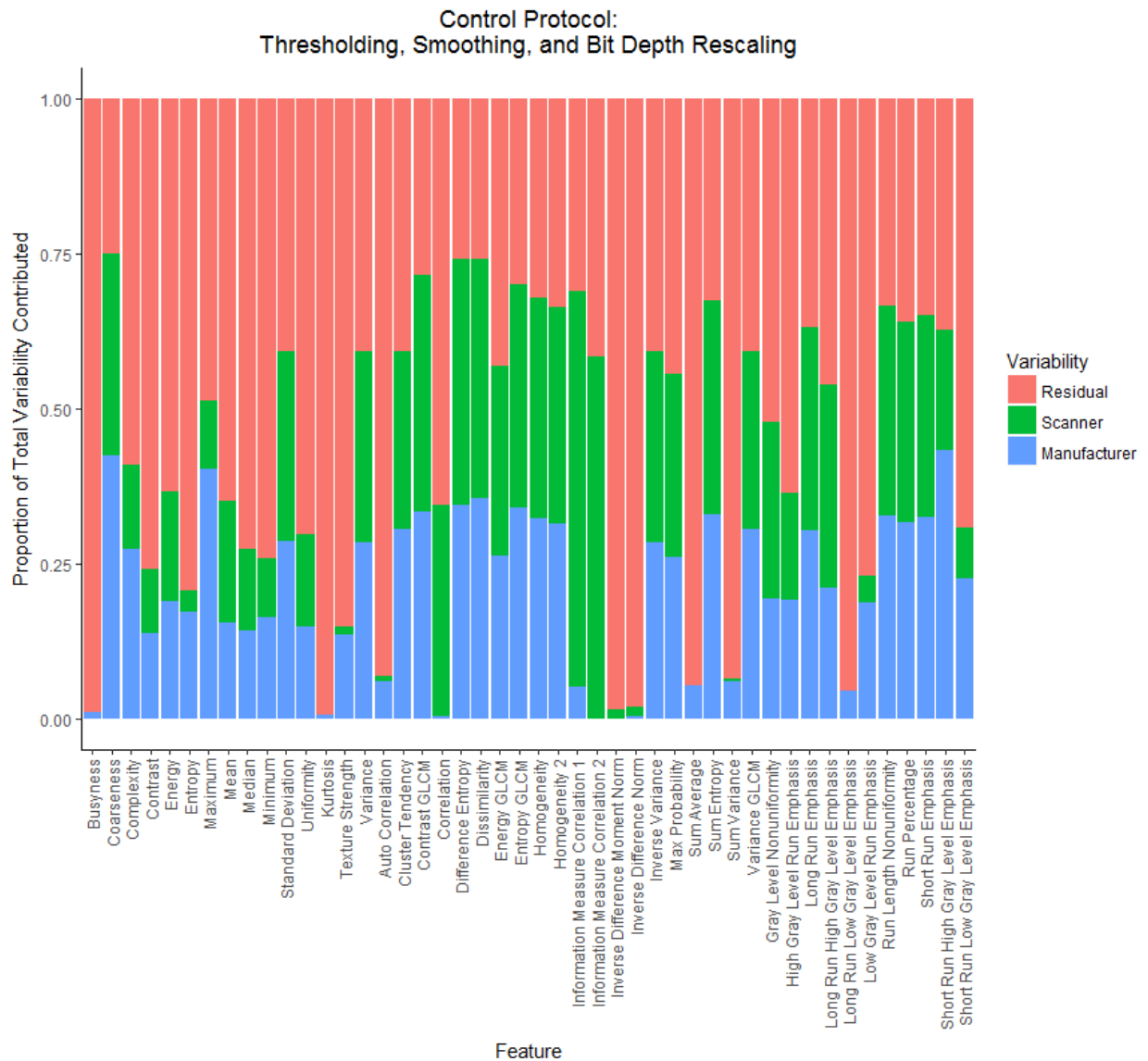


Figure D-4: Bar Plots of Variability in Controlled Protocol with Thresholding, Smoothing, and Bit Depth Rescaling.

Bar plots of the relative contribution of the scanner wise variability, manufacturer wise variability, and residual variability for each feature using thresholding, smoothing, and bit depth rescaling calculated on the controlled protocol.

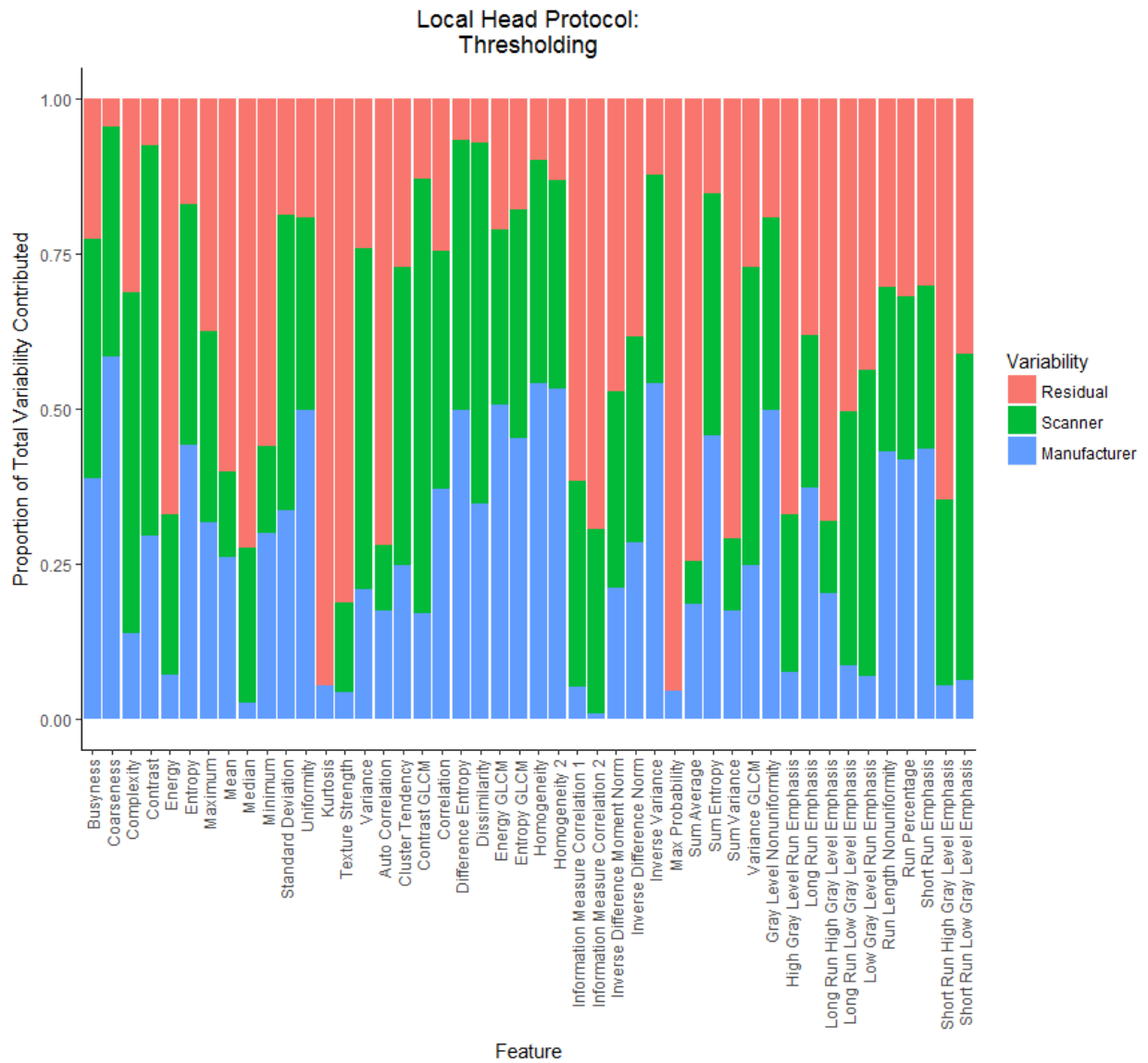


Figure D-5: Bar Plots of Variability in Head Protocol with Thresholding.

Bar plots of the relative contribution of the scanner wise variability, manufacturer wise variability, and residual variability for each feature using thresholding calculated on the local head protocol.

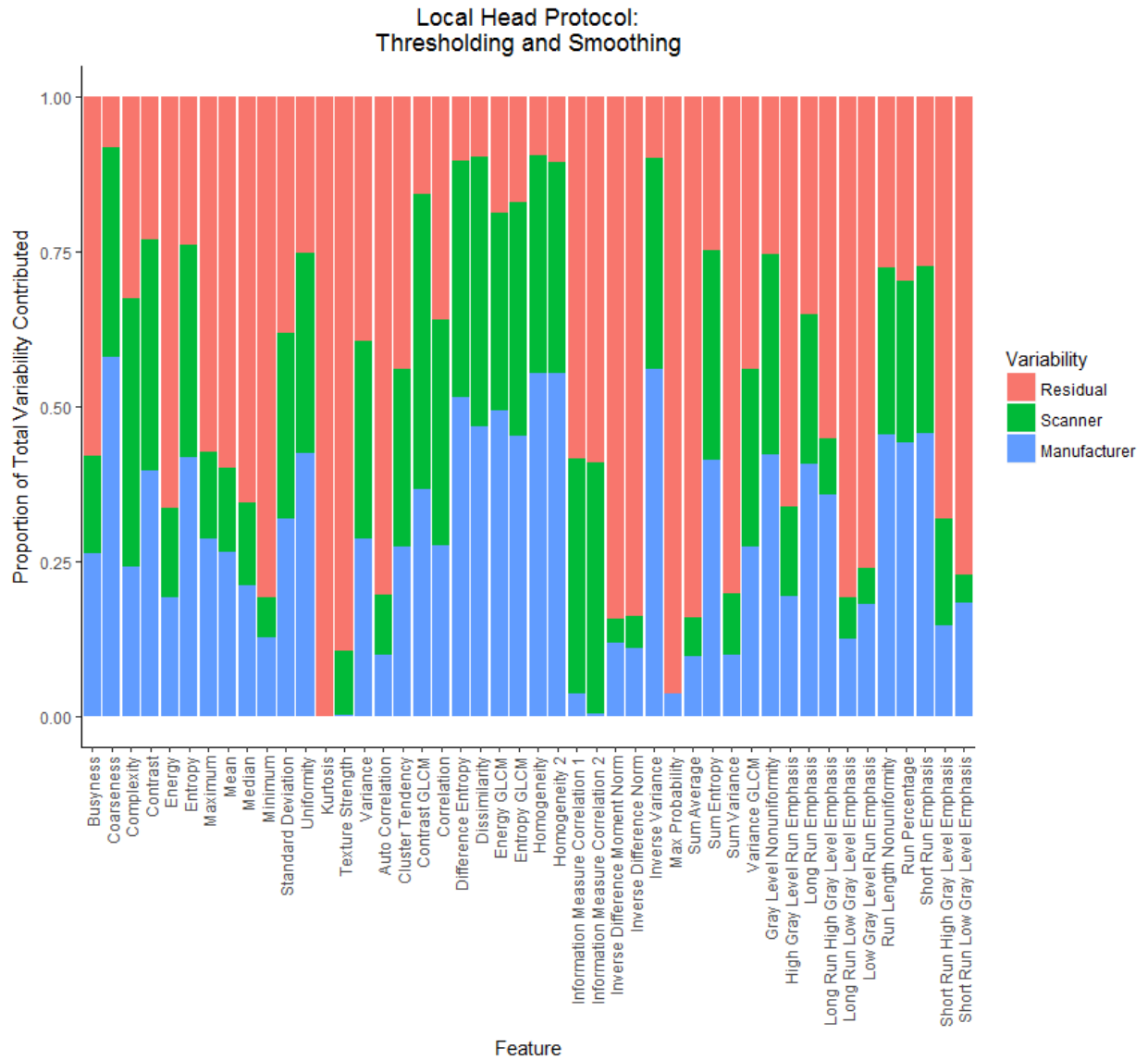


Figure D-6: Bar Plots of Variability in Head Protocol with Thresholding and Smoothing.

Bar plots of the relative contribution of the scanner wise variability, manufacturer wise variability, and residual variability for each feature using thresholding and smoothing calculated on the local head protocol.

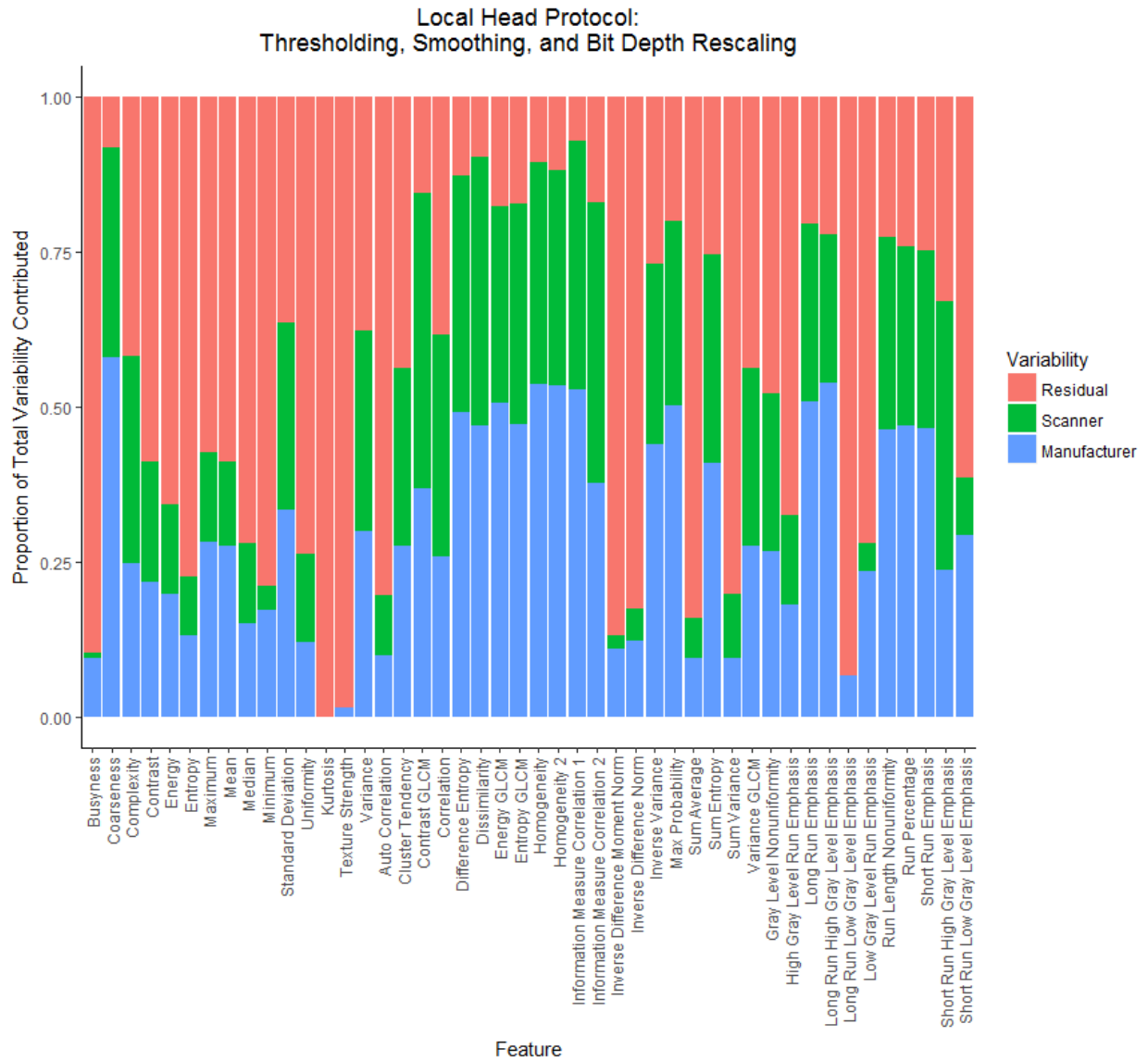


Figure D-7: Bar Plots of Variability in Head Protocol with Thresholding, Smoothing, and Bit Depth Rescaling.

Bar plots of the relative contribution of the scanner wise variability, manufacturer wise variability, and residual variability for each feature using thresholding, smoothing, and bit depth rescaling calculated on the local head protocol.

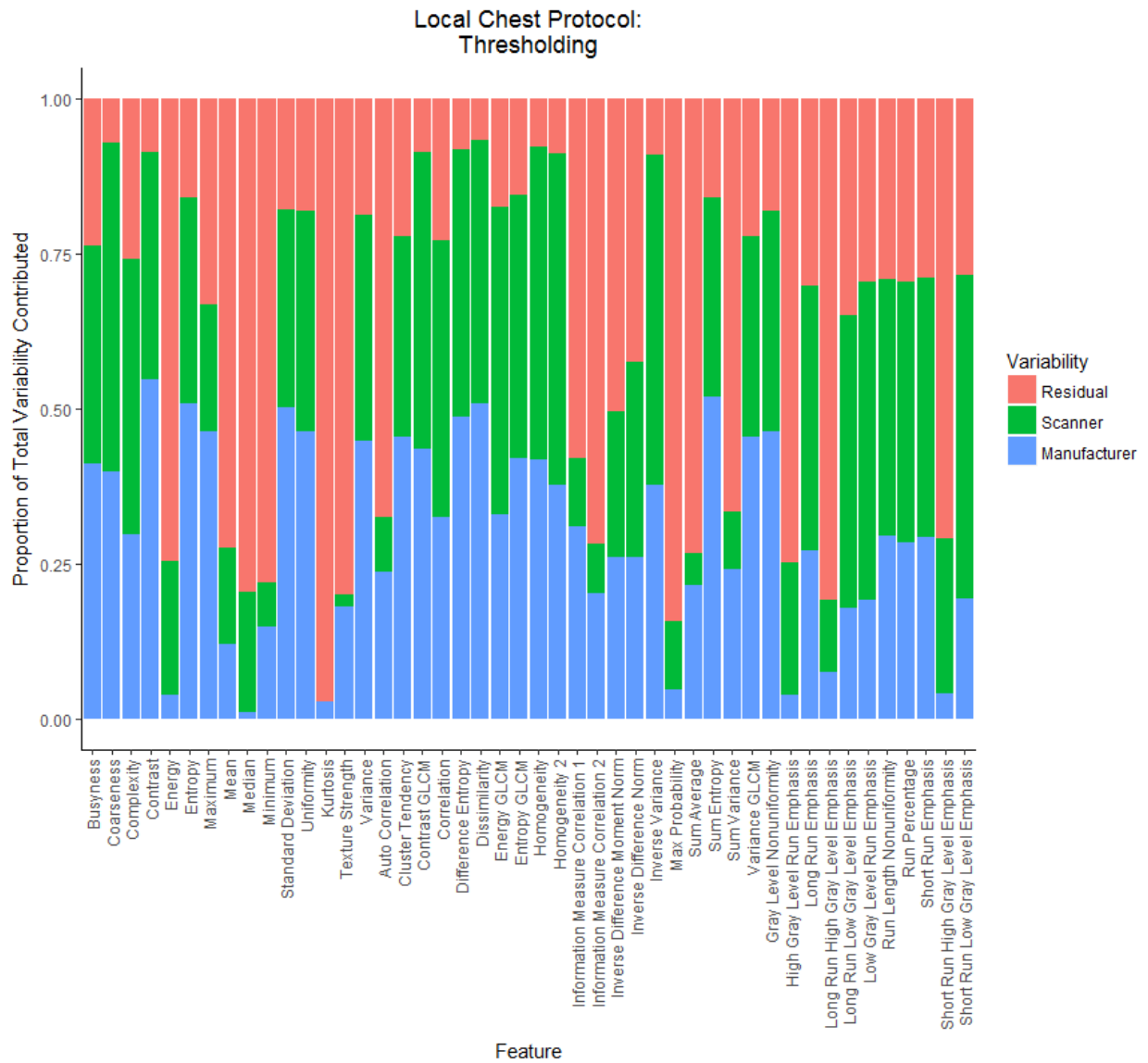


Figure D-8: Bar Plots of Variability in Thoracic Protocol with Thresholding.

