

5-2017

## Detecting and Evaluating Therapy Induced Changes in Radiomics Features Measured from Non-Small Cell Lung Cancer to Predict Patient Outcomes

Xenia J. Fave

Follow this and additional works at: [https://digitalcommons.library.tmc.edu/utgsbs\\_dissertations](https://digitalcommons.library.tmc.edu/utgsbs_dissertations)



Part of the [Medical Biomathematics and Biometrics Commons](#), [Multivariate Analysis Commons](#), and the [Other Physics Commons](#)

---

### Recommended Citation

Fave, Xenia J., "Detecting and Evaluating Therapy Induced Changes in Radiomics Features Measured from Non-Small Cell Lung Cancer to Predict Patient Outcomes" (2017). *The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences Dissertations and Theses (Open Access)*. 778.

[https://digitalcommons.library.tmc.edu/utgsbs\\_dissertations/778](https://digitalcommons.library.tmc.edu/utgsbs_dissertations/778)

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](mailto:digitalcommons@library.tmc.edu).

DETECTING AND EVALUATING THERAPY INDUCED CHANGES IN RADIOMICS  
FEATURES MEASURED FROM NON-SMALL CELL LUNG CANCER TO  
PREDICT PATIENT OUTCOMES

by

*Xenia Janice Favè, B.S.*

APPROVED:

\_\_\_\_\_  
Laurence E. Court, Ph.D.  
Advisory Professor

\_\_\_\_\_  
Peter Balter, Ph.D.

\_\_\_\_\_  
David Followill, Ph.D.

\_\_\_\_\_  
Daniel Gomez, Ph.D.

\_\_\_\_\_  
Aaron Kyle Jones, Ph.D.

\_\_\_\_\_  
Christine Peterson, Ph.D.

APPROVED:

\_\_\_\_\_  
Dean, The University of Texas  
MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences

DETECTING AND EVALUATING THERAPY INDUCED CHANGES IN RADIOMICS  
FEATURES MEASURED FROM NON-SMALL CELL LUNG CANCER TO  
PREDICT PATIENT OUTCOMES

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

Xenia Janice Favè, B.S.

Houston, Texas

August, 2017

## Dedication

Dedicated to my mom.

You were my first role model and are my biggest fan.

To you, I owe my work ethic, positivity, and scientific curiosity.

Thank you for always supporting me and giving me the foundation I needed to succeed.

## Acknowledgments

I would like to start by thanking my advisor, Dr. Laurence Court for his constant support and wisdom over the past five years. This work would not be possible without your help. I am so grateful to have been a part of your lab and hope to one day be as strong and positive of a mentor to future students.

Thank you to my committee, Drs. Peter Balter, David Followill, Daniel Gomez, Kyle Jones, Christine Peterson, and Francesco Stingo. At every meeting your input has been invaluable and has helped shape this project in innumerable ways. I appreciate all the support you have given me.

Thank you to my family who I have always been able to count on and who have helped me to continually look on the bright side of things. To my mom who always believed in me, my dad who gave me my faith, and my brothers, Francis and Sebastian, who have served as both inspiring heroes and comic relief.

Thank you to every member of the Courtyard, past and present who have all contributed to my success. Our culture of support and encouragement (with the occasional cathartic grumbles) have made arriving at work a pleasure. I will miss you all and our lunches, idea-fests, and happy hours but will do my best to spread our motto to every clinic I ever work in, "Friendship, science, and the greater good! For all mankind and beyond!"

Thank you to my great Houston friends, Olivia Popnoe, Erica Fried, Mattie McInnis, Rachel McCarroll, Joshua Loucks, Haden Popnoe, David Fried, and Travis Ray. My Houston life transformed when we all met and not just because I finally learned how to play football. My fondest memories of this city will always include our impromptu field days, adventures at run club, pool-Barbeques, Rockets games, beach rentals, and playing celebrity.

Thank you again to David Fried, who deserves his own special shout-out despite being in the last two groups, for telling me Dr. Court was looking for another student and thereby irrevocably shaping my graduate career. Thank you for introducing me to radiomics and R and most importantly for sitting next to me at work for three and a half years and being a willing sounding board every day.

Thank you to Hannah Lee. I cannot picture graduate school without you, and even if everything else had been a failure, I would have felt blessed to have met you and been able to count you among my friends. Thank you for all the fun, all the study sessions, and all the lunches. You are remarkable, and I cannot wait to see where life takes you next.

Thank you to Katherine Kellogg. I will never stop being thankful that you moved to 82<sup>nd</sup> street and shared my bus stop 16 years ago. You have been my best friend since then and have always been there when I needed someone to talk to. Thank you for listening to my many tedious descriptions of research problems and always having the best advice.

And finally, I want to thank Travis Ray. I have never met anyone who believed in me as much as you have. Your unshakable confidence in me and the sheer joy you have taken in each of my accomplishments continues to inspire me. Thank you for all the 5k's, movie nights, rodeos, concerts, trips, and picnics. I am thrilled to be starting this next big adventure with you.

# DETECTING AND EVALUATING THERAPY INDUCED CHANGES IN RADIOMICS FEATURES MEASURED FROM NON-SMALL CELL LUNG CANCER TO PREDICT PATIENT OUTCOMES

Xenia Janice Favè, B.S.

Advisory Professor: Laurence E. Court, Ph.D.

The purpose of this study was to investigate whether radiomics features measured from weekly 4-dimensional computed tomography (4DCT) images of non-small cell lung cancers (NSCLC) change during treatment and if those changes are prognostic for patient outcomes or dependent on treatment modality. Radiomics features are quantitative metrics designed to evaluate tumor heterogeneity from routine medical imaging. Features that are prognostic for patient outcome could be used to monitor tumor response and identify high-risk patients for adaptive treatment. This would be especially valuable for NSCLC due to the high prevalence and mortality of this disease.

A novel process was designed to select feature-specific image preprocessing and remove features that were not robust to differences in CT model or tumor volumes. These features were then measured from weekly 4DCT images. These features were evaluated to determine at which point in treatment they first begin changing if those changes were different for patients treated with protons versus photons. A subset of features demonstrated significant changes by the second or third week of treatment, however changes were never significantly different between patient groups. Delta-radiomics features were defined as relative net changes, linear regression slopes, and end of treatment feature values. Features were then evaluated in univariate and multivariate

models for overall survival, distant metastases, and local-regional recurrence. In general, the delta-radiomics features were not more prognostic than models built using clinical factors or features at pre-treatment. However one shape descriptor measured at pre-treatment significantly improved model fit and performance for overall survival and distant metastases. Additionally for local-regional recurrence, the only significant covariate was texture strength measured at the end of treatment. A separate study characterized radiomics feature variability in cone-beam CT images to increased scatter, increased motion, and different scanners. Features were affected by all three parameters and specifically by motion amplitudes greater than 1 cm.

This study resulted in strong evidence that a set of robust radiomics features change significantly during treatment. While these changes were not prognostic or dependent on treatment modality, future studies may benefit from the methodologies described here to explore delta-radiomics in alternative tumor sites or imaging modalities.

## Table of Contents

Dedication .....	iii
Acknowledgments .....	iv
Abstract .....	vi
Table of Contents .....	viii
List of Illustrations .....	xii
List of Tables .....	xvii
Chapter 1 : Introduction .....	1
Chapter 2 : Purpose and Central Hypothesis .....	5
Central Hypothesis .....	5
Specific Aim 1: Calculating Radiomics Features .....	5
Specific Aim 2: Prognostic Potential of Radiomics Features .....	5
Specific Aim 3: Reproducibility of Radiomics Features from CBCT Images .....	6
Specific Aim 4: Treatment Modality Dependence of Radiomics Features .....	6
Chapter 3 : General Methodology .....	8
Patient Cohort .....	8
Imaging Parameters .....	9
Tumor Region of Interest .....	10
Exclusion Criteria .....	11
Clinical Factors .....	12
Radiomics Calculation Software .....	13
Image Preprocessing .....	15
Radiomics Features .....	15
Histogram Features .....	15
Co-occurrence Matrix Features .....	16
Gray-Level Run-Length Matrix Features .....	16
Neighborhood Gray-Tone Difference Matrix Features .....	17
Laplacian of Gaussian Filtered Features .....	18
Shape Features .....	18
Feature Naming Conventions .....	18
Statistical Software .....	26
Chapter 4 : Calculating Radiomics Features .....	27
Introduction .....	27

Methods .....	29
Images .....	29
Exploratory Analysis of Image Pre-processing Techniques .....	30
Image Pre-processing .....	34
Features.....	36
Volume-Dependence.....	38
CT Model Dependence .....	41
Prognostic Potential .....	41
Results .....	42
Exploratory Analysis of Image Preprocessing Techniques .....	42
Volume Dependence .....	53
CT Model Dependence .....	60
Prognostic Potential .....	62
Prognostic Potential versus Volume Dependence .....	68
Discussion .....	69
Conclusions.....	73
Chapter 5 : Prognostic Potential of Radiomics Features .....	75
Introduction.....	75
Methods .....	76
Patient Data .....	76
Landmark Analysis.....	78
Features.....	78
Delta-radiomics Features .....	84
Univariate Analysis .....	85
Multivariate Analysis .....	86
Results .....	89
Feature Selection.....	89
Delta-radiomics Features .....	99
Univariate Analysis .....	99
Multivariate analysis .....	105
Discussion .....	116
Conclusions.....	120
Chapter 6 : Reproducibility of Radiomics Features from CBCT Images.....	122
Introduction.....	122

Methods .....	123
Patient Test-Retest CBCT Images .....	123
Texture Phantom.....	125
Texture Features & Pre-Processing .....	126
Effect of Scanners.....	128
Effect of Scatter .....	131
Effect of Motion .....	132
Results .....	134
Patient Test-Retest CBCT Images .....	134
Effect of Different Scanners .....	138
Effect of Scatter .....	142
Effect of Motion .....	147
Discussion .....	152
Patient Test-Retest .....	152
Inter-Scanner Analysis .....	153
Effect of Scatter .....	154
Effect of Motion .....	155
Overall Best Performing Features .....	156
Conclusions.....	157
Chapter 7 : Treatment Modality Dependence of Radiomics Features.....	158
Introduction.....	158
Methods .....	159
Features.....	159
Patients.....	161
ANOVA Classification.....	161
Analysis.....	162
Clinical Factors.....	166
Results .....	167
ANOVA Results .....	167
Treatment Dependent Features.....	172
Treatment & Model Dependent Features.....	177
Discussion .....	183
Conclusions.....	185
Chapter 8 : Discussion .....	186

Future Directions .....	190
Conclusions.....	194
Appendix A: IBEX Parameter Sheets.....	196
Bibliography.....	210
Vita .....	226