Bar plots of the relative contribution of the scanner wise variability, manufacturer wise variability, and residual variability for each feature using thresholding calculated on the local chest protocol.

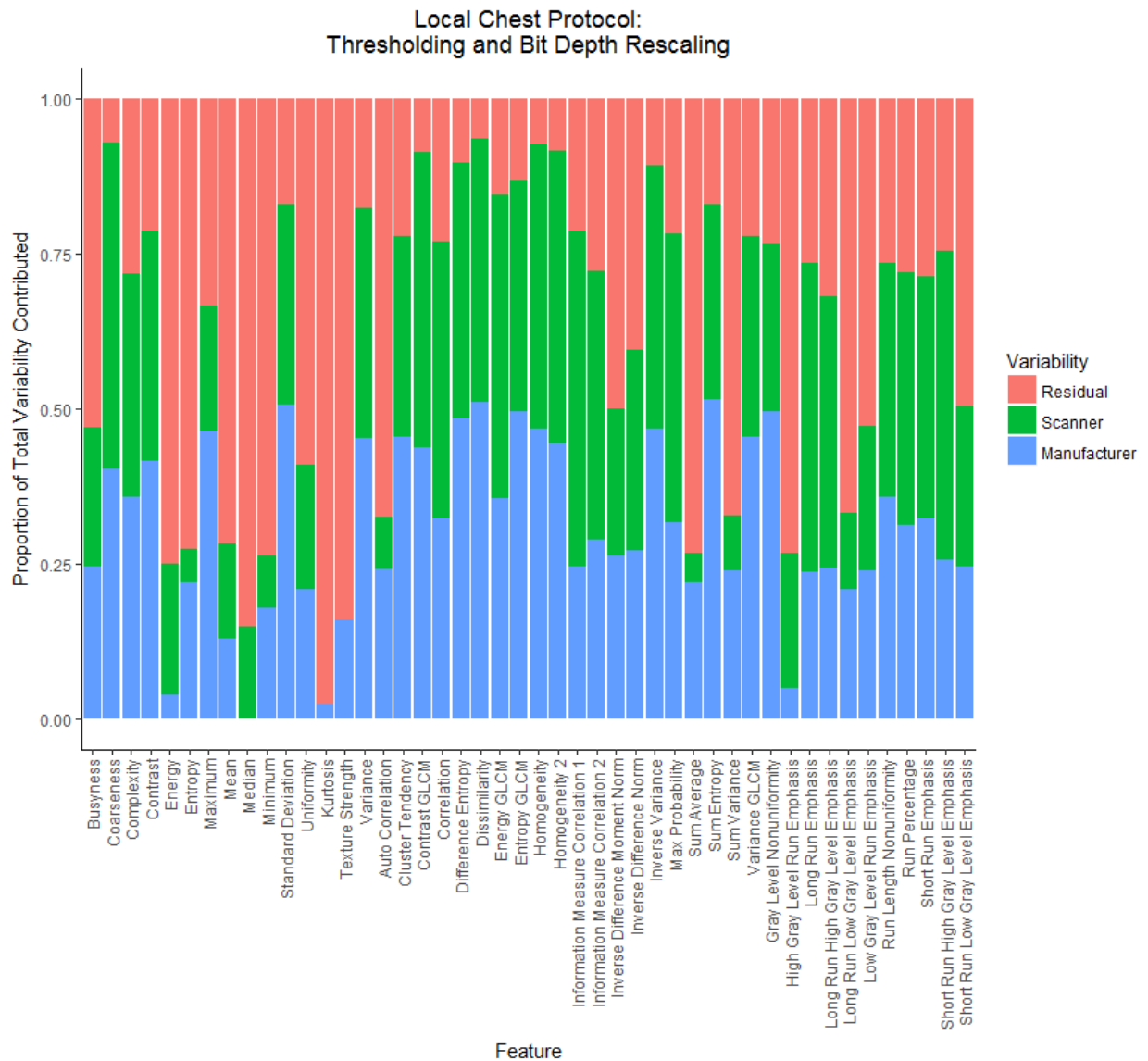


Figure D-9: Bar Plots of Variability in Thoracic Protocol with Thresholding and Bit Depth Rescaling.

Bar plots of the relative contribution of the scanner wise variability, manufacturer wise variability, and residual variability for each feature using thresholding and bit depth rescaling calculated on the local chest protocol.

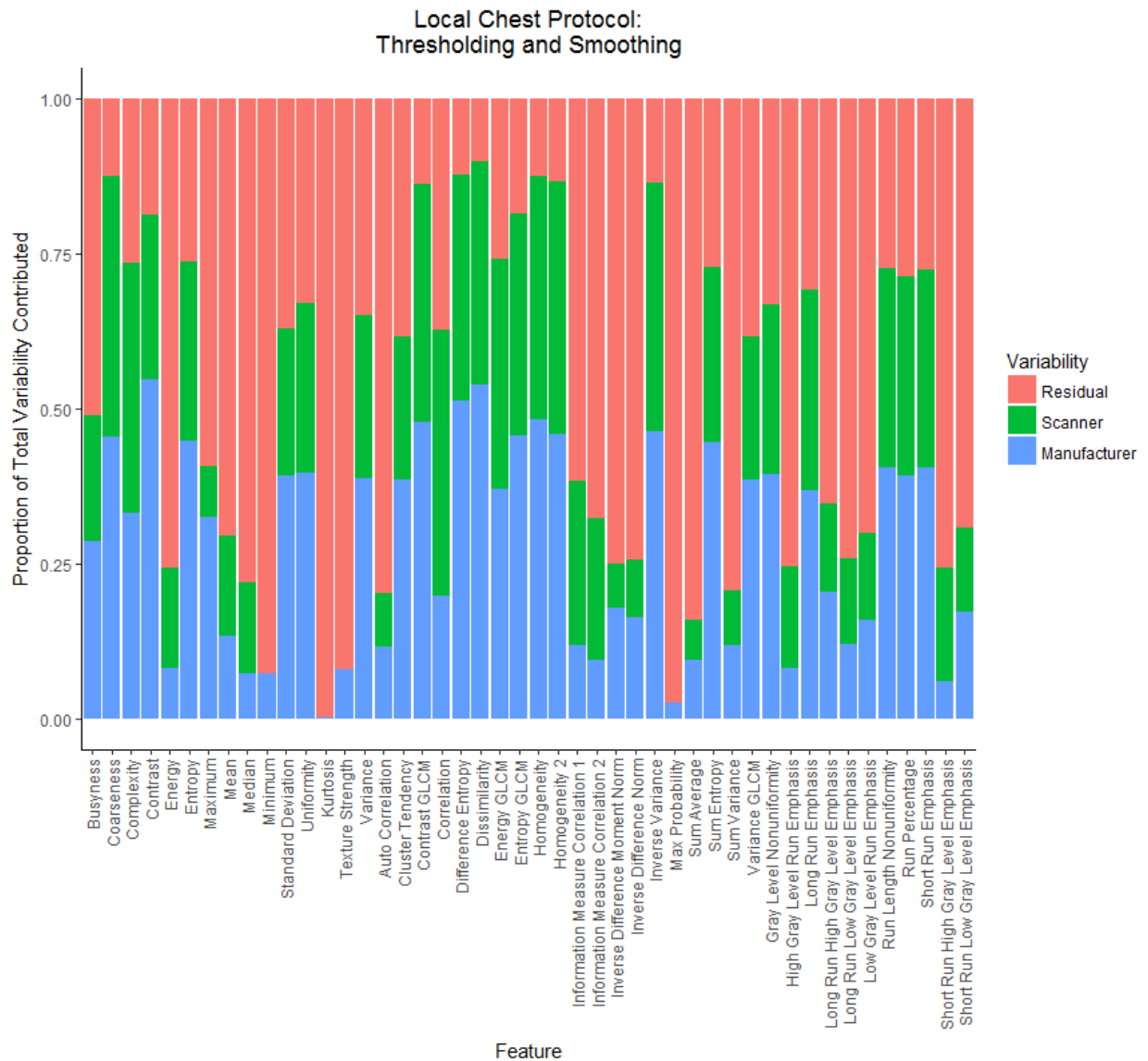


Figure D-10: Bar Plots of Variability in Thoracic Protocol with Thresholding and Smoothing.

Bar plots of the relative contribution of the scanner wise variability, manufacturer wise variability, and residual variability for each feature using thresholding and smoothing calculated on the local chest protocol.

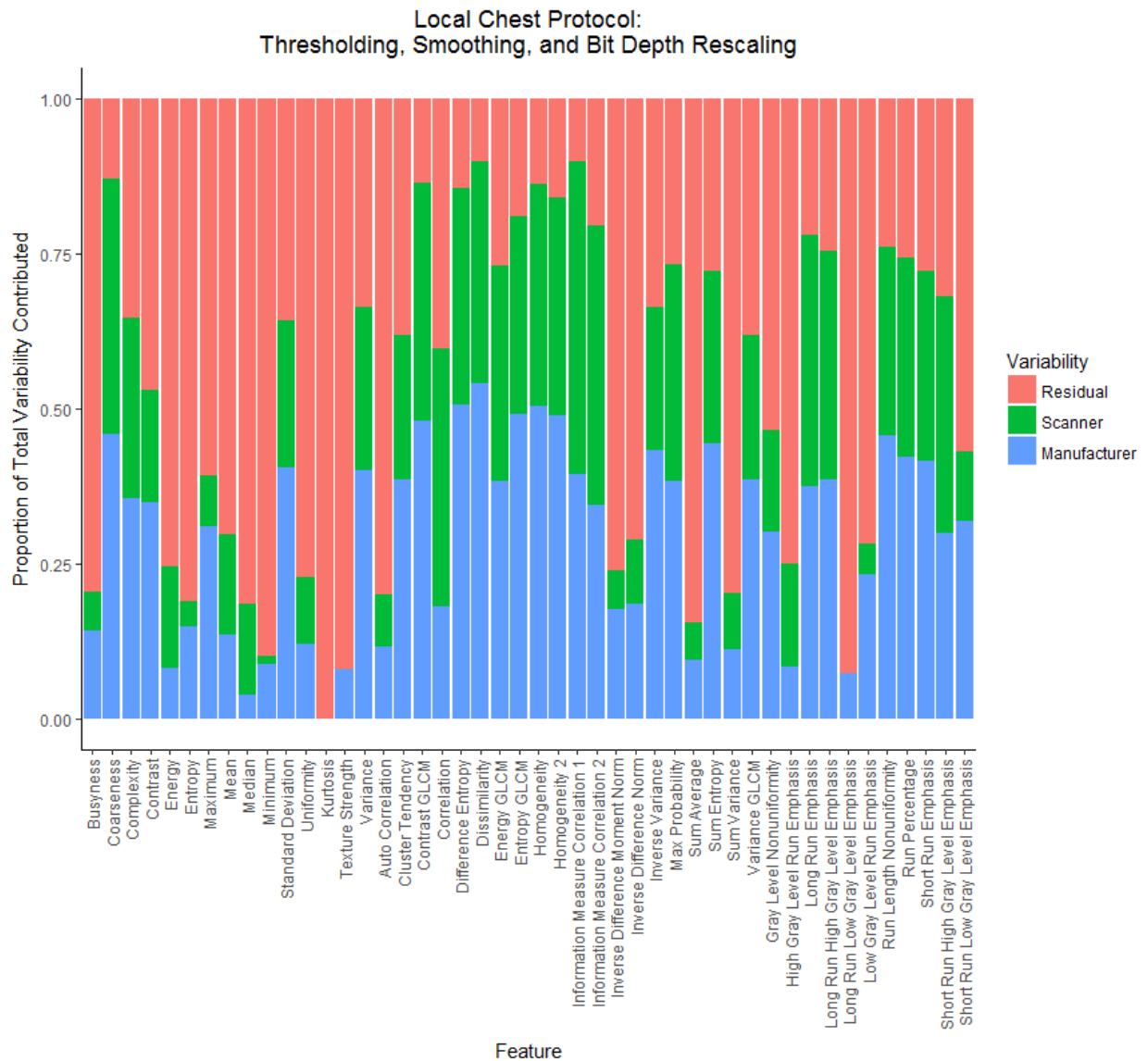


Figure D-11: Bar Plots of Variability in Thoracic Protocol with Thresholding, Smoothing, and Bit Depth Rescaling.

Bar plots of the relative contribution of the scanner wise variability, manufacturer wise variability, and residual variability for each feature using thresholding, smoothing, and bit depth rescaling calculated on the local chest protocol.

Table D-1. P-values for One Sided Pairwise T-tests between Control and Local Protocol Scans

	Control vs Chest Protocol				Control vs Head Protocol			
	Thresholding	Thresholding and Smoothing	Thresholding and Bit Depth Rescaling	Thresholding, Smoothing, and Bit Depth Rescaling	Thresholding	Thresholding and Smoothing	Thresholding and Bit Depth Rescaling	Thresholding, Smoothing, and Bit Depth Rescaling
σ_{β}	0.037	0.032	0.00073	0.0042	0.035	0.54	0.0015	0.0024
σ_{γ}	0.027	0.018	0.00022	0.00049	0.0072	0.90	0.0000093	0.000097
σ_{ϵ}	0.038	0.046	0.0050	0.0049	0.036	0.42	0.0014	0.073

σ_{β} : scanner-wise variability, σ_{γ} : manufacturer-wise variability, σ_{ϵ} : residual variability

Table D-2. P-values for Pairwise T-tests between Variability from Model with All Manufacturers vs GE Only

	Scanner-wise Variability			Residual variability		
	Control Protocol	Local Chest Protocol	Local Head Protocol	Control Protocol	Local Chest Protocol	Local Head Protocol
Thresholding	0.088	0.084	0.076	0.29	0.070	0.071
Thresholding and Smoothing	0.45	0.0078	0.0024	0.051	0.13	0.48
Thresholding and Bit Depth Resampling	0.93	0.21	0.025	0.0022	0.028	0.0081
Thresholding, Smoothing, and Bit Depth Resampling	0.049	0.00022	0.0016	0.0024	0.083	0.0014

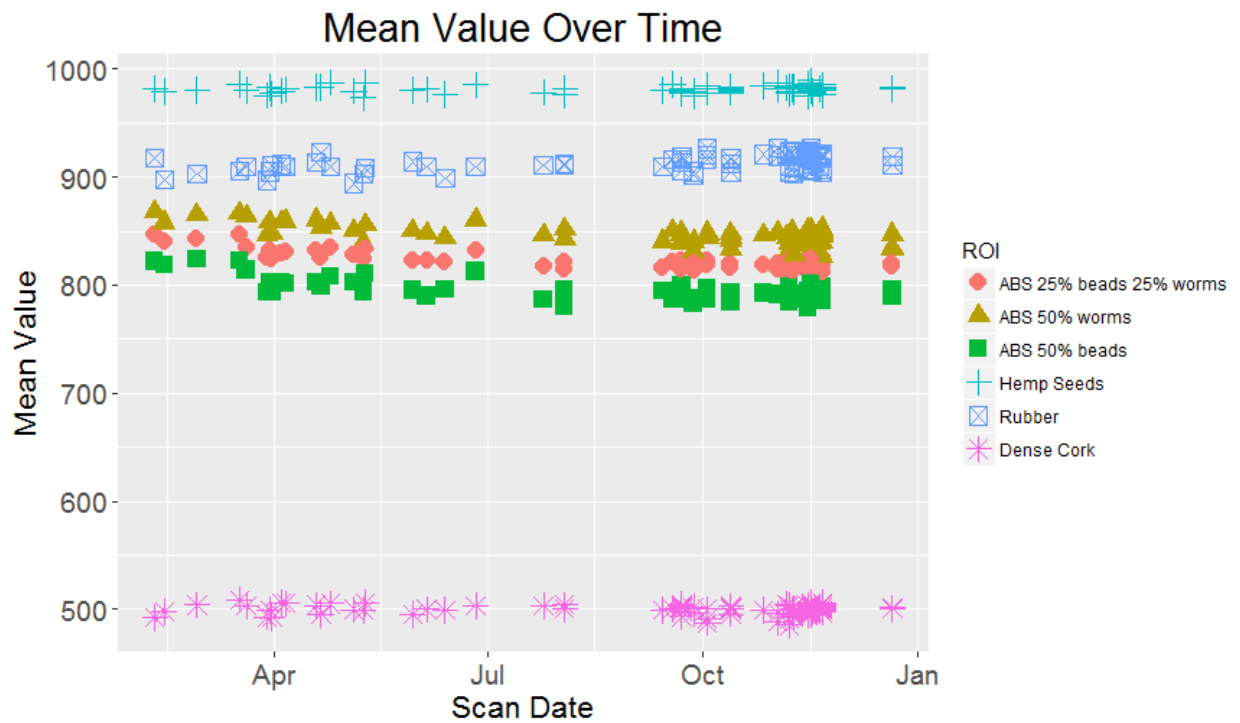


Figure D-12: Mean Value of Radiomics Cartridges Over Time.

Mean value over time for each ROI, shown as a different colored and shaped point. The three cartridges with ABS demonstrated a downward trend with time while the other cartridges did not demonstrate any trend with time.

Percentage of Features Outside 1/3 Scaled Patient Standard Deviation Using All Robust Features

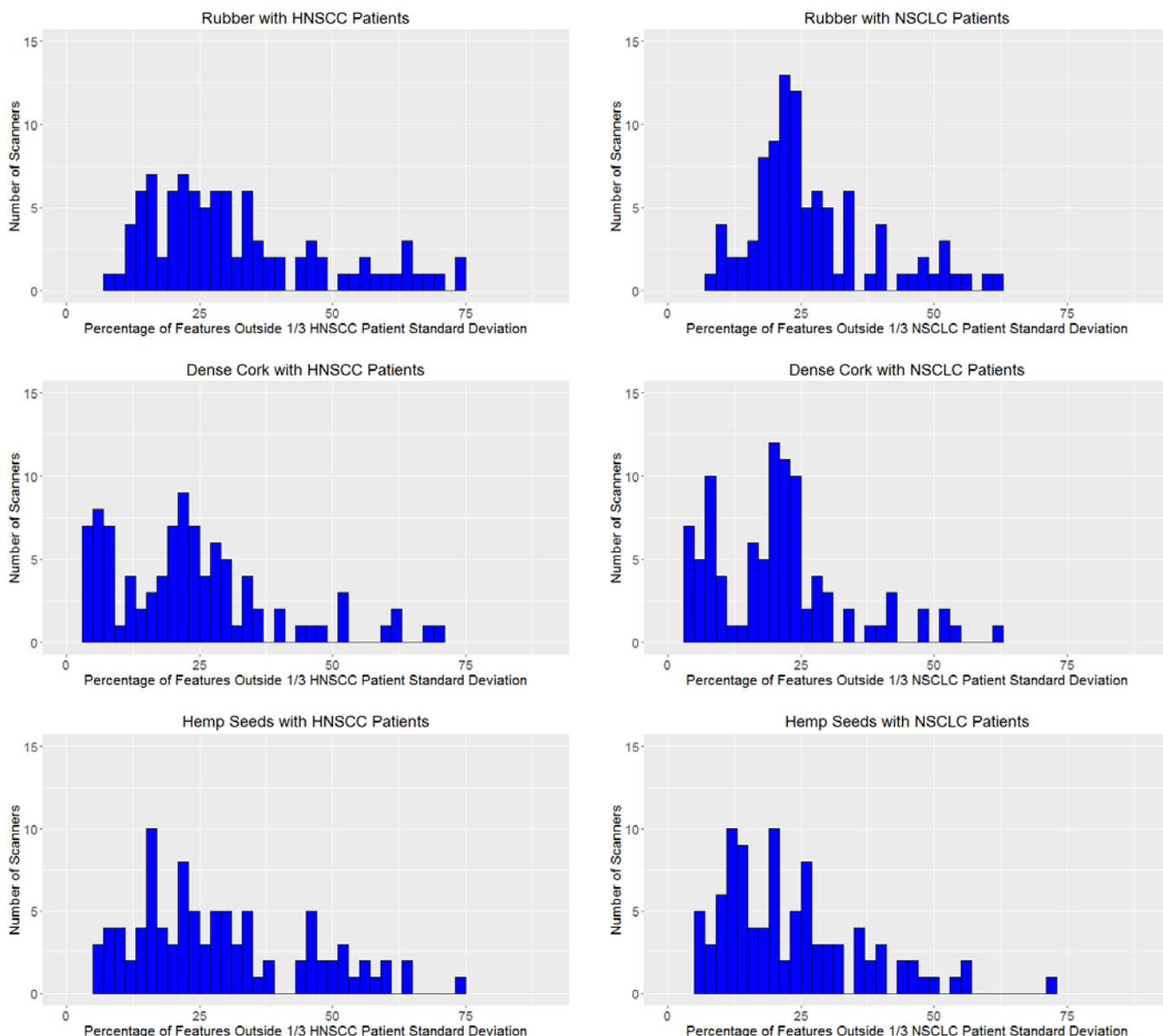


Figure D-13: Percentage of Features Outside Patient Bounds.

The percentage of features outside 1/3 scaled patient standard deviation for rubber, dense cork, and hemp seeds using the head and neck squamous cell carcinoma (HNSCC) patient cohort and the non-small cell lung cancer (NSCLC) patient cohort. For all plots, there are a large percentage of features outside of a third of the scaled patient standard deviation.

Table D-3. Preprocessing and Features From Previous Studies Used for QA Analysis

Preprocessing	Feature Group	Feature
Thresholding and Smoothing	NGTDM	Texture Strength
Thresholding and Smoothing	IH	Variance
Thresholding and Smoothing	IH	Standard Deviation
Thresholding and Smoothing	IH	Mean
Thresholding and Smoothing	IH	Kurtosis
Thresholding, Smoothing, and Bit Depth Resampling	IH	Entropy
Thresholding and Smoothing	IH	Energy
Thresholding and Smoothing	GLCM	Inverse Difference Norm
Thresholding and Smoothing	GLCM	Inverse Difference Moment Norm
Thresholding	GLCM	Information Measure Correlation 1
Thresholding	GLCM	Information Measure Correlation 2
Thresholding and Smoothing	GLCM	Dissimilarity
Thresholding, Smoothing, and Bit Depth Resampling	GLCM	Difference Entropy
Thresholding and Smoothing	GLCM	Contrast
Thresholding	NGTDM	Busyness
Thresholding	IH	Kurtosis

NGTDM: neighborhood gray tone difference matrix; IH: intensity histogram; GLCM: gray level co-occurrence matrix

Appendix E: Supplemental Material for Chapter 8

Table E-1. Philips Scanner ICC Values for Protocol Parameters Changed

Feature	Category	Time per bed position		Filter cutoff		Filter cutoff (subset)		Iterations and subsets		Iterations and subsets (subset)	
		64 levels	fixed bin	64 levels	fixed bin	64 levels	fixed bin	64 levels	fixed bin	64 levels	fixed bin
AutoCorrelation	GLCM	0.99	1.00	0.91	0.99	0.92	1.00	0.91	0.99	0.99	1.00
ClusterProminence	GLCM	0.98	1.00	0.66	0.83	0.89	0.95	0.52	0.96	0.93	0.99
ClusterShade	GLCM	1.00	1.00	0.93	0.95	0.98	0.98	0.65	0.86	0.97	0.98
ClusterTendency	GLCM	0.99	1.00	0.68	0.88	0.91	0.97	0.41	0.90	0.93	0.99
Contrast	GLCM	0.97	1.00	0.89	0.90	0.89	0.98	0.84	0.88	0.97	0.99
Correlation	GLCM	1.00	1.00	0.82	0.82	0.92	0.92	0.79	0.79	0.94	0.94
DifferenceEntropy	GLCM	0.98	1.00	0.76	0.88	0.88	0.93	0.77	0.90	0.95	0.98
Dissimilarity	GLCM	0.98	1.00	0.86	0.95	0.91	0.97	0.86	0.94	0.96	0.99
Energy	GLCM	0.99	1.00	0.85	0.79	0.91	0.87	0.25	0.25	0.88	0.87
Entropy	GLCM	0.98	1.00	0.85	0.91	0.93	0.96	0.71	0.84	0.95	0.97
Homogeneity	GLCM	0.99	0.99	0.88	0.83	0.97	0.96	0.77	0.82	0.98	0.97
Homogeneity2	GLCM	0.99	1.00	0.88	0.81	0.98	0.95	0.78	0.82	0.98	0.97
InformationMeasureCorr1	GLCM	0.99	1.00	0.80	0.75	0.96	0.93	0.78	0.75	0.97	0.95
InformationMeasureCorr2	GLCM	1.00	1.00	0.74	0.72	0.91	0.90	0.73	0.73	0.91	0.91
InverseDiffMomentNorm	GLCM	0.97	0.97	0.89	0.89	0.89	0.90	0.84	0.81	0.97	0.96
InverseDiffNorm	GLCM	0.99	0.98	0.85	0.84	0.92	0.91	0.86	0.84	0.96	0.96
InverseVariance	GLCM	0.99	0.98	0.68	0.77	0.89	0.78	0.62	0.00	0.91	0.87
MaxProbability	GLCM	0.99	1.00	0.84	0.88	0.88	0.88	0.54	0.62	0.92	0.94
SumAverage	GLCM	0.99	1.00	0.95	0.98	0.96	1.00	0.91	0.98	0.99	1.00
SumEntropy	GLCM	0.98	1.00	0.86	0.95	0.88	0.97	0.61	0.77	0.93	0.97
SumVariance	GLCM	0.99	1.00	0.90	0.98	0.91	0.99	0.90	0.99	0.98	1.00
Variance	GLCM	0.99	1.00	0.68	0.88	0.91	0.97	0.41	0.90	0.93	0.99
GrayLevelNonuniformity	GLRLM	0.99	1.00	0.96	0.76	0.97	0.98	0.85	0.69	0.99	0.98
HighGrayLevelRunEmpha	GLRLM	0.98	1.00	0.90	0.99	0.91	0.99	0.90	0.99	0.98	1.00
LongRunEmphasis	GLRLM	0.98	0.97	0.94	0.62	0.96	0.95	0.39	0.10	0.98	0.94
LongRunHighGrayLevelEmpha	GLRLM	0.97	0.96	0.86	0.81	0.96	0.99	0.79	0.79	0.97	0.98
LongRunLowGrayLevelEmpha	GLRLM	0.99	1.00	0.91	0.70	0.96	0.92	0.53	0.11	0.95	0.89
LowGrayLevelRunEmpha	GLRLM	1.00	1.00	0.96	0.95	0.98	1.00	0.96	0.94	0.99	0.99
RunLengthNonuniformity	GLRLM	0.98	0.99	0.91	0.82	0.97	0.92	0.77	0.85	0.97	0.96
RunPercentage	GLRLM	0.98	0.99	0.93	0.76	0.97	0.94	0.72	0.80	0.98	0.96
ShortRunEmphasis	GLRLM	0.98	0.99	0.90	0.61	0.97	0.87	0.63	0.68	0.96	0.94
ShortRunHighGrayLevelEmpha	GLRLM	0.98	1.00	0.92	0.99	0.91	0.99	0.93	0.96	0.98	1.00
ShortRunLowGrayLevelEmpha	GLRLM	0.99	0.99	0.93	0.80	0.94	0.96	0.94	0.70	0.98	0.96
EnergyNorm	IH	0.99	1.00	0.92	0.99	0.93	1.00	0.92	0.99	0.99	1.00
GlobalEntropy	IH	0.92	0.98	0.67	0.74	0.83	0.88	0.60	0.51	0.93	0.97
GlobalStd	IH	0.98	1.00	0.67	0.92	0.91	0.99	0.45	0.91	0.93	0.99
GlobalUniformity	IH	0.96	0.98	0.72	0.75	0.87	0.89	0.46	0.56	0.94	0.97
Kurtosis	IH	1.00	1.00	0.67	0.67	0.86	0.86	0.19	0.20	0.93	0.93
Skewness	IH	1.00	1.00	0.92	0.92	0.98	0.98	0.75	0.75	0.99	0.99
Variance	IH	0.98	1.00	0.66	0.89	0.90	0.98	0.42	0.92	0.93	0.99
Busyness	NGTDM	0.99	0.98	0.81	0.89	0.86	0.92	0.86	0.85	0.94	0.92
Coarseness	NGTDM	0.98	1.00	0.62	0.69	0.79	0.77	0.61	0.50	0.89	0.87
Complexity	NGTDM	0.92	0.93	0.77	0.72	0.82	0.84	0.80	0.70	0.97	0.87
Contrast	NGTDM	0.96	0.95	0.63	0.71	0.83	0.87	0.53	0.71	0.91	0.91
TextureStrength	NGTDM	1.00	1.00	0.98	0.90	0.99	0.98	0.98	0.86	0.99	0.98
Feature	Category	64 levels	fixed bin	64 levels	fixed bin	64 levels	fixed bin	64 levels	fixed bin	64 levels	fixed bin
AutoCorrelation	GLCM	0.99	1.00	0.91	0.99	0.92	1.00	0.91	0.99	0.99	1.00
ClusterProminence	GLCM	0.98	1.00	0.66	0.83	0.89	0.95	0.52	0.96	0.93	0.99
ClusterShade	GLCM	1.00	1.00	0.93	0.95	0.98	0.98	0.65	0.86	0.97	0.98

GLCM: gray level co-occurrence matrix; GLRLM: gray level run length matrix; IH: intensity histogram; NGTDM: neighborhood gray tone difference matrix
Filter cutoff (subset): cutoff values < 6 mm
Iterations and subsets (subset): $16 \leq \text{effective iterations} \leq 45$

Table E-2. GE Scanner ICC Values for Protocol Parameters Changed

Feature	Category	Time per bed position		Filter cutoff		Filter cutoff (subset)		Field of view		Field of view resampled pixel	
		64 levels	fixed bin	64 levels	fixed bin	64 levels	fixed bin	64 levels	fixed bin	64 levels	fixed bin
AutoCorrelation	GLCM	0.99	1.00	0.98	0.99	0.97	1.00	1.00	1.00	1.00	1.00
ClusterProminence	GLCM	0.96	1.00	0.90	0.97	0.91	0.99	0.99	0.99	0.99	0.99
ClusterShade	GLCM	1.00	1.00	0.96	0.99	0.98	1.00	0.99	1.00	1.00	1.00
ClusterTendency	GLCM	0.98	1.00	0.95	0.97	0.97	0.99	0.97	0.98	0.99	0.99
Contrast	GLCM	1.00	1.00	0.86	0.78	0.98	0.95	0.57	0.56	1.00	1.00
Correlation	GLCM	1.00	1.00	0.87	0.87	0.98	0.98	0.55	0.56	1.00	1.00
DifferenceEntropy	GLCM	1.00	1.00	0.98	0.98	0.99	1.00	0.99	0.98	1.00	1.00
Dissimilarity	GLCM	1.00	1.00	0.95	0.94	0.99	0.99	0.83	0.82	1.00	1.00
Energy	GLCM	1.00	1.00	0.99	0.98	0.99	0.99	0.92	0.97	1.00	1.00
Entropy	GLCM	1.00	1.00	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00
Homogeneity	GLCM	1.00	1.00	0.99	0.99	0.99	1.00	1.00	0.99	1.00	1.00
Homogeneity2	GLCM	1.00	1.00	0.99	0.99	0.99	1.00	1.00	0.98	1.00	1.00
InformationMeasureCorr1	GLCM	1.00	1.00	0.99	0.98	1.00	0.99	1.00	0.99	1.00	1.00
InformationMeasureCorr2	GLCM	1.00	1.00	0.86	0.83	0.97	0.97	0.42	0.35	0.99	0.99
InverseDiffMomentNorm	GLCM	1.00	0.99	0.87	0.87	0.98	0.98	0.60	0.59	1.00	0.99
InverseDiffNorm	GLCM	1.00	1.00	0.96	0.95	0.99	0.99	0.86	0.86	1.00	1.00
InverseVariance	GLCM	1.00	0.98	0.94	0.71	0.99	0.90	0.50	0.37	0.99	0.97
MaxProbability	GLCM	1.00	1.00	0.98	0.99	0.99	0.99	0.97	0.99	0.99	1.00
SumAverage	GLCM	1.00	1.00	0.99	1.00	0.99	1.00	1.00	1.00	1.00	1.00
SumEntropy	GLCM	1.00	1.00	0.99	1.00	0.99	1.00	0.99	0.99	1.00	1.00
SumVariance	GLCM	0.99	1.00	0.97	0.99	0.97	1.00	1.00	1.00	1.00	1.00
Variance	GLCM	0.98	1.00	0.95	0.97	0.97	0.99	0.97	0.98	0.99	0.99
GrayLevelNonuniformity	GLRLM	1.00	1.00	0.94	0.82	1.00	0.98	0.75	0.25	1.00	1.00
HighGrayLevelRunEmpha	GLRLM	0.99	1.00	0.97	0.99	0.97	1.00	1.00	1.00	1.00	1.00
LongRunEmphasis	GLRLM	1.00	1.00	0.99	0.97	1.00	0.99	0.56	0.44	1.00	1.00
LongRunHighGrayLevelEmpha	GLRLM	0.99	0.99	0.95	0.86	0.98	0.97	0.77	0.51	0.99	0.99
LongRunLowGrayLevelEmpha	GLRLM	1.00	1.00	0.99	0.98	0.99	0.99	0.60	0.45	1.00	1.00
LowGrayLevelRunEmpha	GLRLM	1.00	1.00	0.99	1.00	1.00	1.00	0.98	0.99	1.00	1.00
RunLengthNonuniformity	GLRLM	1.00	1.00	0.98	0.98	0.99	1.00	0.99	0.83	0.99	0.99
RunPercentage	GLRLM	1.00	1.00	0.99	0.99	0.99	1.00	0.96	0.98	1.00	1.00
ShortRunEmphasis	GLRLM	0.99	1.00	0.97	0.98	0.99	0.99	0.93	0.96	0.99	0.99
ShortRunHighGrayLevelEmpha	GLRLM	0.99	1.00	0.97	0.98	0.97	1.00	0.99	0.97	1.00	1.00
ShortRunLowGrayLevelEmpha	GLRLM	1.00	1.00	0.98	0.93	1.00	0.98	0.87	0.48	1.00	0.99
EnergyNorm	IH	1.00	1.00	0.98	0.99	0.98	1.00	1.00	1.00	1.00	1.00
GlobalEntropy	IH	1.00	0.99	0.96	0.98	0.98	0.99	1.00	1.00	1.00	0.99
GlobalStd	IH	0.98	1.00	0.88	0.96	0.96	0.99	1.00	1.00	0.99	1.00
GlobalUniformity	IH	1.00	0.99	0.97	0.98	1.00	0.99	1.00	1.00	1.00	0.99
Kurtosis	IH	1.00	1.00	0.94	0.94	0.99	0.99	1.00	1.00	1.00	1.00
Skewness	IH	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00
Variance	IH	0.98	1.00	0.89	0.95	0.96	0.99	1.00	1.00	0.99	1.00
Busyness	NGTDM	0.98	0.90	0.89	0.57	0.94	0.65	0.86	0.60	0.99	0.98
Coarseness	NGTDM	1.00	1.00	0.97	0.91	0.98	0.96	0.90	0.95	1.00	1.00
Complexity	NGTDM	0.66	0.93	0.25	0.66	0.76	0.84	0.12	0.43	0.97	0.98
Contrast	NGTDM	0.98	0.97	0.78	0.79	0.94	0.95	0.69	0.59	0.99	0.99
TextureStrength	NGTDM	0.99	0.99	0.97	0.97	0.99	0.99	0.69	0.76	1.00	0.99
Feature	Category	0.99	1.00	0.98	0.99	0.97	1.00	1.00	1.00	1.00	1.00
AutoCorrelation	GLCM	0.96	1.00	0.90	0.97	0.91	0.99	0.99	0.99	0.99	0.99
ClusterProminence	GLCM	1.00	1.00	0.96	0.99	0.98	1.00	0.99	1.00	1.00	1.00
ClusterShade	GLCM	0.98	1.00	0.95	0.97	0.97	0.99	0.97	0.98	0.99	0.99

Filter cutoff (subset): cutoff values < 6 mm

Table E-3. GE Scanner ICC Values for Protocol Parameters Changed Contd.