## List of Illustrations

Figure 3.1: Axial image slices for two patients at the beginning and end of treatment. ....	11
Figure 4.1: Sample image of a patient tumor ROI using each preprocessing technique. ....	35
Figure 4.2: Visualization of the co-occurrence matrix for the same tumor ROI after four different image preprocessing techniques. ....	35
Figure 4.3: Axial views of the 3D spherical digital phantoms created for this study. ....	40
Figure 4.4: Correlation between features calculated after bit depth resampling between 11 and 6 bits and their default values without any bit depth resampling (BD12). ....	45
Figure 4.5: Correlation between features calculated after different smoothing filters to their default values from images smoothed with a Butterworth filter with a cutoff of 125. ....	47
Figure 4.6: Correlation between features calculated after Butterworth smoothing filters with different cutoffs (75,100,150) and an 8 bit depth resample to their default values from images smoothed with a Butterworth filter with a cutoff of 125 and an 8 bit depth resample. ....	49
Figure 4.7: Correlation between features calculated after three filters to their default values from images without any image preprocessing (BD12). ....	51
Figure 4.8: The volume correlation for the features measured using the spearman rank correlation coefficient. ....	55
Figure 4.9: The absolute value of Spearman correlation coefficients plotted as a histogram for each preprocessing technique. ....	57
Figure 4.10: Radiomics feature values measured from digital phantoms before and after volume normalization of the feature algorithms. ....	59
Figure 4.11: The Benjamini-hochberg corrected p-values for the Wilcoxon rank-sum test comparing values measured on two different CT models. ....	61
Figure 4.12: The p-values after Benjamini-Hochberg correction for the univariate cox proportional hazards model for each feature and preprocessing combination. ....	63

Figure 4.13: Harrell's concordance index (c-index) for each feature and preprocessing combinations.....	65
Figure 4.14: The Benjamini-Hochberg corrected p-values for the log-likelihood ratio between cox proportional hazards models with only volume as a covariate and models with volume and one radiomics feature at a time. ....	67
Figure 4.15: Effect of preprocessing on the volume correlation versus the added prognostic value of the features.....	69
Figure 5.1: Workflow for the selection of feature specific image preprocessing. ....	83
Figure 5.2: Workflow for building of multivariate models. ....	88
Figure 5.3: Impact of image preprocessing on the univariate significance of radiomics features. ....	91
Figure 5.4: Impact of image preprocessing on the significance of CT Model in a Wilcoxon rank sum test for each radiomics feature.....	93
Figure 5.5: Impact of image preprocessing on the univariate significance and significance of the CT Model in a Wilcoxon rank sum test for each radiomics feature.....	95
Figure 5.6: Final image preprocessing that was selected for each radiomics feature used in the prognostic analysis.....	97
Figure 5.7: The Benjamini-Hochberg corrected p-values for the log-likelihood ratio of each univariate model using one of the clinical factors.....	101
Figure 5.8: C-indices for univariate clinical models calculated using a LOOCV loop to generate patient-specific predictions for three outcomes.....	102
Figure 5.9: The Bonferroni corrected p-values for the log-likelihood ratio of each univariate model using four different versions of each radiomics feature. ....	103
Figure 5.10: C-indices for univariate radiomics models calculated using a LOOCV loop to generate patient-specific predictions for three outcomes.....	104

Figure 5.11: Kaplan-Meier curves for overall survival using clinical factors alone to generate Cox proportional hazards models. ....	110
Figure 5.12: Kaplan-Meier curves for overall survival using clinical factors and pre-treatment radiomics features to generate Cox proportional hazards models. ....	111
Figure 5.13: Kaplan-Meier curves for overall survival using clinical factors, pre-treatment features, and delta-radiomics features to generate Cox proportional hazards models. ....	112
Figure 5.14: Kaplan-Meier curves for distant metastases using clinical factors alone to generate Cox proportional hazards models. ....	113
Figure 5.15: Kaplan-Meier curves for distant metastases using clinical factors and pre-treatment features to generate Cox proportional hazards models. ....	114
Figure 5.16: Kaplan-Meier curves for local-regional recurrence using clinical factors, pre-treatment features, and delta-radiomics features to generate Cox proportional hazards models. ....	115
Figure 6.1: The Credence Cartridge Radiomics Phantom. ....	126
Figure 6.2: Axial images of the radiomics phantom with added scatter material. ....	132
Figure 6.3: Images of the CIRS dynamic motion phantom. ....	134
Figure 6.4: The absolute differences between pairs of scans plotted by the types of groups being compared. ....	139
Figure 6.5: The impact of scatter on the reproducibility of radiomics features extracted from the shredded rubber cartridge. ....	143
Figure 6.6: The impact of scatter on the reproducibility of radiomics features extracted from the dense cork cartridge. ....	145
Figure 6.7: The impact of motion on the reproducibility of radiomics features extracted from the dense cork insert in the motion phantom using the full 3D ROI. ....	148
Figure 6.8: The impact of motion on the reproducibility of radiomics features extracted from the dense cork insert in the motion phantom using a 2D ROI from an axial image slice. ....	150

Figure 7.1: List of features included in the analysis of modality dependence and the image preprocessing used to calculate each. ....	160
Figure 7.2: Results of the ANOVA analysis for each feature to determine dependence on treatment and CT scanner model. ....	168
Figure 7.3: Example features for each of the ANOVA results. ....	170
Figure 7.4: Results of Wilcox sign-rank test comparing patients' radiomics feature values at each week of treatment to their values at week 1 for the treatment dependent features. ....	172
Figure 7.5: Boxplots of the radiomics features values through treatment for the treatment dependent features. ....	173
Figure 7.6: Results of Wilcox rank-sum test comparing the patients' radiomics feature values by treatment modality at each week for the treatment dependent features. ....	174
Figure 7.7: Results of Wilcox rank-sum test comparing the net changes between the patients' radiomics feature values at each week of treatment to their values at week 1 by modality for the treatment dependent features. ....	175
Figure 7.8: Boxplots of the radiomics feature values through treatment and by treatment modality for the treatment dependent features. ....	176
Figure 7.9: Results of Wilcox sign-rank test comparing patients' radiomics feature values at week 5 to their value at week 1 for the treatment & model dependent features. ....	177
Figure 7.10: Boxplots of the radiomics features values at weeks 0 and 5 for the treatment & model dependent features. ....	178
Figure 7.11: Results of Wilcox rank-sum test comparing the patients' radiomics feature values by treatment modality at weeks 0 and 5 for the treatment & model dependent features. ....	180
Figure 7.12: Results of Wilcox rank sum test comparing the net changes between the patients' radiomics feature values at week 0 to week 5 by modality for the treatment & model dependent features. ....	181