Feature	Category	Matrix size		Matrix size resampled pixel		Type of recon		Type of recon w/o Q.Clear	
		64 levels	fixed bin	64 levels	fixed bin	64 levels	fixed bin	64 levels	fixed bin
AutoCorrelation	GLCM	0.95	1.00	0.99	1.00	0.96	0.98	0.99	1.00
ClusterProminence	GLCM	0.92	0.99	0.98	0.99	0.79	0.77	0.88	0.90
ClusterShade	GLCM	0.96	0.98	1.00	1.00	0.98	0.96	0.98	0.98
ClusterTendency	GLCM	0.94	0.95	0.98	0.99	0.76	0.75	0.90	0.91
Contrast	GLCM	0.68	0.61	0.99	1.00	0.92	0.91	0.99	0.99
Correlation	GLCM	0.64	0.64	1.00	1.00	0.94	0.94	0.98	0.98
DifferenceEntropy	GLCM	0.99	1.00	1.00	1.00	0.99	0.99	1.00	1.00
Dissimilarity	GLCM	0.90	0.88	1.00	1.00	0.98	0.98	0.99	1.00
Energy	GLCM	0.95	0.95	0.99	1.00	0.97	0.99	0.99	0.99
Entropy	GLCM	0.99	1.00	1.00	1.00	0.99	0.99	1.00	1.00
Homogeneity	GLCM	0.99	1.00	1.00	1.00	0.99	0.99	1.00	1.00
Homogeneity2	GLCM	0.99	0.99	1.00	1.00	0.99	1.00	1.00	1.00
InformationMeasureCorr1	GLCM	0.99	0.99	1.00	1.00	0.98	0.97	1.00	1.00
InformationMeasureCorr2	GLCM	0.63	0.49	0.99	0.99	0.91	0.87	0.95	0.94
InverseDiffMomentNorm	GLCM	0.74	0.74	0.99	0.99	0.95	0.94	0.99	0.98
InverseDiffNorm	GLCM	0.94	0.94	1.00	1.00	0.99	0.98	1.00	0.99
InverseVariance	GLCM	0.96	0.49	0.99	0.96	0.98	0.88	0.99	0.91
MaxProbability	GLCM	0.97	0.97	0.99	1.00	0.99	0.99	0.99	1.00
SumAverage	GLCM	0.98	1.00	1.00	1.00	0.98	0.99	1.00	1.00
SumEntropy	GLCM	0.97	0.98	0.99	1.00	0.99	0.99	1.00	1.00
SumVariance	GLCM	0.95	1.00	0.99	1.00	0.95	0.98	0.99	1.00
Variance	GLCM	0.94	0.95	0.98	0.99	0.76	0.75	0.90	0.91
GrayLevelNonuniformity	GLRLM	0.94	0.47	0.99	1.00	0.99	0.91	1.00	0.99
HighGrayLevelRunEmpha	GLRLM	0.97	1.00	0.99	1.00	0.96	0.99	0.99	0.99
LongRunEmphasis	GLRLM	0.83	0.72	1.00	0.99	0.93	0.99	0.95	0.99
LongRunHighGrayLevelEmpha	GLRLM	0.94	0.77	0.99	0.99	0.95	0.98	0.98	0.99
LongRunLowGrayLevelEmpha	GLRLM	0.89	0.74	0.99	0.99	0.90	0.99	0.92	0.99
LowGrayLevelRunEmpha	GLRLM	1.00	0.99	1.00	1.00	0.99	0.99	1.00	1.00
RunLengthNonuniformity	GLRLM	0.98	0.98	0.99	1.00	0.98	0.99	0.99	0.99
RunPercentage	GLRLM	0.97	0.99	0.99	1.00	0.97	0.99	0.99	1.00
ShortRunEmphasis	GLRLM	0.95	0.95	0.98	1.00	0.98	0.99	0.99	1.00
ShortRunHighGrayLevelEmpha	GLRLM	0.97	0.99	0.99	1.00	0.96	0.99	0.99	0.99
ShortRunLowGrayLevelEmpha	GLRLM	0.99	0.63	1.00	1.00	0.99	0.90	0.99	0.95
EnergyNorm	IH	0.97	1.00	0.99	1.00	0.96	0.98	0.99	1.00
GlobalEntropy	IH	0.98	0.99	0.99	0.99	0.96	0.97	0.99	0.98
GlobalStd	IH	0.94	0.99	0.96	1.00	0.81	0.88	0.93	0.97
GlobalUniformity	IH	1.00	0.99	0.99	0.99	0.97	0.97	0.99	0.98
Kurtosis	IH	1.00	1.00	1.00	1.00	0.97	0.97	0.99	0.99
Skewness	IH	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00
Variance	IH	0.94	0.99	0.97	1.00	0.77	0.83	0.92	0.96
Busyness	NGTDM	0.76	0.40	0.98	0.93	0.83	0.75	0.96	0.87
Coarseness	NGTDM	0.98	0.87	1.00	1.00	0.97	0.94	0.99	0.99
Complexity	NGTDM	0.00	0.33	0.92	0.94	0.51	0.72	0.84	0.90
Contrast	NGTDM	0.72	0.65	0.97	0.97	0.73	0.77	0.95	0.95
TextureStrength	NGTDM	0.85	0.92	1.00	0.98	0.99	0.96	0.99	0.95
Feature	Category	64 levels	fixed bin	64 levels	fixed bin	64 levels	fixed bin	64 levels	fixed bin
AutoCorrelation	GLCM	0.95	1.00	0.99	1.00	0.96	0.98	0.99	1.00
ClusterProminence	GLCM	0.92	0.99	0.98	0.99	0.79	0.77	0.88	0.90
ClusterShade	GLCM	0.96	0.98	1.00	1.00	0.98	0.96	0.98	0.98

Table E-4. GE Scanner ICC Values for Protocol Parameters Changed Contd.

Feature	Category	Iterations and subsets		Iterations and subsets (subset)		Z smoothing	
		64 levels	fixed bin	64 levels	fixed bin	64 levels	fixed bin
AutoCorrelation	GLCM	0.94	0.98	0.98	1.00	0.99	1.00
ClusterProminence	GLCM	0.66	0.89	0.89	1.00	0.87	0.84
ClusterShade	GLCM	0.78	0.86	0.97	0.99	0.99	0.97
ClusterTendency	GLCM	0.40	0.76	0.90	0.99	0.89	0.88
Contrast	GLCM	0.79	0.72	0.96	0.93	0.99	1.00
Correlation	GLCM	0.76	0.76	0.95	0.95	1.00	1.00
DifferenceEntropy	GLCM	0.97	0.98	1.00	1.00	1.00	1.00
Dissimilarity	GLCM	0.89	0.86	0.98	0.97	1.00	1.00
Energy	GLCM	0.97	0.97	1.00	1.00	0.98	0.99
Entropy	GLCM	0.98	0.98	1.00	1.00	0.99	1.00
Homogeneity	GLCM	0.97	0.97	1.00	1.00	0.99	1.00
Homogeneity2	GLCM	0.97	0.96	1.00	1.00	0.99	1.00
InformationMeasureCorr1	GLCM	0.97	0.97	1.00	1.00	1.00	1.00
InformationMeasureCorr2	GLCM	0.66	0.62	0.92	0.90	1.00	1.00
InverseDiffMomentNorm	GLCM	0.80	0.80	0.96	0.95	0.99	0.99
InverseDiffNorm	GLCM	0.91	0.91	0.98	0.98	1.00	1.00
InverseVariance	GLCM	0.85	0.35	0.98	0.85	0.99	0.96
MaxProbability	GLCM	0.97	0.97	1.00	1.00	1.00	1.00
SumAverage	GLCM	0.96	0.98	0.99	1.00	1.00	1.00
SumEntropy	GLCM	0.98	0.99	1.00	1.00	0.99	1.00
SumVariance	GLCM	0.93	0.97	0.98	1.00	0.98	0.99
Variance	GLCM	0.40	0.76	0.90	0.99	0.89	0.88
GrayLevelNonuniformity	GLRLM	0.78	0.38	0.96	0.88	0.99	1.00
HighGrayLevelRunEmpha	GLRLM	0.95	0.96	0.98	1.00	0.98	0.98
LongRunEmphasis	GLRLM	0.85	0.42	0.98	0.98	0.95	1.00
LongRunHighGrayLevelEmpha	GLRLM	0.84	0.43	0.97	0.96	0.98	0.98
LongRunLowGrayLevelEmpha	GLRLM	0.89	0.43	1.00	0.98	0.92	1.00
LowGrayLevelRunEmpha	GLRLM	0.89	0.97	0.99	1.00	0.99	1.00
RunLengthNonuniformity	GLRLM	0.95	0.85	0.99	0.97	0.99	1.00
RunPercentage	GLRLM	0.94	0.96	0.99	0.99	0.98	1.00
ShortRunEmphasis	GLRLM	0.91	0.98	0.99	0.99	0.98	1.00
ShortRunHighGrayLevelEmpha	GLRLM	0.96	0.89	0.99	0.99	0.98	0.99
ShortRunLowGrayLevelEmpha	GLRLM	0.80	0.44	0.98	0.95	1.00	0.99
EnergyNorm	IH	0.95	0.97	0.98	1.00	0.99	1.00
GlobalEntropy	IH	0.93	0.93	0.98	0.98	0.99	0.98
GlobalStd	IH	0.61	0.90	0.90	0.99	0.90	0.94
GlobalUniformity	IH	0.94	0.93	0.99	0.98	0.99	0.98
Kurtosis	IH	0.84	0.84	0.99	0.99	0.98	0.98
Skewness	IH	0.92	0.92	0.99	0.99	1.00	1.00
Variance	IH	0.58	0.91	0.89	0.99	0.89	0.91
Busyness	NGTDM	0.78	0.55	0.97	0.87	0.93	0.78
Coarseness	NGTDM	0.94	0.79	0.99	0.97	1.00	1.00
Complexity	NGTDM	0.30	0.62	0.50	0.88	0.85	0.90
Contrast	NGTDM	0.79	0.74	0.91	0.89	0.94	0.95
TextureStrength	NGTDM	0.96	0.79	0.99	0.92	1.00	0.95
Feature	Category	64 levels	fixed bin	64 levels	fixed bin	64 levels	fixed bin
AutoCorrelation	GLCM	0.94	0.98	0.98	1.00	0.99	1.00
ClusterProminence	GLCM	0.66	0.89	0.89	1.00	0.87	0.84
ClusterShade	GLCM	0.78	0.86	0.97	0.99	0.99	0.97

Iterations and subsets (subset): $16 \leq \text{effective iterations} \leq 36$

Table E-5. Siemens Continuous Bed Motion Scanner ICC Values for Protocol Parameters Changed

		Time per bed position		Filter cutoff		Filter cutoff (subset)		Matrix size		Matrix size resampled pixel	
Feature	Category	64 levels	fixed bin	64 levels	fixed bin	64 levels	fixed bin	64 levels	fixed bin	64 levels	fixed bin
AutoCorrelation	GLCM	0.99	1.00	0.89	0.98	0.96	0.99	0.94	1.00	0.94	1.00
ClusterProminence	GLCM	0.98	1.00	0.76	0.73	0.91	0.91	0.94	0.95	0.92	0.99
ClusterShade	GLCM	1.00	1.00	0.88	0.93	0.95	0.98	0.95	0.96	1.00	1.00
ClusterTendency	GLCM	0.99	1.00	0.80	0.76	0.93	0.94	0.94	0.95	0.95	0.98
Contrast	GLCM	0.99	1.00	0.91	0.86	0.94	0.98	0.58	0.50	0.92	0.98
Correlation	GLCM	1.00	1.00	0.91	0.91	0.98	0.98	0.48	0.48	0.99	0.99
DifferenceEntropy	GLCM	0.99	1.00	0.87	0.95	0.91	0.98	0.98	0.99	0.96	1.00
Dissimilarity	GLCM	0.99	1.00	0.89	0.97	0.93	0.99	0.86	0.80	0.95	0.99
Energy	GLCM	1.00	1.00	0.77	0.98	0.91	0.99	0.72	0.74	0.98	0.99
Entropy	GLCM	0.99	1.00	0.86	0.97	0.94	0.99	0.98	1.00	0.98	0.99
Homogeneity	GLCM	1.00	1.00	0.83	0.92	0.96	0.98	0.95	0.98	0.98	0.99
Homogeneity2	GLCM	1.00	1.00	0.83	0.90	0.96	0.98	0.93	0.96	0.98	0.99
InformationMeasureCorr1	GLCM	1.00	1.00	0.93	0.88	0.99	0.98	0.91	0.99	0.99	0.99
InformationMeasureCorr2	GLCM	1.00	1.00	0.86	0.84	0.97	0.96	0.43	0.39	0.99	0.98
InverseDiffMomentNorm	GLCM	0.99	0.98	0.91	0.90	0.93	0.93	0.65	0.65	0.93	0.92
InverseDiffNorm	GLCM	0.99	0.99	0.86	0.86	0.93	0.92	0.90	0.90	0.96	0.95
InverseVariance	GLCM	0.99	1.00	0.85	0.84	0.96	0.97	0.86	0.72	0.96	0.98
MaxProbability	GLCM	0.97	1.00	0.62	0.94	0.87	0.99	0.75	0.84	0.82	0.98
SumAverage	GLCM	0.99	1.00	0.94	0.99	0.98	1.00	0.98	1.00	0.97	1.00
SumEntropy	GLCM	0.98	1.00	0.77	0.98	0.88	0.99	0.77	0.96	0.89	1.00
SumVariance	GLCM	0.99	1.00	0.88	0.98	0.95	0.99	0.95	1.00	0.93	1.00
Variance	GLCM	0.99	1.00	0.80	0.76	0.93	0.94	0.94	0.95	0.95	0.98
GrayLevelNonuniformity	GLRLM	1.00	1.00	0.85	0.93	0.98	0.99	0.80	0.54	0.95	0.99
HighGrayLevelRunEmpha	GLRLM	0.99	1.00	0.88	0.98	0.96	0.99	0.96	1.00	0.94	1.00
LongRunEmphasis	GLRLM	0.96	0.99	0.82	0.88	0.95	0.98	0.58	0.57	0.93	0.97
LongRunHighGrayLevelEmpha	GLRLM	0.99	0.99	0.87	0.87	0.96	0.98	0.71	0.66	0.91	0.97
LongRunLowGrayLevelEmpha	GLRLM	0.98	1.00	0.77	0.95	0.91	0.98	0.70	0.62	0.91	0.99
LowGrayLevelRunEmpha	GLRLM	1.00	1.00	0.84	0.99	0.92	1.00	0.95	0.95	0.97	1.00
RunLengthNonuniformity	GLRLM	0.98	1.00	0.80	0.92	0.94	0.98	0.86	0.94	0.93	0.97
RunPercentage	GLRLM	0.98	1.00	0.83	0.91	0.95	0.98	0.80	0.95	0.95	0.98
ShortRunEmphasis	GLRLM	0.98	1.00	0.80	0.89	0.93	0.96	0.81	0.88	0.93	0.97
ShortRunHighGrayLevelEmpha	GLRLM	0.99	1.00	0.88	0.99	0.95	0.99	0.96	0.99	0.95	0.99
ShortRunLowGrayLevelEmpha	GLRLM	1.00	1.00	0.86	0.97	0.92	0.99	0.81	0.51	0.97	0.98
EnergyNorm	IH	0.99	1.00	0.90	0.99	0.96	1.00	0.96	1.00	0.94	1.00
GlobalEntropy	IH	0.96	0.97	0.61	0.74	0.72	0.98	0.87	0.95	0.83	0.91
GlobalStd	IH	0.98	1.00	0.72	0.87	0.90	0.97	0.94	0.99	0.90	0.98
GlobalUniformity	IH	0.98	0.97	0.60	0.76	0.77	0.98	0.92	0.96	0.90	0.92
Kurtosis	IH	1.00	1.00	0.78	0.78	0.89	0.89	0.98	0.98	0.96	0.96
Skewness	IH	1.00	1.00	0.93	0.93	0.99	0.99	1.00	1.00	0.99	0.99
Variance	IH	0.98	1.00	0.70	0.83	0.90	0.95	0.94	0.99	0.90	0.98
Busyness	NGTDM	0.99	0.97	0.89	0.84	0.93	0.86	0.73	0.59	0.97	0.93
Coarseness	NGTDM	1.00	1.00	0.77	0.61	0.93	0.88	0.96	0.84	0.98	0.97
Complexity	NGTDM	0.97	0.98	0.83	0.72	0.77	0.87	0.17	0.40	0.95	0.96
Contrast	NGTDM	0.95	0.97	0.64	0.81	0.87	0.94	0.67	0.53	0.74	0.88
TextureStrength	NGTDM	1.00	0.98	0.96	0.87	0.99	0.94	0.78	0.83	0.99	0.98
Feature	Category	64 levels	fixed bin	64 levels	fixed bin	64 levels	fixed bin	64 levels	fixed bin	64 levels	fixed bin
AutoCorrelation	GLCM	0.99	1.00	0.89	0.98	0.96	0.99	0.94	1.00	0.94	1.00
ClusterProminence	GLCM	0.98	1.00	0.76	0.73	0.91	0.91	0.94	0.95	0.92	0.99
ClusterShade	GLCM	1.00	1.00	0.88	0.93	0.95	0.98	0.95	0.96	1.00	1.00

Filter cutoff (subset): cutoff values < 6 mm

Table E-6. Siemens Continuous Bed Motion Scanner ICC Values for Protocol Parameters Changed Contd.

		Type of recon		Iterations		Iterations and subsets non-TOF		Iterations and subsets non-TOF (subset)	
Feature	Category	64 levels	fixed bin	64 levels	fixed bin	64 levels	fixed bin	64 levels	fixed bin
AutoCorrelation	GLCM	0.89	0.98	0.93	1.00	0.62	0.97	0.92	1.00
ClusterProminence	GLCM	0.91	0.94	0.86	0.99	0.60	0.86	0.83	0.98
ClusterShade	GLCM	0.97	0.97	0.92	0.95	0.50	0.68	0.90	0.95
ClusterTendency	GLCM	0.95	0.96	0.89	0.97	0.17	0.66	0.90	0.94
Contrast	GLCM	0.94	0.96	0.91	0.96	0.76	0.70	0.95	0.94
Correlation	GLCM	0.97	0.97	0.97	0.97	0.73	0.72	0.94	0.94
DifferenceEntropy	GLCM	0.94	0.99	0.95	0.99	0.82	0.96	0.95	1.00
Dissimilarity	GLCM	0.94	0.99	0.93	0.98	0.86	0.88	0.95	0.97
Energy	GLCM	0.44	0.56	0.96	0.90	0.31	0.23	0.84	0.82
Entropy	GLCM	0.84	0.96	0.96	0.99	0.80	0.92	0.96	0.99
Homogeneity	GLCM	0.90	0.95	0.96	0.98	0.65	0.84	0.96	0.98
Homogeneity2	GLCM	0.90	0.95	0.97	0.97	0.60	0.83	0.95	0.98
InformationMeasureCorr1	GLCM	0.97	0.96	0.99	0.98	0.86	0.85	0.99	0.99
InformationMeasureCorr2	GLCM	0.98	0.97	0.91	0.90	0.59	0.56	0.90	0.90
InverseDiffMomentNorm	GLCM	0.95	0.94	0.92	0.92	0.79	0.80	0.95	0.95
InverseDiffNorm	GLCM	0.94	0.94	0.93	0.93	0.86	0.87	0.95	0.95
InverseVariance	GLCM	0.96	0.96	0.96	0.95	0.78	0.08	0.92	0.93
MaxProbability	GLCM	0.49	0.56	0.86	0.93	0.34	0.47	0.86	0.89
SumAverage	GLCM	0.93	0.98	0.97	1.00	0.71	0.96	0.96	1.00
SumEntropy	GLCM	0.64	0.94	0.92	0.98	0.56	0.79	0.90	0.99
SumVariance	GLCM	0.89	0.99	0.91	1.00	0.62	0.97	0.91	1.00
Variance	GLCM	0.95	0.96	0.89	0.97	0.17	0.66	0.90	0.94
GrayLevelNonuniformity	GLRLM	0.84	0.92	0.98	0.96	0.98	0.55	0.99	0.93
HighGrayLevelRunEmpha	GLRLM	0.91	0.99	0.94	1.00	0.70	0.97	0.92	1.00
LongRunEmphasis	GLRLM	0.63	0.86	0.89	0.93	0.18	0.10	0.85	0.91
LongRunHighGrayLevelEmpha	GLRLM	0.89	0.98	0.92	0.96	0.50	0.51	0.88	0.96
LongRunLowGrayLevelEmpha	GLRLM	0.52	0.81	0.92	0.92	0.36	0.12	0.94	0.91
LowGrayLevelRunEmpha	GLRLM	0.87	0.97	0.97	0.99	0.81	0.94	0.97	1.00
RunLengthNonuniformity	GLRLM	0.81	0.96	0.90	0.95	0.35	0.90	0.91	0.97
RunPercentage	GLRLM	0.75	0.94	0.92	0.96	0.36	0.83	0.93	0.98
ShortRunEmphasis	GLRLM	0.81	0.94	0.89	0.92	0.24	0.70	0.92	0.99
ShortRunHighGrayLevelEmpha	GLRLM	0.91	0.99	0.94	1.00	0.71	0.90	0.92	0.99
ShortRunLowGrayLevelEmpha	GLRLM	0.91	0.91	0.97	0.92	0.81	0.29	0.97	0.93
EnergyNorm	IH	0.89	0.98	0.93	1.00	0.66	0.97	0.92	1.00
GlobalEntropy	IH	0.67	0.86	0.73	0.98	0.47	0.30	0.90	0.92
GlobalStd	IH	0.86	0.96	0.82	0.98	0.10	0.74	0.86	0.96
GlobalUniformity	IH	0.73	0.86	0.81	0.98	0.51	0.38	0.93	0.92
Kurtosis	IH	0.93	0.93	0.90	0.90	0.00	0.00	0.81	0.81
Skewness	IH	0.96	0.96	0.97	0.97	0.65	0.65	0.95	0.95
Variance	IH	0.85	0.95	0.82	0.98	0.09	0.79	0.85	0.96
Busyness	NGTDM	0.95	0.92	0.92	0.90	0.74	0.79	0.98	0.93
Coarseness	NGTDM	0.95	0.97	0.91	0.85	0.81	0.70	0.96	0.94
Complexity	NGTDM	0.92	0.94	0.94	0.92	0.58	0.66	0.96	0.89
Contrast	NGTDM	0.69	0.86	0.67	0.87	0.10	0.66	0.56	0.82
TextureStrength	NGTDM	1.00	0.94	0.98	0.97	0.97	0.75	0.99	0.97
Feature	Category	64 levels	fixed bin	64 levels	fixed bin	64 levels	fixed bin	64 levels	fixed bin
AutoCorrelation	GLCM	0.89	0.98	0.93	1.00	0.62	0.97	0.92	1.00
ClusterProminence	GLCM	0.91	0.94	0.86	0.99	0.60	0.86	0.83	0.98
ClusterShade	GLCM	1.00	1.00	0.88	0.93	0.95	0.98	0.95	0.96

Iterations and subsets (subset): $16 \leq \text{effective iterations} \leq 24$

Table E-7. Siemens Scanner ICC Values for Protocol Parameters Changed

		Time per bed position		Filter cutoff		Filter cutoff (subset)		Matrix size		Matrix size resampled pixel	
Feature	Category	64 levels	fixed bin	64 levels	fixed bin	64 levels	fixed bin	64 levels	fixed bin	64 levels	fixed bin
AutoCorrelation	GLCM	0.99	1.00	0.92	0.99	0.97	0.99	0.97	1.00	0.98	1.00
ClusterProminence	GLCM	0.99	1.00	0.70	0.51	0.97	0.78	0.90	0.94	0.92	0.98
ClusterShade	GLCM	1.00	1.00	0.95	0.88	0.98	0.95	0.97	0.97	1.00	1.00
ClusterTendency	GLCM	0.99	1.00	0.71	0.60	0.98	0.87	0.87	0.93	0.94	0.98
Contrast	GLCM	1.00	1.00	0.92	0.87	0.97	0.98	0.60	0.51	0.95	0.98
Correlation	GLCM	1.00	1.00	0.90	0.90	0.98	0.98	0.48	0.48	0.99	0.99
DifferenceEntropy	GLCM	1.00	1.00	0.93	0.94	0.96	0.98	0.98	0.99	0.99	0.99
Dissimilarity	GLCM	1.00	1.00	0.93	0.96	0.97	0.99	0.86	0.81	0.98	0.99
Energy	GLCM	1.00	1.00	0.77	0.93	0.87	0.96	0.74	0.74	0.99	0.99
Entropy	GLCM	1.00	1.00	0.90	0.96	0.96	0.99	0.99	1.00	0.99	0.99
Homogeneity	GLCM	1.00	1.00	0.92	0.94	0.98	0.99	0.97	0.99	0.99	0.99
Homogeneity2	GLCM	1.00	1.00	0.92	0.93	0.98	0.99	0.95	0.97	0.99	0.99
InformationMeasureCorr1	GLCM	1.00	1.00	0.93	0.91	0.99	0.98	0.96	0.99	0.99	0.99
InformationMeasureCorr2	GLCM	1.00	1.00	0.86	0.85	0.97	0.97	0.46	0.40	0.99	0.98
InverseDiffMomentNorm	GLCM	1.00	0.99	0.93	0.92	0.97	0.96	0.66	0.65	0.96	0.95
InverseDiffNorm	GLCM	1.00	1.00	0.92	0.92	0.97	0.97	0.90	0.91	0.98	0.98
InverseVariance	GLCM	1.00	1.00	0.90	0.87	0.98	0.98	0.85	0.63	0.99	0.97
MaxProbability	GLCM	0.97	0.99	0.75	0.93	0.82	0.90	0.76	0.78	0.87	0.97
SumAverage	GLCM	1.00	1.00	0.97	0.99	0.99	1.00	0.98	1.00	0.99	1.00
SumEntropy	GLCM	0.99	1.00	0.82	0.96	0.93	0.99	0.76	0.93	0.94	0.99
SumVariance	GLCM	0.99	1.00	0.92	0.98	0.97	0.99	0.97	1.00	0.98	1.00
Variance	GLCM	0.99	1.00	0.71	0.60	0.98	0.87	0.87	0.93	0.94	0.98
GrayLevelNonuniformity	GLRLM	1.00	1.00	0.87	0.92	0.97	0.99	0.80	0.53	0.97	0.97
HighGrayLevelRunEmpha	GLRLM	0.99	1.00	0.92	0.97	0.97	0.98	0.98	0.99	0.98	1.00
LongRunEmphasis	GLRLM	0.99	1.00	0.93	0.94	0.96	0.99	0.62	0.59	0.97	0.99
LongRunHighGrayLevelEmpha	GLRLM	0.99	0.99	0.88	0.82	0.98	0.99	0.80	0.62	0.95	0.96
LongRunLowGrayLevelEmpha	GLRLM	1.00	1.00	0.80	0.95	0.85	0.97	0.73	0.62	0.96	0.99
LowGrayLevelRunEmpha	GLRLM	1.00	1.00	0.85	0.99	0.93	1.00	0.98	0.96	0.98	1.00
RunLengthNonuniformity	GLRLM	0.99	1.00	0.90	0.92	0.98	0.99	0.92	0.94	0.97	0.98
RunPercentage	GLRLM	0.99	1.00	0.93	0.93	0.98	0.99	0.87	0.96	0.98	0.99
ShortRunEmphasis	GLRLM	0.99	1.00	0.91	0.90	0.98	0.98	0.88	0.90	0.97	0.97
ShortRunHighGrayLevelEmpha	GLRLM	0.99	1.00	0.92	0.98	0.97	0.99	0.98	0.99	0.98	1.00
ShortRunLowGrayLevelEmpha	GLRLM	0.99	1.00	0.85	0.97	0.93	1.00	0.93	0.56	0.98	0.99
EnergyNorm	IH	0.99	1.00	0.94	0.99	0.98	1.00	0.98	1.00	0.98	1.00
GlobalEntropy	IH	0.97	1.00	0.65	0.68	0.83	0.86	0.87	0.92	0.85	0.96
GlobalStd	IH	0.99	1.00	0.63	0.72	0.96	0.91	0.89	0.99	0.88	0.98
GlobalUniformity	IH	0.99	1.00	0.59	0.70	0.79	0.86	0.90	0.93	0.90	0.96
Kurtosis	IH	1.00	1.00	0.74	0.74	0.85	0.85	0.98	0.98	0.96	0.96
Skewness	IH	1.00	1.00	0.95	0.95	0.99	0.99	1.00	1.00	1.00	1.00
Variance	IH	0.99	1.00	0.61	0.68	0.96	0.88	0.89	0.99	0.89	0.98
Busyness	NGTDM	0.99	0.99	0.90	0.73	0.93	0.71	0.71	0.50	0.97	0.92
Coarseness	NGTDM	1.00	1.00	0.84	0.60	0.95	0.87	0.94	0.81	0.99	0.97
Complexity	NGTDM	0.97	0.97	0.80	0.78	0.76	0.89	0.19	0.35	0.93	0.95
Contrast	NGTDM	0.98	0.98	0.76	0.85	0.92	0.96	0.69	0.56	0.91	0.94
TextureStrength	NGTDM	1.00	1.00	0.97	0.85	0.99	0.94	0.75	0.91	0.99	0.97
Feature	Category	0.99	1.00	0.92	0.99	0.97	0.99	0.97	1.00	0.98	1.00
AutoCorrelation	GLCM	0.99	1.00	0.70	0.51	0.97	0.78	0.90	0.94	0.92	0.98
ClusterProminence	GLCM	1.00	1.00	0.95	0.88	0.98	0.95	0.97	0.97	1.00	1.00
ClusterShade	GLCM	0.99	1.00	0.71	0.60	0.98	0.87	0.87	0.93	0.94	0.98

Filter cutoff (subset): cutoff values < 6 mm

Table E-8. Siemens Scanner ICC Values for Protocol Parameters Changed Contd.

		Type of recon		Iterations		Iterations and subsets non-TOF		Iterations and subsets non-TOF (subset)	
Feature	Category	64 levels	fixed bin	64 levels	fixed bin	64 levels	fixed bin	64 levels	fixed bin
AutoCorrelation	GLCM	0.96	0.99	0.95	1.00	0.80	0.97	0.95	1.00
ClusterProminence	GLCM	0.88	0.83	0.87	0.98	0.34	0.77	0.65	0.97
ClusterShade	GLCM	0.99	0.96	0.96	0.97	0.52	0.74	0.93	0.96
ClusterTendency	GLCM	0.91	0.88	0.82	0.96	0.09	0.48	0.69	0.86
Contrast	GLCM	0.97	0.97	0.93	0.96	0.77	0.69	0.94	0.95
Correlation	GLCM	0.97	0.97	0.97	0.97	0.72	0.71	0.93	0.93
DifferenceEntropy	GLCM	0.99	0.99	0.98	0.99	0.96	0.97	0.98	1.00
Dissimilarity	GLCM	0.99	0.99	0.96	0.98	0.90	0.87	0.97	0.98
Energy	GLCM	0.86	0.84	0.91	0.86	0.36	0.20	0.83	0.76
Entropy	GLCM	0.98	0.99	0.97	0.98	0.91	0.92	0.98	0.99
Homogeneity	GLCM	0.99	0.99	0.99	0.99	0.89	0.89	0.99	0.98
Homogeneity2	GLCM	0.99	0.99	0.99	0.98	0.87	0.89	0.99	0.98
InformationMeasureCorr1	GLCM	0.99	0.99	0.99	0.99	0.90	0.88	0.99	0.99
InformationMeasureCorr2	GLCM	0.98	0.98	0.91	0.91	0.58	0.56	0.88	0.88
InverseDiffMomentNorm	GLCM	0.98	0.97	0.94	0.94	0.80	0.78	0.95	0.94
InverseDiffNorm	GLCM	0.99	0.99	0.97	0.97	0.92	0.92	0.97	0.97
InverseVariance	GLCM	0.99	0.98	0.98	0.94	0.90	0.10	0.97	0.91
MaxProbability	GLCM	0.71	0.77	0.78	0.83	0.40	0.40	0.82	0.81
SumAverage	GLCM	0.98	0.99	0.98	1.00	0.83	0.97	0.98	1.00
SumEntropy	GLCM	0.94	0.98	0.91	0.98	0.75	0.76	0.93	0.98
SumVariance	GLCM	0.96	0.99	0.94	1.00	0.80	0.97	0.94	1.00
Variance	GLCM	0.91	0.88	0.82	0.96	0.09	0.48	0.69	0.86
GrayLevelNonuniformity	GLRLM	0.93	0.91	0.98	0.94	0.97	0.49	0.99	0.92
HighGrayLevelRunEmpha	GLRLM	0.96	0.99	0.95	1.00	0.83	0.96	0.94	1.00
LongRunEmphasis	GLRLM	0.97	0.98	0.95	0.97	0.62	0.19	0.97	0.96
LongRunHighGrayLevelEmpha	GLRLM	0.95	0.96	0.94	0.96	0.59	0.41	0.90	0.97
LongRunLowGrayLevelEmpha	GLRLM	0.92	0.98	0.95	0.95	0.67	0.15	0.96	0.92
LowGrayLevelRunEmpha	GLRLM	0.95	0.98	0.98	0.99	0.88	0.96	0.96	1.00
RunLengthNonuniformity	GLRLM	0.97	0.99	0.98	0.97	0.74	0.88	0.96	0.98
RunPercentage	GLRLM	0.98	0.99	0.98	0.98	0.76	0.88	0.97	0.98
ShortRunEmphasis	GLRLM	0.96	0.98	0.98	0.95	0.63	0.78	0.95	0.98
ShortRunHighGrayLevelEmpha	GLRLM	0.96	0.99	0.95	0.99	0.84	0.89	0.94	0.99
ShortRunLowGrayLevelEmpha	GLRLM	0.94	0.95	0.97	0.93	0.79	0.38	0.92	0.93
EnergyNorm	IH	0.97	0.99	0.95	1.00	0.81	0.97	0.95	1.00
GlobalEntropy	IH	0.87	0.83	0.70	0.95	0.55	0.19	0.80	0.84
GlobalStd	IH	0.81	0.90	0.71	0.97	0.11	0.58	0.62	0.91
GlobalUniformity	IH	0.90	0.84	0.73	0.95	0.54	0.23	0.88	0.85
Kurtosis	IH	0.95	0.95	0.86	0.86	0.18	0.18	0.86	0.86
Skewness	IH	0.99	0.99	0.98	0.98	0.71	0.71	0.97	0.97
Variance	IH	0.82	0.90	0.71	0.97	0.11	0.66	0.61	0.90
Busyness	NGTDM	0.97	0.89	0.94	0.91	0.81	0.74	0.97	0.93
Coarseness	NGTDM	0.99	0.97	0.95	0.86	0.89	0.72	0.97	0.95
Complexity	NGTDM	0.91	0.88	0.91	0.91	0.51	0.64	0.92	0.90
Contrast	NGTDM	0.88	0.93	0.80	0.89	0.54	0.68	0.70	0.87
TextureStrength	NGTDM	0.99	0.92	0.97	0.98	0.97	0.79	0.99	0.98
Feature	Category	0.96	0.99	0.95	1.00	0.80	0.97	0.95	1.00
AutoCorrelation	GLCM	0.88	0.83	0.87	0.98	0.34	0.77	0.65	0.97
ClusterProminence	GLCM	0.99	0.96	0.96	0.97	0.52	0.74	0.93	0.96
ClusterShade	GLCM	0.91	0.88	0.82	0.96	0.09	0.48	0.69	0.86

Iterations and subsets (subset): 16 ≤ effective iterations ≤ 24

Table E-9. Interscanner Standard Deviation Compared to NSCLC Interpatient Standard Deviation

Feature	Category	64 levels	fixed bin
AutoCorrelation	GLCM	0.835	0.418
ClusterProminence	GLCM	1.856	0.099
ClusterShade	GLCM	2.987	0.576
ClusterTendendcy	GLCM	1.160	0.296
Contrast	GLCM	0.383	0.188
Correlation	GLCM	0.479	0.478
DifferenceEntropy	GLCM	0.776	0.602
Dissimilarity	GLCM	0.482	0.389
Energy	GLCM	3.697	3.117
Entropy	GLCM	0.702	0.708
Homogeneity	GLCM	1.107	0.890
Homogeneity2	GLCM	1.301	0.949
InformationMeasureCorr1	GLCM	0.421	0.732
InformationMeasureCorr2	GLCM	0.422	0.364
InverseDiffMomentNorm	GLCM	0.363	0.379
InverseDiffNorm	GLCM	0.510	0.533
InverseVariance	GLCM	0.884	0.474
MaxProbability	GLCM	2.584	3.498
SumAverage	GLCM	0.912	0.569
SumEntropy	GLCM	0.843	0.529
SumVariance	GLCM	0.805	0.377
Variance	GLCM	1.160	0.296
GrayLevelNonuniformity	GLRLM	1.354	0.266
HighGrayLevelRunEmpha	GLRLM	0.813	0.404
LongRunEmphasis	GLRLM	3.035	2.419
LongRunHighGrayLevelEmpha	GLRLM	0.738	0.508
LongRunLowGrayLevelEmpha	GLRLM	1.991	3.633
LowGrayLevelRunEmpha	GLRLM	1.447	1.023
RunLengthNonuniformity	GLRLM	1.257	0.759
RunPercentage	GLRLM	1.628	1.001
ShortRunEmphasis	GLRLM	1.347	0.888
ShortRunHighGrayLevelEmpha	GLRLM	0.826	0.395
ShortRunLowGrayLevelEmpha	GLRLM	1.400	0.625
EnergyNorm	IH	0.841	0.391
GlobalEntropy	IH	0.554	0.265
GlobalMax	IH	0.000	0.327
GlobalMean	IH	0.928	0.600
GlobalMedian	IH	1.264	0.785
GlobalMin	IH	0.127	0.002
GlobalStd	IH	0.518	0.342
GlobalUniformity	IH	0.591	0.334
Kurtosis	IH	0.378	0.382
Skewness	IH	1.014	1.019
Variance	IH	0.708	0.279
Busyness	NGTDM	0.290	0.121