Figure 7.13: Boxplots of the radiomics feature values through treatment and by treatment  
modality for the treatment & model dependent features. .... 182

## List of Tables

Table 3.1: Clinical factors used in this study. ....	14
Table 3.2: Features used in this study and their abbreviations used in subsequent tables and figures. ....	19
Table 4.1: Summary of the clinical characteristics of the study population, n=107. ....	30
Table 4.2: The radiomics features used in the analysis of image preprocessing techniques. ....	37
Table 4.3: The original and normalized algorithms for the volume-dependent features. ....	56
Table 5.1: Clinical characteristics of the NSCLC patient population used for modeling. ....	77
Table 5.2: List of features used in the analysis of univariate and multivariate prognostic potential. ....	80
Table 5.3: Final comparison of the three models for each outcome. ....	105
Table 5.4: The number of times each clinical factor and radiomics feature was selected in the first leave-one-out cross validation for each outcome. ....	106
Table 6.1: Clinical characteristics for the test-retest CBCT patient population. ....	123
Table 6.2: Features that were included in the analysis of CBCT feature reproducibility. ....	127
Table 6.3: Scan characteristics for the phantom CBCT images used in this study. ....	130
Table 6.4: The results of the CCC and $r_s$ tests for the patient test-retest data. ....	136
Table 6.5: Results of the inter-scanner variability test for the shredded rubber ROI. ....	140
Table 6.6: Results of the inter-scanner variability test for the dense cork ROI. ....	141
Table 7.1: Number of patients with an image available at each week for each cohort. ....	165
Table 7.2: Number of patients at each week in each cohort used in calculations. ....	165
Table 7.3: Clinical characteristics of the patients treated with protons versus photons. ....	166
Table 7.4: Differences in clinical factors for patients treated with photons versus protons (n=110). ....	167
Table 7.5: P-values from the ANOVA analysis. ....	171

# Chapter 1 : Introduction

Lung cancer is responsible for the majority of cancer deaths in the United States for both men and women<sup>1</sup>. In 2016 there were an estimated 224,390 new lung cancer cases<sup>1</sup>. Of these, up to 85% were diagnosed as non-small cell lung cancer (NSCLC)<sup>1,2</sup>. While advances in cancer care have substantially improved outcomes for particular tumor sites over the past 20 years, little change has been seen in the NSCLC patient outcomes<sup>3</sup>. Additionally, patients with similar clinical factors and pathological characteristics can have very different outcomes<sup>4</sup>. As a result, multiple studies have sought prognostic biomarkers that would identify high-risk patients in order to tailor their treatment<sup>5-9</sup>. One novel approach for identifying high-risk NSCLC patients is the use of radiomics features.

Radiomics is the process of calculating quantitative imaging features from medical images and using these values to characterize the tumor or predict a clinical outcome<sup>10</sup>. These features can be simple intensity-derived metrics such as the mean value or standard deviation of the pixels within the tumor. More sophisticated metrics also exist that aim to capture the spatial heterogeneity of pixels in the tumor. These include descriptive features such as contrast or busyness and are measured from radiomics matrices such as the co-occurrence matrix<sup>11,12</sup>, run length matrix<sup>13</sup>, or neighborhood difference matrix<sup>14</sup>. A variety of shape features also exist to quantitatively describe how smooth or spiculated the tumor appears. Useful features are believed to reflect tumor-specific phenotypes and biology. Radiomics features have the advantage of measuring regional heterogeneity differences from the entire 3 dimensional tumor which are known to vary spatially in solid tumors<sup>15-17</sup>. This advantage contrasts with biopsies that can only sample specific points and may miss the heterogeneous areas linked to malignancy. Additionally, repeated tumor biopsies can lead to patient complications and particularly for lung cancer can result in pneumothorax<sup>18</sup> while radiomics analyses are inherently non-invasive. In fact, features are typically measured from medical images that are

already routinely acquired so large datasets for radiomics analyses can be acquired without interrupting the clinical workflow.

Radiomics studies have had success using computed tomography (CT), fluorodeoxyglucose(FDG) positron emission tomography (PET), and magnetic resonance (MR) images for a multitude of tumor sites. Radiomics has been particularly successful in studies using CT images of NSCLC tumors. Much of the early groundwork for this area focused on classifying tumors either by histology or malignancy. For example, an early study tested the ability of 102 2D and 215 3D features to classify 74 lung tumors as adenocarcinoma or squamous-cell carcinoma<sup>19</sup>. They achieved 68% classification accuracy by using a decisions tree based classifier<sup>19</sup>. Song *et al* also investigated the ability of 592 radiomics features to correctly classify histology and by using support vector machines were able to achieve 75% classification accuracy<sup>20</sup>. Another study used radiomics features from 72 mediastinal lung cancer nodes to classify the nodes as benign or malignant and achieved 81% sensitivity with 80% specificity<sup>21</sup>. A separate study also classified small nodules as benign or malignant using radiomics features and identified five features from the co-occurrence matrix that could be used as independent classifiers<sup>22</sup>.

There have also been a variety of radiomics studies focused on predicting outcomes for NSCLC patients. The majority of these have examined predicting overall survival such as a study by Aerts *et al* which developed and validated a radiomics signature in both lung cancer and head and neck cancer patient cohorts<sup>23</sup>. Other studies have indicated that radiomics features can act as independent predictors of overall survival<sup>24,25</sup> and significantly improve risk stratification compared to clinical factors alone<sup>26,27</sup>. Studies predicting risk of distant metastases were able to develop a radiomics-signature with a c-index of 0.61<sup>28</sup> and improve risk stratification compared to clinical models alone<sup>26</sup>. Studies have also investigated radiomics for predicting recurrence in NSCLC. One study used 101 early stage patients with adenocarcinoma and predicted tumor recurrence with a c-index of 0.81 and AUC of 0.79<sup>29</sup>. Koo

*et al* used radiomics features measured from the preoperative CT images to predict tumor recurrence<sup>30</sup> while Mattonen *et al* used the post-treatment follow-up CT images to predict if the patient was developing radiation induced lung injury or recurrence<sup>31</sup>. In summary, many groups have reported success in using radiomics features to either classify tumor types or predict patient outcomes.

These studies have prompted research into the underlying biology that drives these features. Recently genetic mutations in NSCLC tumors have been linked to radiomics features. Specifically, kurtosis and skewness from the histogram were shown to be related to KRAS mutations<sup>32</sup> while features from the gray-level co-occurrence matrix were linked to EGFR mutations<sup>33</sup>. Both of these mutations are connected to the MAPK pathway which is known to be prognostic for survival<sup>34</sup>. These results suggest that there is a biological explanation for the success of certain features measured from NSCLC in predicting patient outcomes or classifying tumors.

In all of the afore-mentioned studies the radiomics features were measured at one time point and typically prior to treatment. If these features are measuring tumor phenotypes then the changes in the features during treatment may be useful as biomarkers of response. Currently, tumor response is typically measured using the Response Evaluation Criteria In Solid Tumor (RECIST) guidelines<sup>35,36</sup>. These guidelines primarily evaluate response through the change in size of the tumor which is measured as either the largest diameter or overall volume<sup>35</sup>. Studies using changes in radiomics features (delta-radiomics) have been used in other areas of radiomics research. One study used delta-radiomics features to measure and predict pneumonitis from CT images<sup>37</sup>. For CT images of colorectal liver metastases, the relative differences in uniformity and entropy measured pre- and post- chemotherapy were more prognostic of tumor response than changes in size or volume<sup>38</sup>. Another study examined the role of a feature, mean of positive pixels, measured from contrast enhanced CT images of soft tissue sarcomas to predict pathological response and showed that it outperformed size,

density, and tumor blood flow<sup>39</sup>. However no study has yet examined the potential for delta-radiomics to assess NSCLC tumor response.

While radiomics features have been successfully used in a variety of studies, there are also several challenges associated with their use. Because they are statistical metrics calculated from the pixels of the tumor image, they are highly dependent on the parameters used to acquire the images as well as any artifacts within the image. For CT images in particular, features have been shown to be affected by the scanner used to acquire the images<sup>40</sup> and the imaging parameters<sup>41</sup>. Even in test-retest studies of the same patient on the same scanner or studies using multiple contours on the same images, features have shown variability<sup>42,43</sup>. Additionally, radiomics features can be calculated using different parameters or image processing techniques, such as gray-level discretization, and the specific choice of parameters can impact the feature reproducibility<sup>44</sup>. Controlling for these uncertainties is challenging because of the lack of a ground truth for the feature values.

The main goal of this study was to evaluate whether delta-radiomics features measured from CT images of NSCLC are independent, useful biomarkers of tumor response when compared to clinical factors and radiomics features measured prior to treatment. An identified subset of useful features would be beneficial for evaluating an individual patient's tumor response and potentially altering their treatment or follow-up care.

## Chapter 2 : Purpose and Central Hypothesis

### **Central Hypothesis**

Radiomics features measured from NSCLC tumors change significantly during treatment, and the magnitude of those changes are characteristic of treatment modality and prognostic for outcome.

### **Specific Aim 1: Calculating Radiomics Features**

**Aim:** Analyze the effect of image preprocessing on a set of radiomics features measured from CT images and identify the optimal preprocessing for each.

**Hypothesis:** The image preprocessing technique used has a significant impact on the prognostic ability and volume independence of radiomics features.

**Project 1.1:** Analyze the impact of image pre-processing on the volume dependence of radiomics features.

**Project 1.2:** Analyze the impact of image pre-processing on the univariate significance of radiomics features.

**Project 1.3:** Analyze the impact of image pre-processing on the scanner independence of radiomics features.

### **Specific Aim 2: Prognostic Potential of Radiomics Features**

**Aim:** Determine which radiomics features, measured from CT images, change significantly during the course of radiation therapy and the relationship between these changes and outcome.

**Hypothesis:** Radiomics features can be identified that change during the course of treatment and are predictive for patient outcome.

**Project 2.1:** Identify radiomics features that change significantly with dose.

**Project 2.2:** Determine the univariate significance of the radiomics features and clinical factors in predicting 3 outcomes.

**Project 2.3:** Perform a multivariate analysis including radiomics features and clinical factors in order to predict 3 outcomes.

### **Specific Aim 3: Reproducibility of Radiomics Features from CBCT Images**

**Aim:** Identify which radiomics features can be reproducibly measured from cone-beam CT (CBCT) images.

**Hypothesis:** A subset of radiomics features can be identified that are robust to the increased scatter and motion present during CBCT imaging.

**Project 3.1:** Evaluate the impact of using different CBCT imagers to acquire images on the extracted radiomics features.