Coarseness	NGTDM	0.983	1.739
Complexity	NGTDM	0.832	0.120
Contrast	NGTDM	0.182	0.432
TextureStrength	NGTDM	0.534	0.219

Each value is: standard deviation from standard-protocol scans of phantom/standard deviation from NSCLC patients

Highlighted values are those features with the phantom-scan mean more than 2 standard deviations away from patient-scan mean

Appendix F: Supplemental Material for Chapter 9

Table F-1. Radiomics Features used in CT Analysis

Gray Level Co-occurrence Matrix	Gray Level Run Length Matrix	Intensity Histogram	Neighborhood Gray Tone Difference Matrix
Auto Correlation	Gray Level Nonuniformity	Energy	Busyness
Cluster Prominence	High Gray Level Run Emphasis	Entropy	Coarseness
Cluster Shade	Long Run Emphasis	Kurtosis	Complexity
Cluster Tendency	Long Run High Gray Level Emphasis	Mean	Contrast
Contrast	Long Run Low Gray Level Emphasis	Median	Texture Strength
Correlation	Low Gray Level Run Emphasis	Minimum	
Difference Entropy	Run Length Nonuniformity	Skewness	
Dissimilarity	Run Percentage	Standard Deviation	
Energy	Short Run Emphasis	Uniformity	
Entropy	Short Run High Gray Level Emphasis	Variance	
Homogeneity	Short Run Low Gray Level Emphasis		
Homogeneity 2			
Information Measure Correlation 1			
Information Measure Correlation 2			
Inverse Difference Moment Norm			
Inverse Difference Norm			
Inverse Variance			
Max Probability			
Sum Average			
Sum Entropy			
Sum Variance			
Variance			

Table F-2. Radiomics Features used in PET Analysis

Gray Level Co-occurrence Matrix	Gray Level Run Length Matrix	Intensity Histogram	Neighborhood Gray Tone Difference Matrix
Auto Correlation	Gray Level Nonuniformity	Energy	Busyness
Cluster Prominence	High Gray Level Run Emphasis	Entropy	Coarseness
Cluster Shade	Long Run Emphasis	Kurtosis	Complexity
Cluster Tendency	Long Run High Gray Level Emphasis	Maximum	Contrast
Contrast	Long Run Low Gray Level Emphasis	Mean	Texture Strength
Correlation	Low Gray Level Run Emphasis	Median	
Difference Entropy	Run Length Nonuniformity	Minimum	
Dissimilarity	Run Percentage	Skewness	
Energy	Short Run Emphasis	Standard Deviation	
Entropy	Short Run High Gray Level Emphasis	Uniformity	
Homogeneity	Short Run Low Gray Level Emphasis	Variance	
Homogeneity 2			
Information Measure Correlation 1			
Information Measure Correlation 2			
Inverse Difference Moment Norm			
Inverse Difference Norm			
Inverse Variance			
Max Probability			
Sum Average			
Sum Entropy			
Sum Variance			
Variance			

Table F-3. Results of CT Patient Models

Subset of Patients	Patients in training	Patients in testing	Covariates in final model	p-value of covariates when fit on testing data	AUC
Same imaging protocol	260	251	HPV status Cluster tendency (GLCM) calculated using thresholding, smoothing, and bit depth resampling	p = 0.79 p = 0.90	0.55
Same imaging protocol HPV positive	168	152	Complexity (NGTDM) calculated using thresholding and smoothing	p = 0.16	0.65
Same imaging protocol HPV negative	92	99	Volume	p = 0.001	0.62
HPV positive	224	189	Volume Complexity (NGTDM) calculated using thresholding	p = 2.1×10^{-4} p = 0.26	0.75 (volume alone 0.76)
HPV negative	153	160	Volume Sum entropy (GLCM) calculated using thresholding	p = 3.6×10^{-4} p = 0.021	0.65 (volume alone 0.72)

Oropharynx	362	324	Volume	p = 0.013	0.69 (volume alone 0.71)
			HPV status	p = 0.19	
			Contrast (GLCM) calculated using thresholding	p = 0.94	
			Information measure correlation 2 (GLCM) calculated using thresholding, smoothing, and bit depth resampling	p = 0.060	
Oropharynx HPV positive	224	189	Volume	p = 2.1×10^{-4}	0.75 (volume alone 0.76)
			Complexity (NGTDM) calculated using thresholding	p = 0.26	
Oropharynx HPV negative	138	135	Volume	p = 0.012	0.64 (volume alone 0.68)
			Sum entropy (GLCM) calculated using thresholding	p = 0.13	

Table F-4. Results of PET Patient Models

Subset of Patients	Patients in training	Patients in testing	Covariates in final model	p-value of covariates when fit on testing data	AUC
Same imaging protocol	144	167	None		
Same imaging protocol HPV positive	117	137	None		
Same imaging protocol HPV negative	27	30	None		
HPV positive	207	206	Coarseness (NGTDM) calculated using 64 gray levels Sum average (GLCM) calculated using 64 gray levels	p = 0.28 p = 0.85	0.55
HPV negative	138	135	None		
Oropharynx	318	310	HPV status Coarseness (NGTDM) calculated using fixed bin width	p = 0.61 p = 0.20	0.58
Oropharynx HPV positive	206	206	Coarseness (NGTDM) calculated using 64 gray levels	p = 0.28	0.59
Oropharynx HPV negative	112	104	None		

Bibliography:

1. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D: **Global cancer statistics**. *CA: a cancer journal for clinicians* 2011, **61**(2):69-90.
2. **SEER Cancer Statistics Review, 1975-2014** [https://seer.cancer.gov/csr/1975_2014/]
3. Jensen SB, Pedersen AM, Vissink A, Andersen E, Brown CG, Davies AN, Dutilh J, Fulton JS, Jankovic L, Lopes NN *et al*: **A systematic review of salivary gland hypofunction and xerostomia induced by cancer therapies: prevalence, severity and impact on quality of life**. *Supportive care in cancer : official journal of the Multinational Association of Supportive Care in Cancer* 2010, **18**(8):1039-1060.
4. Bernstein JM, Bernstein CR, West CM, Homer JJ: **Molecular and cellular processes underlying the hallmarks of head and neck cancer**. *European archives of oto-rhino-laryngology : official journal of the European Federation of Oto-Rhino-Laryngological Societies* 2013, **270**(10):2585-2593.
5. Horsman MR, Mortensen LS, Petersen JB, Busk M, Overgaard J: **Imaging hypoxia to improve radiotherapy outcome**. *Nature reviews Clinical oncology* 2012, **9**(12):674-687.
6. Houweling AC, Schakel T, van den Berg CA, Philippons ME, Roesink JM, Terhaard CH, Raaijmakers CP: **MRI to quantify early radiation-induced changes in the salivary glands**. *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology* 2011, **100**(3):386-389.
7. Cheng CC, Chiu SC, Jen YM, Chang HC, Chung HW, Liu YJ, Chiu HC, Chen CY, Huang GS, Juan CJ: **Parotid perfusion in nasopharyngeal carcinoma patients in early-to-intermediate stage after low-dose intensity-modulated radiotherapy: evaluated by fat-saturated dynamic contrast-enhanced magnetic resonance imaging**. *Magnetic resonance imaging* 2013, **31**(8):1278-1284.
8. Juan CJ, Chen CY, Jen YM, Liu HS, Liu YJ, Hsueh CJ, Wang CY, Chou YC, Chai YT, Huang GS *et al*: **Perfusion characteristics of late radiation injury of parotid glands: quantitative evaluation with dynamic contrast-enhanced MRI**. *European radiology* 2009, **19**(1):94-102.
9. Bernstein JM, Homer JJ, West CM: **Dynamic contrast-enhanced magnetic resonance imaging biomarkers in head and neck cancer: potential to guide treatment? A systematic review**. *Oral oncology* 2014, **50**(10):963-970.
10. Noij DP, de Jong MC, Mulders LG, Marcus JT, de Bree R, Lavini C, de Graaf P, Castelijns JA: **Contrast-enhanced perfusion magnetic resonance imaging for head and neck squamous cell carcinoma: a systematic review**. *Oral oncology* 2015, **51**(2):124-138.
11. Lee FK, King AD, Kam MK, Ma BB, Yeung DK: **Radiation injury of the parotid glands during treatment for head and neck cancer: assessment using dynamic contrast-enhanced MR imaging**. *Radiation research* 2011, **175**(3):291-296.
12. van Luijk P, Pringle S, Deasy JO, Moiseenko VV, Faber H, Hovan A, Baanstra M, van der Laan HP, Kierkels RG, van der Schaaf A *et al*: **Sparing the region of the salivary gland containing stem cells preserves saliva production after radiotherapy for head and neck cancer**. *Science translational medicine* 2015, **7**(305):305ra147.
13. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, Zegers CM, Gillies R, Boellard R, Dekker A *et al*: **Radiomics: extracting more information from medical images using advanced feature analysis**. *European journal of cancer* 2012, **48**(4):441-446.
14. Fave X, Zhang L, Yang J, Mackin D, Balter P, Gomez D, Followill D, Jones AK, Stingo F, Liao Z *et al*: **Delta-radiomics features for the prediction of patient outcomes in non-small cell lung cancer**. *Scientific reports* 2017, **7**(1):588.
15. Fried DV, Mawlawi O, Zhang L, Fave X, Zhou S, Ibbott G, Liao Z, Court LE: **Stage III Non-Small Cell Lung Cancer: Prognostic Value of FDG PET Quantitative Imaging Features Combined with Clinical Prognostic Factors**. *Radiology* 2016, **278**(1):214-222.

16. Zhang H, Graham CM, Elci O, Griswold ME, Zhang X, Khan MA, Pitman K, Caudell JJ, Hamilton RD, Ganeshan B *et al*: **Locally advanced squamous cell carcinoma of the head and neck: CT texture and histogram analysis allow independent prediction of overall survival in patients treated with induction chemotherapy.** *Radiology* 2013, **269**(3):801-809.
17. Bogowicz M, Riesterer O, Ikenberg K, Stieb S, Moch H, Studer G, Guckenberger M, Tanadini-Lang S: **Computed Tomography Radiomics Predicts HPV Status and Local Tumor Control After Definitive Radiochemotherapy in Head and Neck Squamous Cell Carcinoma.** *Int J Radiat Oncol Biol Phys* 2017.
18. Bogowicz M, Riesterer O, Stark LS, Studer G, Unkelbach J, Guckenberger M, Tanadini-Lang S: **Comparison of PET and CT radiomics for prediction of local tumor control in head and neck squamous cell carcinoma.** *Acta oncologica* 2017:1-6.
19. Leijenaar RT, Carvalho S, Hoebbers FJ, Aerts HJ, van Elmpt WJ, Huang SH, Chan B, Waldron JN, O'Sullivan B, Lambin P: **External validation of a prognostic CT-based radiomic signature in oropharyngeal squamous cell carcinoma.** *Acta oncologica* 2015, **54**(9):1423-1429.
20. Parmar C, Leijenaar RT, Grossmann P, Rios Velazquez E, Bussink J, Rietveld D, Rietbergen MM, Haibe-Kains B, Lambin P, Aerts HJ: **Radiomic feature clusters and prognostic signatures specific for Lung and Head & Neck cancer.** *Scientific reports* 2015, **5**:11044.
21. Hunter LA, Chen YP, Zhang L, Matney JE, Choi H, Kry SF, Martel MK, Stingo F, Liao Z, Gomez D *et al*: **NSCLC tumor shrinkage prediction using quantitative image features.** *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society* 2016, **49**:29-36.
22. Fried DV, Tucker SL, Zhou S, Liao Z, Mawlawi O, Ibbott G, Court LE: **Prognostic value and reproducibility of pretreatment CT texture features in stage III non-small cell lung cancer.** *International journal of radiation oncology, biology, physics* 2014, **90**(4):834-842.
23. Thawani R, McLane M, Beig N, Ghose S, Prasanna P, Velcheti V, Madabhushi A: **Radiomics and radiogenomics in lung cancer: A review for the clinician.** *Lung Cancer* 2017.
24. Fried DV, Mawlawi O, Zhang L, Fave X, Zhou S, Ibbott G, Liao Z, Court LE: **Potential Use of (18)F-fluorodeoxyglucose Positron Emission Tomography-Based Quantitative Imaging Features for Guiding Dose Escalation in Stage III Non-Small Cell Lung Cancer.** *International journal of radiation oncology, biology, physics* 2016, **94**(2):368-376.
25. Cook GJ, Azad G, Owczarczyk K, Siddique M, Goh V: **Challenges and Promises of PET Radiomics.** *International Journal of Radiation Oncology* Biology* Physics* 2018.
26. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, Bussink J, Monshouwer R, Haibe-Kains B, Rietveld D *et al*: **Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach.** *Nat Commun* 2014, **5**:4006.
27. Parmar C, Grossmann P, Rietveld D, Rietbergen MM, Lambin P, Aerts HJ: **Radiomic machine-learning classifiers for prognostic biomarkers of head and neck cancer.** *Frontiers in oncology* 2015, **5**:272.
28. Ou D, Blanchard P, Rosellini S, Levy A, Nguyen F, Leijenaar RT, Garberis I, Gorphe P, Bidault F, Féré C: **Predictive and prognostic value of CT based radiomics signature in locally advanced head and neck cancers patients treated with concurrent chemoradiotherapy or bioradiotherapy and its added value to Human Papillomavirus status.** *Oral oncology* 2017, **71**:150-155.
29. Vallieres M, Kay-Rivest E, Perrin LJ, Liem X, Furstoss C, Aerts H, Khaouam N, Nguyen-Tan PF, Wang CS, Sultanem K *et al*: **Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer.** *Scientific reports* 2017, **7**(1):10117.
30. El Naqa I, Grigsby P, Apte A, Kidd E, Donnelly E, Khullar D, Chaudhari S, Yang D, Schmitt M, Laforest R: **Exploring feature-based approaches in PET images for predicting cancer treatment outcomes.** *Pattern recognition* 2009, **42**(6):1162-1171.

31. Folkert MR, Setton J, Apte AP, Grkovski M, Young RJ, Schöder H, Thorstad WL, Lee NY, Deasy JO, Oh JH: **Predictive modeling of outcomes following definitive chemoradiotherapy for oropharyngeal cancer based on FDG-PET image characteristics.** *Physics in Medicine & Biology* 2017, **62**(13):5327.
32. Shafiq-ul-Hassan M, Zhang GG, Latifi K, Ullah G, Hunt DC, Balagurunathan Y, Abdalah MA, Schabath MB, Goldgof DG, Mackin D: **Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels.** *Medical physics* 2017, **44**(3):1050-1062.
33. Lu L, Ehmke RC, Schwartz LH, Zhao B: **Assessing Agreement between Radiomic Features Computed for Multiple CT Imaging Settings.** *PloS one* 2016, **11**(12):e0166550.
34. Zhao B, Tan Y, Tsai WY, Qi J, Xie C, Lu L, Schwartz LH: **Reproducibility of radiomics for deciphering tumor phenotype with imaging.** *Scientific reports* 2016, **6**:23428.
35. Zhao B, Tan Y, Tsai WY, Schwartz LH, Lu L: **Exploring Variability in CT Characterization of Tumors: A Preliminary Phantom Study.** *Transl Oncol* 2014, **7**(1):88-93.
36. Larue R, van Timmeren JE, de Jong EEC, Feliciani G, Leijenaar RTH, Schreurs WMJ, Sosef MN, Raat F, van der Zande FHR, Das M *et al*: **Influence of gray level discretization on radiomic feature stability for different CT scanners, tube currents and slice thicknesses: a comprehensive phantom study.** *Acta oncologica* 2017:1-10.
37. Mackin D, Fave X, Zhang L, Yang J, Jones AK, Ng CS, Court L: **Harmonizing the pixel size in retrospective computed tomography radiomics studies.** *PLoS One* 2017, **12**(9):e0178524.
38. Mackin D, Ger R, Dodge C, Fave X, Chi P-C, Zhang L, Yang J, Bache S, Dodge C, Jones AK: **Effect of tube current on computed tomography radiomic features.** *Scientific reports* 2018, **8**(1):2354.
39. Bailly C, Bodet-Milin C, Couespel S, Necib H, Kraeber-Bodéré F, Ansquer C, Carlier T: **Revisiting the robustness of PET-based textural features in the context of multi-centric trials.** *PloS one* 2016, **11**(7):e0159984.
40. Doumou G, Siddique M, Tsoumpas C, Goh V, Cook GJ: **The precision of textural analysis in 18F-FDG-PET scans of oesophageal cancer.** *European radiology* 2015, **25**(9):2805-2812.
41. Forgacs A, Jonsson HP, Dahlbom M, Daver F, DiFranco MD, Opposits G, Krizsan AK, Garai I, Czernin J, Varga J: **A study on the basic criteria for selecting heterogeneity parameters of F18-FDG PET images.** *PloS one* 2016, **11**(10):e0164113.
42. Galavis PE, Hollensen C, Jallow N, Paliwal B, Jeraj R: **Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters.** *Acta oncologica* 2010, **49**(7):1012-1016.
43. Lasnon C, Majdoub M, Lavigne B, Do P, Madelaine J, Visvikis D, Hatt M, Aide N: **18 F-FDG PET/CT heterogeneity quantification through textural features in the era of harmonisation programs: a focus on lung cancer.** *European journal of nuclear medicine and molecular imaging* 2016, **43**(13):2324-2335.
44. Orlhac F, Nioche C, Soussan M, Buvat I: **Understanding changes in tumor textural indices in PET: a comparison between visual assessment and index values in simulated and patient data.** *Journal of nuclear medicine : official publication, Society of Nuclear Medicine* 2017, **58**(3):387-392.
45. Orlhac F, Thézé B, Soussan M, Boisgard R, Buvat I: **Multiscale texture analysis: from 18F-FDG PET images to histologic images.** *Journal of nuclear medicine : official publication, Society of Nuclear Medicine* 2016, **57**:1823-1828.
46. Shiri I, Rahmim A, Ghaffarian P, Geramifar P, Abdollahi H, Bitarafan-Rajabi A: **The impact of image reconstruction settings on 18F-FDG PET radiomic features: multi-scanner phantom and patient studies.** *European radiology* 2017, **27**(11):4498-4509.
47. van Velden FH, Kramer GM, Frings V, Nissen IA, Mulder ER, de Langen AJ, Hoekstra OS, Smit EF, Boellaard R: **Repeatability of radiomic features in non-small-cell lung cancer [18F] FDG-PET/CT**