**Project 3.2:** Evaluate the impact of different thicknesses of scatter material on the extracted radiomics features.

**Project 3.3:** Evaluate the impact of motion during imaging on the extracted radiomics features.

### **Specific Aim 4: Treatment Modality Dependence of Radiomics Features**

**Aim:** Compare the changes in radiomics features from patients treated with intensity modulated radiation therapy (IMRT) to those treated with passive scatter proton therapy (PSPT).

**Hypothesis:** The changes in radiomics features measured from patients treated with protons will occur earlier in treatment compared to those treated with photons due to the increased RBE.

**Project 4.1:** Compare the distribution of values at the beginning and end of treatment for the two treatment modalities

**Project 4.2:** Quantify when in treatment radiomics features first exhibit significant changes from baseline and if those changes are modality-dependent at any point in treatment.

## Chapter 3 : General Methodology

Portions of this chapter are written or based on the following publications:

Fave, X, Mackin, D, Yang, J, Zhang, J, Fried, D, Balter, P, Followill, D. , Gomez, D, Jones, AK, Stingo, F, Fontenot, J, and Court, L. Can radiomics features be reproducibly measured from CBCT images for patients with non-small cell lung cancer? Medical Physics doi: 10.1118/1.4934826. Volume 42, Issue 12, pages 6784-6797. 2015. ©John Wiley and Sons.

Fave X, Zhang L, Yang J, Mackin D, Balter P, Gomez D, Followill D, Jones AK, Stingo F, Court LE. Impact of image preprocessing on the volume dependence and prognostic potential of radiomics features in non-small cell lung cancer. Translational Cancer Research. Doi: 10.21037/tcr.2016.07.11. Volume 5, Issue 4, pages 349-363.(c) AME Publishing Company.

The permission for reuse of this material was obtained from John Wiley and Sons © and AME Publishing Company© respectively.

The typical workflow for any radiomics study is to gather a set of patient images, define the region of interest (ROI) on each, extract a set of radiomics features from these ROIs, and then perform the statistical analysis. In this chapter we describe these parts of the methodology that were universal through the analyses for the 4 aims. First the initial patient cohort used in this study and the exclusion criteria we used are described. Then our methodology for delineating the tumor ROI on these patient's images is explained. Finally details of the radiomics software and radiomics features are given. Further details that are particular to a specific study are described in the relevant subsequent chapters.

### **Patient Cohort**

The following analyses were performed using the images, clinical factors, and outcomes data from a set of 157 patients treated at the University of M.D. Anderson between 2009-2014 as part of an institutional review board (IRB) approved clinical trial comparing outcomes and normal tissue toxicity between patients treated with intensity modulated radiation therapy

(IMRT) and passive scatter proton therapy (PSPT)<sup>45</sup>. The retrospective review conducted for this study was approved by the University of Texas M.D. Anderson Cancer Center IRB with a waiver of informed consent. Patients on the original trial were treated with concurrent chemotherapy. The inclusion criteria for patients on that trial were the presence of a pathologically proven, unresected, locoregionally advanced, stage II-IIIb NSCLC tumor, a Karnofsky performance score  $>70$  or ECOG score 0-1, measurable disease on a chest x-ray, contrast enhanced CT, or PET scan, a forced expiratory volume in the first second  $\geq 1$  liters, acquisition of a fluorodeoxyglucose (FDG)-PET scan 3 months before trial registration, between 18-85 years old, and a signed informed consent form<sup>45</sup>. The exclusion criteria were diagnosis of small cell tumor histology, prior radiation to the treatment field, pregnancy, body weight that exceeded the limits of the treatment couch, or being oxygen dependent due to preexistent lung disease<sup>45</sup>. For the analyses in this dissertation, extra exclusion criteria were implemented for each specific aim to ensure uniformity in the dataset and/or robustness of the radiomics features. These specific criteria are addressed independently where they apply in chapters 4-7 and are summarized in the Exclusion Criteria section below.

## **Imaging Parameters**

As part of the aforementioned clinical trial, patients were imaged during each week of their treatment with a 4 dimensional CT (4DCT) using the institutional protocol: peak tube voltage of 120 kVp, tube current of 100 or 200 mA, and rotation time of 0.5 or 0.8 seconds. Images were reconstructed into a 512x512 pixel matrix with an image thickness of 2.5 mm and in-plane resolution of 0.98 mm. Additionally, a pre-treatment 4DCT was acquired for treatment planning purposes using the same institutional protocol. Images were acquired using a GE Discovery ST (GE Medical, Waukesha, WI), GE LightSpeed RT16, Phillips Brilliance Big Bore (Philips Medical Systems, Cleveland, OH), or Philips Brilliance 64 CT scanner.

Patients also received cone-beam CT (CBCT) images for setup verification periodically through treatment. All of the CBCT patient images were acquired using the thoracic imaging

protocol on a Varian linac: peak tube voltage of 110 kVp, tube current of 20 mA, and exposure time (total pulsed beam-on time) of 7-14 seconds. Images were reconstructed as a 512x512 grid with pixel dimensions of 0.8 mm and a 2.5mm slice thickness.

## **Tumor Region of Interest**

In the subsequent analyses the end-of-exhale phase images for each 4DCT scan were used for feature extraction. This phase was selected because it was considered the most reproducible for patients and has been used in other radiomics studies<sup>26,46,41</sup>. Additionally, one study demonstrated that radiomics features consistently order patients regardless of phase from T20 to T90, where T50 represents the end of exhale phase<sup>41</sup>. The three-dimensional gross tumor volume contour from the treatment plan was used as the region of interest (ROI) for feature extraction. The gross tumor volume contour from the treatment plan was deformably registered to each subsequent weekly 4DCT scan using a clinical software, CT-assisted targeting, developed in-house<sup>47-49</sup>. During this step, only the contour is deformed, the images themselves are not changed. Examples of deformed contours are in Figure 3.1. After the contours were deformed, each image was examined to ensure consistency and make minor adjustments. Adjustments were most frequently needed for tumors connected to the mediastinum. In all cases, adjustments were made conservatively: that is, contouring only tumor was prioritized over contouring anything that could be tumor. This approach was successfully used in previous radiomics studies in our group<sup>26</sup>. Furthermore, to ensure that normal lung and bone were excluded from the final ROI used for feature calculation, a thresholding step was also applied to each CT image with a lower threshold of -100 Hounsfield Units (HU) and upper threshold of 200 HU.