- studies: impact of reconstruction and delineation.** *Molecular imaging and biology* 2016, **18**(5):788-795.
48. Yan J, Chu-Shern JL, Loi HY, Khor LK, Sinha AK, Quek ST, Tham IW, Townsend D: **Impact of image reconstruction settings on texture features in 18F-FDG PET.** *Journal of nuclear medicine* 2015, **56**(11):1667-1673.
 49. Nyflot MJ, Yang F, Byrd D, Bowen SR, Sandison GA, Kinahan PE: **Quantitative radiomics: impact of stochastic effects on textural feature analysis implies the need for standards.** *Journal of Medical Imaging* 2015, **2**(4):041002.
 50. Reuzé S, Orlhac F, Chargari C, Nioche C, Limkin E, Riet F, Escande A, Haie-Meder C, Dercle L, Gouy S: **Prediction of cervical cancer recurrence using textural features extracted from 18F-FDG PET images acquired with different scanners.** *Oncotarget* 2017, **8**(26):43169.
 51. Shan ZY, Kwiatek R, Burnet R, Del Fante P, Staines DR, Marshall-Gradisnik SM, Barnden LR: **Progressive brain changes in patients with chronic fatigue syndrome: A longitudinal MRI study.** *Journal of magnetic resonance imaging : JMIR* 2016, **44**(5):1301-1311.
 52. Chou PC, Shunmugavel A, El Sayed H, Desouki MM, Nguyen SA, Khan M, Singh I, Bilgen M: **Preclinical use of longitudinal MRI for screening the efficacy of S-nitrosoglutathione in treating spinal cord injury.** *Journal of magnetic resonance imaging : JMIR* 2011, **33**(6):1301-1311.
 53. Sharma U, Danishad KK, Seenu V, Jagannathan NR: **Longitudinal study of the assessment by MRI and diffusion-weighted imaging of tumor response in patients with locally advanced breast cancer undergoing neoadjuvant chemotherapy.** *NMR in biomedicine* 2009, **22**(1):104-113.
 54. Caruana EJ, Roman M, Hernandez-Sanchez J, Solli P: **Longitudinal studies.** *Journal of thoracic disease* 2015, **7**(11):E537-540.
 55. Cooperative JHaNR-MD: **Dynamic contrast-enhanced MRI detects acute radiotherapy-induced alterations in mandibular microvasculature: prospective assessment of imaging biomarkers of normal tissue injury.** *Scientific reports* 2016, **6**:29864.
 56. Mohamed AS, Ruangskul MN, Awan MJ, Baron CA, Kalpathy-Cramer J, Castillo R, Castillo E, Guerrero TM, Kocak-Uzel E, Yang J *et al*: **Quality assurance assessment of diagnostic and radiation therapy-simulation CT image registration for head and neck radiation therapy: anatomic region of interest-based comparison of rigid and deformable algorithms.** *Radiology* 2015, **274**(3):752-763.
 57. Kirby N, Chuang C, Ueda U, Pouliot J: **The need for application-based adaptation of deformable image registration.** *Medical physics* 2013, **40**(1):011702.
 58. Wang H, Dong L, Lii MF, Lee AL, de Crevoisier R, Mohan R, Cox JD, Kuban DA, Cheung R: **Implementation and validation of a three-dimensional deformable registration algorithm for targeted prostate cancer radiotherapy.** *International journal of radiation oncology, biology, physics* 2005, **61**(3):725-735.
 59. Wang H, Dong L, O'Daniel J, Mohan R, Garden AS, Ang KK, Kuban DA, Bonnen M, Chang JY, Cheung R: **Validation of an accelerated 'demons' algorithm for deformable image registration in radiation therapy.** *Physics in medicine and biology* 2005, **50**(12):2887-2905.
 60. Rigaud B, Simon A, Castelli J, Gobeli M, Ospina Arango JD, Cazoulat G, Henry O, Haigron P, De Crevoisier R: **Evaluation of deformable image registration methods for dose monitoring in head and neck radiotherapy.** *BioMed research international* 2015, **2015**:726268.
 61. Castillo R, Castillo E, Guerra R, Johnson VE, McPhail T, Garg AK, Guerrero T: **A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets.** *Physics in medicine and biology* 2009, **54**(7):1849-1870.
 62. Brock KK, Sharpe MB, Dawson LA, Kim SM, Jaffray DA: **Accuracy of finite element model-based multi-organ deformable image registration.** *Medical physics* 2005, **32**(6):1647-1659.

63. Kaus MR, Brock KK, Pekar V, Dawson LA, Nichol AM, Jaffray DA: **Assessment of a model-based deformable image registration approach for radiation therapy planning.** *International journal of radiation oncology, biology, physics* 2007, **68**(2):572-580.
64. Brock KK, Nichol AM, Menard C, Moseley JL, Warde PR, Catton CN, Jaffray DA: **Accuracy and sensitivity of finite element model-based deformable registration of the prostate.** *Medical physics* 2008, **35**(9):4019-4025.
65. Nichol AM, Brock KK, Lockwood GA, Moseley DJ, Rosewall T, Warde PR, Catton CN, Jaffray DA: **A magnetic resonance imaging study of prostate deformation relative to implanted gold fiducial markers.** *International journal of radiation oncology, biology, physics* 2007, **67**(1):48-56.
66. Lian J, Xing L, Hunjan S, Dumoulin C, Levin J, Lo A, Watkins R, Rohling K, Giaquinto R, Kim D *et al*: **Mapping of the prostate in endorectal coil-based MRI/MRSI and CT: a deformable registration and validation study.** *Medical physics* 2004, **31**(11):3087-3094.
67. Bruckner T, Lucht R, Brix G: **Comparison of rigid and elastic matching of dynamic magnetic resonance mammographic images by mutual information.** *Medical physics* 2000, **27**(10):2456-2461.
68. Schnabel JA, Tanner C, Castellano-Smith AD, Degenhard A, Leach MO, Hose DR, Hill DL, Hawkes DJ: **Validation of nonrigid image registration using finite-element methods: application to breast MR images.** *IEEE transactions on medical imaging* 2003, **22**(2):238-247.
69. Palma DA, van Sornsen de Koste JR, Verbakel WF, Senan S: **A new approach to quantifying lung damage after stereotactic body radiation therapy.** *Acta oncologica* 2011, **50**(4):509-517.
70. Singhrao K, Kirby N, Pouliot J: **A three-dimensional head-and-neck phantom for validation of multimodality deformable image registration for adaptive radiotherapy.** *Medical physics* 2014, **41**(12):121709.
71. Lawson JD, Schreiber E, Jani AB, Fox T: **Quantitative evaluation of a cone-beam computed tomography-planning computed tomography deformable image registration method for adaptive radiation therapy.** *Journal of applied clinical medical physics / American College of Medical Physics* 2007, **8**(4):2432.
72. Lin H, Ayan A, Zhai H, Zhu T, Both S: **SU-GG-I-109: A Quantitative Evaluation of Velocity AI Deformable Image Registration.** *Medical physics* 2010, **37**(6):3126-3126.
73. Stanley N, Zhong H, Glide-Hurst C, Chetty I, Movsas B: **MO-F-BRA-06: Systematic Evaluation of a Deformable Image Registration Algorithm from a Commercial Software Package.** *Medical physics* 2012, **39**(6):3876-3876.
74. Zhang Y, Yang J, Zhang L, Court LE, Balter PA, Dong L: **Modeling respiratory motion for reducing motion artifacts in 4D CT images.** *Medical physics* 2013, **40**(4):041716.
75. Yu ZH, Kudchadker R, Dong L, Zhang Y, Court LE, Mourtada F, Yock A, Tucker SL, Yang J: **Learning anatomy changes from patient populations to create artificial CT images for voxel-level validation of deformable image registration.** *Journal of applied clinical medical physics / American College of Medical Physics* 2016, **17**(1):5888.
76. **Magnetic resonance imaging (MRI) exams** [<https://data.oecd.org/healthcare/magnetic-resonance-imaging-mri-exams.htm>]
77. Amini A, Yang J, Williamson R, McBurney ML, Erasmus J, Jr., Allen PK, Karhade M, Komaki R, Liao Z, Gomez D *et al*: **Dose constraints to prevent radiation-induced brachial plexopathy in patients treated for lung cancer.** *International journal of radiation oncology, biology, physics* 2012, **82**(3):e391-398.
78. Yang J, Zhang Y, Zhang L, Dong L: **Automatic segmentation of parotids from CT scans using multiple atlases.** *Medical Image Analysis for the Clinic: A Grand Challenge* 2010:323-330.
79. Fontanilla HP, Klopp AH, Lindberg ME, Jhingran A, Kelly P, Takiar V, Iyer RB, Levenback CF, Zhang Y, Dong L *et al*: **Anatomic distribution of [(18)F] fluorodeoxyglucose-avid lymph nodes in patients with cervical cancer.** *Practical radiation oncology* 2013, **3**(1):45-53.

80. Ding Y, Mohamed AS, Yang J, Colen RR, Frank SJ, Wang J, Wassal EY, Wang W, Kantor ME, Balter PA *et al*: **Prospective observer and software-based assessment of magnetic resonance imaging quality in head and neck cancer: Should standard positioning and immobilization be required for radiation therapy applications?** *Practical radiation oncology* 2015, **5**(4):e299-308.
81. Huang W, Chen Y, Fedorov A, Li X, Jajamovich GH, Malyarenko DI, Aryal MP, LaViolette PS, Oborski MJ, O'Sullivan F *et al*: **The Impact of Arterial Input Function Determination Variations on Prostate Dynamic Contrast-Enhanced Magnetic Resonance Imaging Pharmacokinetic Modeling: A Multicenter Data Analysis Challenge.** *Tomography : a journal for imaging research* 2016, **2**(1):56-66.
82. Yankeelov TE, Gore JC: **Dynamic Contrast Enhanced Magnetic Resonance Imaging in Oncology: Theory, Data Acquisition, Analysis, and Examples.** *Current medical imaging reviews* 2009, **3**(2):91-107.
83. Yankeelov TE, Rooney WD, Li X, Springer CS, Jr.: **Variation of the relaxographic "shutter-speed" for transcytolemmal water exchange affects the CR bolus-tracking curve shape.** *Magnetic resonance in medicine* 2003, **50**(6):1151-1169.
84. Leach M, Morgan B, Tofts P, Buckley D, Huang W, Horsfield M, Chenevert T, Collins D, Jackson A, Lomas D: **Imaging vascular function for early stage clinical trials using dynamic contrast-enhanced magnetic resonance imaging.** *European radiology* 2012, **22**(7):1451-1464.
85. Schabel MC, Parker DL: **Uncertainty and bias in contrast concentration measurements using spoiled gradient echo pulse sequences.** *Physics in medicine and biology* 2008, **53**(9):2345-2373.
86. Yang C, Karczmar GS, Medved M, Oto A, Zamora M, Stadler WM: **Reproducibility assessment of a multiple reference tissue method for quantitative dynamic contrast enhanced-MRI analysis.** *Magnetic resonance in medicine* 2009, **61**(4):851-859.
87. Heisen M, Fan X, Buurman J, van Riel NA, Karczmar GS, ter Haar Romeny BM: **The influence of temporal resolution in determining pharmacokinetic parameters from DCE-MRI data.** *Magnetic resonance in medicine* 2010, **63**(3):811-816.
88. Di Giovanni P, Azlan CA, Ahearn TS, Semple SI, Gilbert FJ, Redpath TW: **The accuracy of pharmacokinetic parameter measurement in DCE-MRI of the breast at 3 T.** *Physics in medicine and biology* 2010, **55**(1):121-132.
89. Sourbron SP, Buckley DL: **On the scope and interpretation of the Tofts models for DCE-MRI.** *Magnetic resonance in medicine* 2011, **66**(3):735-745.
90. Sourbron SP, Buckley DL: **Tracer kinetic modelling in MRI: estimating perfusion and capillary permeability.** *Physics in medicine and biology* 2012, **57**(2):R1-33.
91. Othman AE, Falkner F, Kessler DE, Martirosian P, Weiss J, Kruck S, Kaufmann S, Grimm R, Kramer U, Nikolaou K *et al*: **Comparison of different population-averaged arterial-input-functions in dynamic contrast-enhanced MRI of the prostate: Effects on pharmacokinetic parameters and their diagnostic performance.** *Magnetic resonance imaging* 2016, **34**(4):496-501.
92. Tofts PS, Kermode AG: **Measurement of the blood-brain barrier permeability and leakage space using dynamic MR imaging. 1. Fundamental concepts.** *Magnetic resonance in medicine* 1991, **17**(2):357-367.
93. Heye T, Davenport MS, Horvath JJ, Feuerlein S, Breault SR, Bashir MR, Merkle EM, Boll DT: **Reproducibility of dynamic contrast-enhanced MR imaging. Part I. Perfusion characteristics in the female pelvis by using multiple computer-aided diagnosis perfusion analysis solutions.** *Radiology* 2013, **266**(3):801-811.
94. Huang W, Li X, Chen Y, Li X, Chang MC, Oborski MJ, Malyarenko DI, Muzi M, Jajamovich GH, Fedorov A *et al*: **Variations of dynamic contrast-enhanced magnetic resonance imaging in evaluation of breast cancer therapy response: a multicenter data analysis challenge.** *Translational oncology* 2014, **7**(1):153-166.

95. Beuzit L, Eliat PA, Brun V, Ferre JC, Gandon Y, Bannier E, Saint-Jalmes H: **Dynamic contrast-enhanced MRI: Study of inter-software accuracy and reproducibility using simulated and clinical data.** *Journal of magnetic resonance imaging : JMRI* 2016, **43**(6):1288-1300.
96. Cron GO, Sourbron S, Barnoriak D, Abdeen R, Hogan M, Nguyen TB: **Bias and precision of three different DCE-MRI analysis software packages: a comparison using simulated data.** In: *International Society for Magnetic Resonance in Medicine: 2014: Proc 22nd Annual Meeting ISMRM, Milan (abstract 4592); 2014.*
97. Tofts PS: **Modeling tracer kinetics in dynamic Gd-DTPA MR imaging.** *Journal of magnetic resonance imaging : JMRI* 1997, **7**(1):91-101.
98. Tofts PS, Brix G, Buckley DL, Evelhoch JL, Henderson E, Knopp MV, Larsson HB, Lee TY, Mayr NA, Parker GJ et al: **Estimating kinetic parameters from dynamic contrast-enhanced T(1)-weighted MRI of a diffusable tracer: standardized quantities and symbols.** *Journal of magnetic resonance imaging : JMRI* 1999, **10**(3):223-232.
99. Korporaal JG, van den Berg CA, van Osch MJ, Groenendaal G, van Vulpen M, van der Heide UA: **Phase-based arterial input function measurements in the femoral arteries for quantification of dynamic contrast-enhanced (DCE) MRI and comparison with DCE-CT.** *Magnetic resonance in medicine* 2011, **66**(5):1267-1274.
100. [<http://www.nordicneurolab.com>]
101. Li X, Huang W, Morris EA, Tudorica LA, Seshan VE, Rooney WD, Tagge I, Wang Y, Xu J, Springer CS, Jr.: **Dynamic NMR effects in breast cancer dynamic-contrast-enhanced MRI.** *Proceedings of the National Academy of Sciences of the United States of America* 2008, **105**(46):17937-17942.
102. Tudorica A, Oh KY, Chui SY, Roy N, Troxell ML, Naik A, Kemmer KA, Chen Y, Holtorf ML, Afzal A et al: **Early Prediction and Evaluation of Breast Cancer Response to Neoadjuvant Chemotherapy Using Quantitative DCE-MRI.** *Translational oncology* 2016, **9**(1):8-17.
103. Coolens C, Driscoll B, Chung C, Shek T, Gorjizadeh A, Menard C, Jaffray D: **Automated voxel-based analysis of volumetric dynamic contrast-enhanced CT data improves measurement of serial changes in tumor vascular biomarkers.** *International journal of radiation oncology, biology, physics* 2015, **91**(1):48-57.
104. Coolens C, Driscoll B, Moseley J, Brock KK, Dawson LA: **Feasibility of 4D perfusion CT imaging for the assessment of liver treatment response following SBRT and sorafenib.** *Advances in Radiation Oncology* 2016, **1**(3):194-203.
105. Hormuth DA, 2nd, Skinner JT, Does MD, Yankeelov TE: **A comparison of individual and population-derived vascular input functions for quantitative DCE-MRI in rats.** *Magnetic resonance imaging* 2014, **32**(4):397-401.
106. Barnes SL, Whisenant JG, Loveless ME, Yankeelov TE: **Practical dynamic contrast enhanced MRI in small animal models of cancer: data acquisition, data analysis, and interpretation.** *Pharmaceutics* 2012, **4**(3):442-478.
107. Butterworth E, Jardine BE, Raymond GM, Neal ML, Bassingthwaighe JB: **JSim, an open-source modeling system for data analysis.** *F1000Research* 2013, **2**:288.
108. Daniel P. Barboriak Lab [<https://sites.duke.edu/dblab/qibacontent/>]
109. Daniel P. Barboriak Lab [<https://sites.duke.edu/dblab/qibacontent/>]
110. Carletta J: **Assessing agreement on classification tasks: the kappa statistic.** *Computational linguistics* 1996, **22**(2):249-254.
111. Krippendorff K: **Reliability in content analysis.** *Human communication research* 2004, **30**(3):411-433.
112. Neuendorf KA: **The content analysis guidebook:** Sage; 2002.
113. Kim S, Loevner LA, Quon H, Kilger A, Sherman E, Weinstein G, Chalian A, Poptani H: **Prediction of response to chemoradiation therapy in squamous cell carcinomas of the head and neck using dynamic contrast-enhanced MR imaging.** *AJNR American journal of neuroradiology* 2010, **31**(2):262-268.