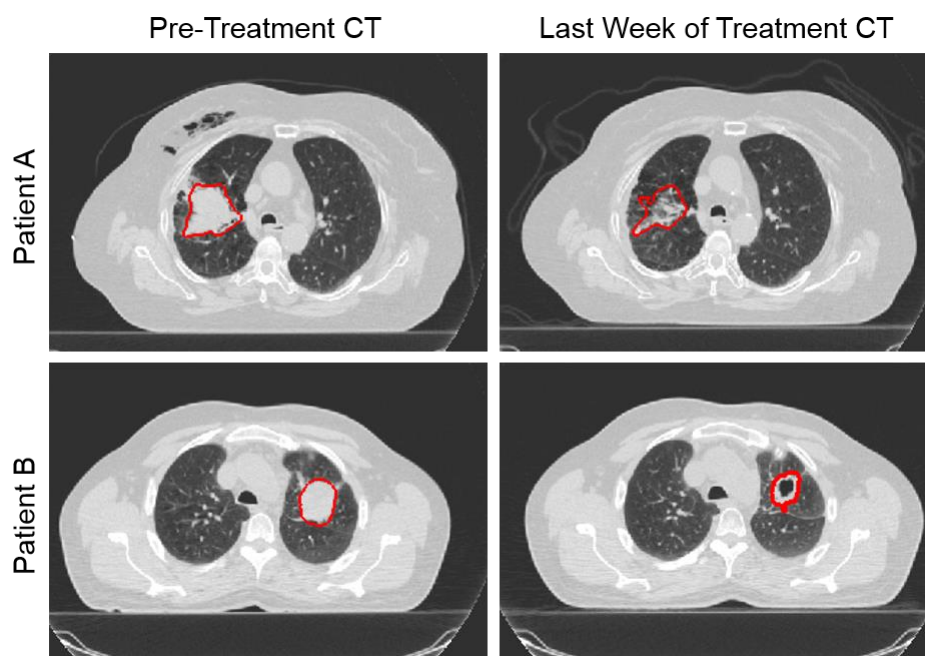


Figure 3.1: Axial image slices for two patients at the beginning and end of treatment. The contour from the pre-treatment CT was deformed to each subsequent weekly 4DCT image. The two examples shown here demonstrate the robustness of this process to tumor shrinkage (Patient A) and cavitation (Patient B).

## Exclusion Criteria

Of the 157 patients, 29 never had their GTV contour deformed to the weekly images and all of their images were left out of the analysis entirely following the following criteria:

1. The patient's weekly images were acquired using breath hold instead of 4DCT, n=9 patients.
2. The patient had no GTVp in their plan (only nodal disease apparent), n=11 patients.
3. The patient's primary tumor was very small at beginning of treatment and disappeared before the end of imaging, n=9 patients.

The remaining 128 patients had the GTV contour deformed to each of their weekly images and radiomics features were calculated from each image. These patients had between 4-9 images each. The following extra exclusion criteria were applied by calculating the volume of the deformed GTV on each image:

4. The patient was excluded if the measured tumor volume was <5cc at the beginning of treatment, n=18 patients.
5. A patient's particular weekly image was excluded if the measured tumor volume at that week fell below 5 cc, n=28 images from 7 patients.

This left 110 patients with between 2-9 images each. When the methods included a survival analysis then a landmark time point was used. This technique is explained in detail in Chapter 5 and required the following extra criterion to be applied:

6. Patients who experienced the event in a survival analysis before the landmark time were excluded, n=3 for overall survival, n=4 for time until distant metastases, and n=4 for time until local-regional recurrence.

This left 107 or 106 patients for survival analyses.

## **Clinical Factors**

Clinical factors were obtained through a retrospective review of the patient charts and are listed in Table 3.1. These factors were used as clinical covariates in Aim 2 and thus are tabulated for the patient subset of 107 patients that was used in Aim 2. Many of the clinical factors were split into levels for ease of model building as well as to reflect actual clinical impact (e.g. age>65 is more meaningful for treatment and outcome than that a patient's age is exactly 68). The levels used for each covariate are also listed in Table 3.1.

## **Radiomics Calculation Software**

The Imaging Biomarker EXplorer (IBEX) software package version 1.0 was used for the calculation of all radiomics features in this study<sup>50</sup>. This software is an open-source, MATLAB based package that is freely available online at [http://bit.ly/IBEX\\_MDAnderson](http://bit.ly/IBEX_MDAnderson). IBEX allows the user a large amount of freedom in how the features are calculated. This includes the ability to preprocess the image with different imaging filters, and the flexibility to adjust parameters of the different radiomics matrices such as the co-occurrence matrix bin size and offset.

Table 3.1: Clinical factors used in this study.