114. Van Cann EM, Rijpkema M, Heerschap A, van der Bilt A, Koole R, Stoelinga PJ: **Quantitative dynamic contrast-enhanced MRI for the assessment of mandibular invasion by squamous cell carcinoma.** *Oral oncology* 2008, **44**(12):1147-1154.
115. Lee FK, King AD, Ma BB, Yeung DK: **Dynamic contrast enhancement magnetic resonance imaging (DCE-MRI) for differential diagnosis in head and neck cancers.** *European journal of radiology* 2012, **81**(4):784-788.
116. Bisdas S, Seitz O, Middendorp M, Chambron-Pinho N, Bisdas T, Vogl TJ, Hammerstingl R, Ernemann U, Mack MG: **An exploratory pilot study into the association between microcirculatory parameters derived by MRI-based pharmacokinetic analysis and glucose utilization estimated by PET-CT imaging in head and neck cancer.** *European radiology* 2010, **20**(10):2358-2366.
117. Tofts PS, Berkowitz B, Schnall MD: **Quantitative analysis of dynamic Gd-DTPA enhancement in breast tumors using a permeability model.** *Magnetic resonance in medicine* 1995, **33**(4):564-568.
118. Ashton E: **Quantitative MR in multi-center clinical trials.** *Journal of magnetic resonance imaging : JMRI* 2010, **31**(2):279-288.
119. Schabel MC, Morrell GR: **Uncertainty in T(1) mapping using the variable flip angle method with two flip angles.** *Physics in medicine and biology* 2009, **54**(1):N1-8.
120. Eklund A, Nichols TE, Knutsson H: **Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates.** *Proceedings of the National Academy of Sciences of the United States of America* 2016, **113**(28):7900-7905.
121. Head J, Cooperative NR-MD, Elhalawani H, Ger RB, Mohamed AS, Awan MJ, Ding Y, Li K, Fave XJ, Beers AL: **Dynamic contrast-enhanced magnetic resonance imaging for head and neck cancers.** *Scientific data* 2018, **5**:180008.
122. Galbraith SM, Lodge MA, Taylor NJ, Rustin GJ, Bentzen S, Stirling JJ, Padhani AR: **Reproducibility of dynamic contrast-enhanced MRI in human muscle and tumours: comparison of quantitative and semi-quantitative analysis.** *NMR in biomedicine* 2002, **15**(2):132-142.
123. Fave X, Cook M, Frederick A, Zhang L, Yang J, Fried D, Stingo F, Court L: **Preliminary investigation into sources of uncertainty in quantitative imaging features.** *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society* 2015, **44**:54-61.
124. Berenguer R, Pastor-Juan MDR, Canales-Vazquez J, Castro-Garcia M, Villas MV, Legorburo FM, Sabater S: **Radiomics of CT Features May Be Nonreproducible and Redundant: Influence of CT Acquisition Parameters.** *Radiology* 2018:172361.
125. McCollough CH, Primak AN, Braun N, Kofler J, Yu L, Christner J: **Strategies for reducing radiation dose in CT.** *Radiologic clinics of North America* 2009, **47**(1):27-40.
126. Kambadakone AR, Chaudhary NA, Desai GS, Nguyen DD, Kulkarni NM, Sahani DV: **Low-dose MDCT and CT enterography of patients with Crohn disease: feasibility of adaptive statistical iterative reconstruction.** *AJR American journal of roentgenology* 2011, **196**(6):W743-752.
127. Singh S, Kalra MK, Moore MA, Shailam R, Liu B, Toth TL, Grant E, Westra SJ: **Dose reduction and compliance with pediatric CT protocols adapted to patient size, clinical indication, and number of prior studies.** *Radiology* 2009, **252**(1):200-208.
128. Yu L, Bruesewitz MR, Thomas KB, Fletcher JG, Kofler JM, McCollough CH: **Optimal tube potential for radiation dose reduction in pediatric CT: principles, clinical implementations, and pitfalls.** *Radiographics : a review publication of the Radiological Society of North America, Inc* 2011, **31**(3):835-848.
129. Yu L, Li H, Fletcher JG, McCollough CH: **Automatic selection of tube potential for radiation dose reduction in CT: a general strategy.** *Medical physics* 2010, **37**(1):234-243.
130. Marin D, Nelson RC, Schindera ST, Richard S, Youngblood RS, Yoshizumi TT, Samei E: **Low-tube-voltage, high-tube-current multidetector abdominal CT: improved image quality and**

- decreased radiation dose with adaptive statistical iterative reconstruction algorithm--initial clinical experience.** *Radiology* 2010, **254**(1):145-153.
131. Lee KH, Lee JM, Moon SK, Baek JH, Park JH, Flohr TG, Kim KW, Kim SJ, Han JK, Choi BI: **Attenuation-based automatic tube voltage selection and tube current modulation for dose reduction at contrast-enhanced liver CT.** *Radiology* 2012, **265**(2):437-447.
 132. Mackin D, Fave X, Zhang L, Fried D, Yang J, Taylor B, Rodriguez-Rivera E, Dodge C, Jones AK, Court L: **Measuring Computed Tomography Scanner Variability of Radiomics Features.** *Investigative radiology* 2015, **50**(11):757-765.
 133. Johns HE: **Physics of radiology**; Charles River Media; 1983.
 134. Chang KP, Hung SH, Chie YH, Shiau AC, Huang RJ: **A comparison of physical and dosimetric properties of lung substitute materials.** *Medical physics* 2012, **39**(4):2013-2020.
 135. Marashdeh MW, Al-Hamarneh IF, Munem EMA, Tajuddin A, Ariffin A, Al-Omari S: **Determining the mass attenuation coefficient, effective atomic number, and electron density of raw wood and binderless particleboards of Rhizophora spp. by using Monte Carlo simulation.** *Results in Physics* 2015, **5**:228-234.
 136. Ger RB, Zhou S, Chi P-CM, Lee HJ, Layman RR, Jones AK, Goff DL, Fuller CD, Howell RM, Li H *et al*: **Comprehensive Investigation on Controlling for CT Imaging Variabilities in Radiomics Studies.** *Scientific reports* 2018, **8**(1):13047.
 137. Zhang L, Fried DV, Fave XJ, Hunter LA, Yang J, Court LE: **IBEX: an open infrastructure software platform to facilitate collaborative work in radiomics.** *Medical physics* 2015, **42**(3):1341-1353.
 138. Ger RB, Cardenas CE, Anderson BM, Yang J, Mackin DS, Zhang L: **Guidelines and Experience Using Imaging Biomarker Explorer (IBEX) for Radiomics.** *Journal of visualized experiments: JoVE* 2018(131).
 139. Haralick RM, Shanmugam K: **Textural features for image classification.** *IEEE Transactions on systems, man, and cybernetics* 1973(6):610-621.
 140. Galloway MM: **Texture analysis using gray level run lengths.** *Computer Graphics and Image Processing* 1975, **4**(2):172-179.
 141. Tang X: **Texture information in run-length matrices.** *IEEE transactions on image processing* 1998, **7**(11):1602-1609.
 142. Amadasun M, King R: **Textural features corresponding to textural properties.** *IEEE Transactions on systems, man, and Cybernetics* 1989, **19**(5):1264-1274.
 143. Fave X, Zhang L, Yang J, Mackin D, Balter P, Gomez D, Followill D, Jones AK, Stingo F: **Impact of image preprocessing on the volume dependence and prognostic potential of radiomics features in non-small cell lung cancer.** *Translational Cancer Research* 2016, **5**(4):349-363.
 144. Noid G, Tai A, Liu Y, Li A: **Enhancement of Early Radiation Treatment Response Assessment by Mono-energetic Decomposition of Dual-Energy Computed Tomography.** *International Journal of Radiation Oncology• Biology• Physics* 2016, **96**(2):S192.
 145. Goodsitt MM, Christodoulou EG, Larson SC: **Accuracies of the synthesized monochromatic CT numbers and effective atomic numbers obtained with a rapid kVp switching dual energy CT scanner.** *Medical physics* 2011, **38**(4):2222-2232.
 146. Yamashita Y, Kimura M, Kitahara M, Hamaguchi T, Kanno I, Ohtaka M, Hashimoto M, Ara K, Onabe H: **Measurement of effective atomic numbers using energy-resolved computed tomography.** *Journal of Nuclear Science and Technology* 2014, **51**(10):1256-1263.
 147. Saion E, Sulaiman ZA, Ahmad A, Wagiran H: **Determination of effective atomic number of rubber.** *Pertanika* 1983, **6**(3):95-98.
 148. Buch K, Li B, Qureshi M, Kuno H, Anderson S, Sakai O: **Quantitative assessment of variation in CT parameters on texture features: pilot study using a nonanatomic phantom.** *American Journal of Neuroradiology* 2017.
 149. Evans A, Evans R: **The composition of a tyre: typical components.** *The Waste & Resources Action Programme* 2006, **5**.

150. Fave X, Cook M, Frederick A, Zhang L, Yang J, Fried D, Stingo F, Court L: **Preliminary investigation into sources of uncertainty in quantitative imaging features.** *Computerized Medical Imaging and Graphics* 2015, **44**:54-61.
151. Ger R, Cardenas C, Anderson B, Yang J, Mackin D, Zhang L: **Guidelines and Experience Using Imaging Biomarker Explorer (IBEX) for Radiomics.** *Journal of visualized experiments: JoVE* 2018(131).
152. Grant RL, Summers PA, Neihart JL, Blatnica AP, Sahoo N, Gillin MT, Followill DS, Ibbott GS: **Relative stopping power measurements to aid in the design of anthropomorphic phantoms for proton radiotherapy.** *Journal of applied clinical medical physics* 2014, **15**(2):121-126.
153. Craft DF, Howell RM: **Preparation and fabrication of a full-scale, sagittal-sliced, 3D-printed, patient-specific radiotherapy phantom.** *Journal of applied clinical medical physics* 2017, **18**(5):285-292.
154. Craft DF, Kry SF, Balter P, Salehpour M, Woodward W, Howell RM: **Material matters: Analysis of density uncertainty in 3D printing and its consequences for radiation oncology.** *Medical physics* 2018, **45**(4):1614-1621.
155. Measurements ICoRUa: **Phantoms And Computational Models In Therapy, Diagnosis, And Protection.** In. Bethesda, Md., U.S.A.; 1992.
156. Ger RB, Craft DF, Mackin DS, Zhou S, Layman RR, Jones AK, Elhalawani H, Fuller CD, Howell RM, Li H: **Practical guidelines for handling head and neck computed tomography artifacts for quantitative image analysis.** *Computerized Medical Imaging and Graphics* 2018, **69**:134-139.
157. Larue RT, Defraene G, De Ruysscher D, Lambin P, van Elmpt W: **Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures.** *The British journal of radiology* 2017, **90**(1070):20160665.
158. Mackin D, Ger R, Zhang L, Yang J, Chi P, Bache S, Dodge C, Jones AK, Court L: **Homogenizing Reconstruction Kernels for CT Radiomics.** *Medical physics* 2018, **AAPM Anual Meeting TU-GH-KDBRC-2.**
159. Shrout PE, Fleiss JL: **Intraclass correlations: uses in assessing rater reliability.** *Psychological bulletin* 1979, **86**(2):420-428.
160. Revelle W, Revelle MW: **Package 'psych'.** In.; 2017.
161. Koo TK, Li MY: **A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research.** *Journal of chiropractic medicine* 2016, **15**(2):155-163.
162. Solomon JB, Christianson O, Samei E: **Quantitative comparison of noise texture across CT scanners from different manufacturers.** *Medical physics* 2012, **39**(10):6048-6055.
163. Winslow J, Zhang Y, Samei E: **A method for characterizing and matching CT image quality across CT scanners from different manufacturers.** *Medical physics* 2017, **44**(11):5705-5717.
164. Shafiq-ul-Hassan M, Zhang GG, Hunt DC, Latifi K, Ullah G, Gillies RJ, Moros EG: **Accounting for reconstruction kernel-induced variability in CT radiomic features using noise power spectra.** *Journal of Medical Imaging* 2017, **5**(1):011013.
165. He L, Huang Y, Ma Z, Liang C, Liang C, Liu Z: **Effects of contrast-enhancement, reconstruction slice thickness and convolution kernel on the diagnostic performance of radiomics signature in solitary pulmonary nodule.** *Scientific reports* 2016, **6**:34921.
166. Leijenaar RT, Nalbantov G, Carvalho S, Van Elmpt WJ, Troost EG, Boellaard R, Aerts HJ, Gillies RJ, Lambin P: **The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis.** *Scientific reports* 2015, **5**:11075.
167. Hatt M, Majdoub M, Vallieres M, Tixier F, Le Rest CC, Groheux D, Hindie E, Martineau A, Pradier O, Hustinx R *et al*: **¹⁸F-FDG PET uptake characterization through texture analysis: investigating the complementary nature of heterogeneity and functional tumor volume in a multi-cancer site patient cohort.** *Journal of nuclear medicine : official publication, Society of Nuclear Medicine* 2015, **56**(1):38-44.

168. Orlhac F, Boughdad S, Philippe C, Stalla-Bourdillon H, Nioche C, Champion L, Soussan M, Frouin F, Frouin V, Buvat I: **A postreconstruction harmonization method for multicenter radiomic studies in PET.** *Journal of Nuclear Medicine* 2018, **59**(8):1321-1328.
169. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Phillips S, Maffitt D, Pringle M: **The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository.** *Journal of digital imaging* 2013, **26**(6):1045-1057.
170. Vallieres M, Kay-Rivest E, Perrin L, Liem X, Furstoss C, Khaouam N, Nguyen-Tan P, Wang C, Sultanem K: **Data from Head-Neck-PET-CT.** In.: The Cancer Imaging Archive; 2017.
171. Feliciani G, Fioroni F, Grassi E, Bertolini M, Rosca A, Timon G, Galaverni M, Iotti C, Versari A, Iori M: **Radiomic Profiling of Head and Neck Cancer: 18F-FDG PET Texture Analysis as Predictor of Patient Survival.** *Contrast media & molecular imaging* 2018, **2018**.
172. Kuno H, Qureshi M, Chapman M, Li B, Andreu-Arasa V, Onoue K, Truong M, Sakai O: **CT texture analysis potentially predicts local failure in head and neck squamous cell carcinoma treated with chemoradiotherapy.** *American Journal of Neuroradiology* 2017, **38**(12):2334-2340.
173. Foy JJ, Mitta P, Nowosatka LR, Mendel KR, Li H, Giger ML, Al-Hallaq H, Armato SG: **Variations in algorithm implementation among quantitative texture analysis software packages.** In: *Medical Imaging 2018: Computer-Aided Diagnosis: 2018*: International Society for Optics and Photonics; 2018: 105751K.
174. Kann BH, Aneja S, Loganadane GV, Kelly JR, Smith SM, Decker RH, James BY, Park HS, Yarbrough WG, Malhotra A: **Pretreatment Identification of Head and Neck Cancer Nodal Metastasis and Extranodal Extension Using Deep Learning Neural Networks.** *Scientific reports* 2018, **8**(1):14036.
175. Vial A, Stirling D, Field M, Ros M, Ritz C, Carolan M, Holloway L, Miller AA: **The role of deep learning and radiomic feature extraction in cancer-specific predictive modelling: a review.** *Translational Cancer Research* 2018, **7**(3):803-816.
176. Jackson A, Li K-L, Zhu X: **Semi-quantitative parameter analysis of DCE-MRI revisited: monte-carlo simulation, clinical comparisons, and clinical validation of measurement errors in patients with type 2 neurofibromatosis.** *PloS one* 2014, **9**(3):e90300.
177. Rouvière O, Raudrant A, Ecochard R, Colin-Pangaud C, Pasquiou C, Bouvier R, Maréchal J, Lyonnet D: **Characterization of time-enhancement curves of benign and malignant prostate tissue at dynamic MR imaging.** *European radiology* 2003, **13**(5):931-942.
178. Lavini C, Verhoeff JJ, Majoie CB, Stalpers LJ, Richel DJ, Maas M: **Model-based, semiquantitative and time intensity curve shape analysis of dynamic contrast-enhanced MRI: A comparison in patients undergoing antiangiogenic treatment for recurrent glioma.** *Journal of Magnetic Resonance Imaging* 2011, **34**(6):1303-1312.
179. Narang J, Jain R, Arbab AS, Mikkelsen T, Scarpace L, Rosenblum ML, Hearshen D, Babajani-Feremi A: **Differentiating treatment-induced necrosis from recurrent/progressive brain tumor using nonmodel-based semiquantitative indices derived from dynamic contrast-enhanced T1-weighted MR perfusion.** *Neuro-oncology* 2011, **13**(9):1037-1046.
180. Vigneswaran N, Williams MD: **Epidemiologic trends in head and neck cancer and aids in diagnosis.** *Oral and Maxillofacial Surgery Clinics* 2014, **26**(2):123-141.
181. ESMO: **Epidemiology, risk factors and pathogenesis of squamous cell tumours.** In: *2017 ESMO Essentials for Clinicians Head and Neck Cancers Chapter 1.* 2017.
182. Guizard A-V, Uhry Z, de Raucourt D, Mazzoleni G, Sánchez M-J, Ligier K, group GE-w: **Trends in net survival from head and neck cancer in six European Latin countries: results from the SUDCAN population-based study.** *European Journal of Cancer Prevention* 2017, **26**:S16-S23.
183. Nie D, Lu J, Zhang H, Adeli E, Wang J, Yu Z, Liu L, Wang Q, Wu J, Shen D: **Multi-Channel 3D Deep Feature Learning for Survival Time Prediction of Brain Tumor Patients Using Multi-Modal Neuroimages.** *Scientific reports* 2019, **9**(1):1103.

184. Hosny A, Parmar C, Coroller TP, Grossmann P, Zeleznik R, Kumar A, Bussink J, Gillies RJ, Mak RH, Aerts HJ: **Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study.** *PLoS medicine* 2018, **15**(11):e1002711.
185. Bibault J-E, Giraud P, Durdux C, Taieb J, Berger A, Coriat R, Chaussade S, Dousset B, Nordlinger B, Burgun A: **Deep Learning and Radiomics predict complete response after neo-adjuvant chemoradiation for locally advanced rectal cancer.** *Scientific reports* 2018, **8**(1):12611.
186. Napel S, Mu W, Jardim-Perassi BV, Aerts HJ, Gillies RJ: **Quantitative imaging of cancer in the postgenomic era: Radio (geno) mics, deep learning, and habitats.** *Cancer* 2018, **124**(24):4633-4649.

Vita

Rachel Beth Ger was born in New York, New York to Marcia Claire Borden and Barry Michael Ger. After completing her work at Marlboro High School, Marlboro, New Jersey in 2010, she entered The University of North Carolina at Chapel Hill in Chapel Hill, North Carolina. In May 2014, Rachel received her Bachelor of Science in Physics and Bachelor of Arts in Mathematics from The University of North Carolina at Chapel Hill. In August 2014, she entered the Medical Physics PhD program at the MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences. She conducted her dissertation research under the guidance of Laurence Court, PhD.

Permanent address:

14610 Ballantyne Lake Road

Apartment 804

Charlotte, NC 28277