Clinical Factors	Number of Patients (n=107)
Sex	
F	45
M	62
Age	
<65	45
>=65	62
T stage	
T1 or T2	49
T3 or T4	58
N stage	
N0 or N1	24
N2 or N3	83
Overall disease stage	
II	12
IIIa	44
IIIb	49
IV	2
Tumor histology	
Squamous cell carcinoma	46
Adenocarcinoma or other	61
Smoking status	
Current	34
Former	64
Never	9
Pack years (continuous)	
0-24	20
25-49	37
50-74	28
75+	22
Karnofsky performance status	
90-100	52
70-80	55
Total radiation dose	
>70 Gy	72
<70 Gy	35

## Image Preprocessing

Prior to radiomics feature calculation, images were often filtered with either a smoothing filter such as the Butterworth filter or with bit depth resampling. In both cases, the aim of the filtration was to reduce the noise in the image and thus produce more informative values for the radiomics features. The impact of different filters on the features was explored as part of Aim 1 in Chapter 4 and full details of the filters are available there. The subsequent filtration selected and used for each feature is detailed under each specific study.

## Radiomics Features

The majority of features currently used in radiomics studies were designed to evaluate visual perception characteristics (i.e. contrast) quantitatively. These include co-occurrence matrix<sup>11,12</sup>, run-length matrix<sup>13</sup>, and neighborhood gray-tone difference matrix features<sup>14</sup>. Other commonly used radiomics feature categories include those calculated from the histogram or histogram features calculated after Laplacian of Gaussian filtration<sup>51,52</sup>. Each radiomics feature category included in this study is explained in more depth below. Different feature subsets were used within each aim and the particular feature selection criteria are thus described independently in Chapters 4-7.

### Histogram Features

Features derived from the histogram category are those that can be calculated from the full distribution of intensities in the ROI with or without binning. These are commonly referred to as first order statistics. Features included in this category are: energy, entropy, kurtosis, maximum, mean, median, minimum, skewness, standard deviation, uniformity, and variance. The majority of histogram features were calculated from the full distribution of intensities (e.g. bin size=1). A few features require a histogram to be calculated (entropy and uniformity) and in these cases the histogram bin size was 16. In both cases the full distribution of values in the

ROI after thresholding was used. Full descriptions and algorithms for each of these features are well documented in the literature<sup>23,53,54</sup>.

### Co-occurrence Matrix Features

The co-occurrence matrix was first defined by Robert Haralick *et al* in 1973<sup>12</sup>. The matrix quantifies the frequency at which each gray-level intensity appears adjacent to each other gray-level intensity in a particular direction. Thus the matrix represents spatial information from the image and is designed to represent what an observer would perceive as the texture of the image. The features derived from the matrix are considered second-order features because they capture more information from the image than first-order histogram based features. In this analysis the bin sizes for gray-level intensities were always set to 1, while the images were often rescaled to 8 bit images prior to feature calculation, thus creating bins of size 16 Hounsfield units. The co-occurrence matrix was always calculated in the four 2D directions (0°, 45°, 90°, and 135°) for each axial slice. The co-occurrence matrices for a particular direction are then summed over the set of axial slices. These direction-specific matrices are then summed and averaged to create the final co-occurrence matrix for the 3D tumor ROI. These steps are performed automatically by the IBEX software when calculating the co-occurrence matrix. Features calculated from this matrix were auto-correlation, cluster prominence, cluster shade, cluster tendency, contrast, correlation, difference entropy, dissimilarity, energy, entropy, homogeneity, 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, and variance. Full descriptions and algorithms for each of these features are well documented in the literature<sup>11,12,23,53,54</sup>.

### Gray-Level Run-Length Matrix Features

The gray-level run-length matrix, hereafter described as the run-length matrix, was defined by Mary Galloway in 1975<sup>13</sup> to classify images of terrain. Runs are defined as

consecutive, collinear image pixels with the same gray level intensity value<sup>13</sup>. The run-length matrix is a tally of the number of runs of each length versus each gray-level intensity. Similar to the co-occurrence matrix, the run-length matrix is calculated for a particular direction (0°, 45°, 90°, or 135°). For this study, the run-length matrices in this study were calculated in the 0° and 90° direction and then summed and averaged to create a global 3D run-length matrix from which to calculate the features. This step is automatically performed within IBEX. The other 2D directions were not used because they have not yet been implemented within IBEX. Bins of size 1 were used in all cases, although the images themselves were often resampled to 8 bits resulting in a de facto bin size of 16 HU. In calculating features from this matrix, different areas can be emphasized to highlight noise (short runs) or signal (long runs). The features calculated from this matrix were short runs emphasis, long runs emphasis, gray level non-uniformity, run length non-uniformity, run percentage, short run low gray level emphasis, short run high gray level emphasis, long run low gray level emphasis, long run high gray level emphasis, low gray level run emphasis, and high gray level run emphasis. Descriptions and algorithms for each of these features are available in the literature<sup>23,13,53,54</sup>.

#### Neighborhood Gray-Tone Difference Matrix Features

The neighborhood gray-tone difference matrix, hereafter described as the neighborhood difference matrix, was designed by Moses Amadasun in 1989 to accurately quantify human perception of five visual characteristics: coarseness, contrast, busyness, complexity, and texture strength<sup>14</sup>. The neighborhood difference matrix is a one-column matrix with an entry for each gray-level intensity in the image. To obtain the values in the matrix, first the average difference between each pixel and its neighbors is calculated. These average differences are then summed for each pixel of the same gray-level intensity. The neighborhood is the set of surrounding pixels at a specified offset (typically 1, 2, or 3 resulting in a square of 3x3, 5x5, or 7x7 centered on the pixel of interest). As a result, the pixels at the periphery of the image do not contribute directly to the neighborhood difference matrix. In this study, gray-level bins of 1

and a neighborhood size of 5x5 were typically used. All five of the defined features for the neighborhood difference matrix were calculated: coarseness, contrast, busyness, complexity, and texture strength. Algorithms and descriptions for each can be found in the literature<sup>14,53,54</sup>.

#### Laplacian of Gaussian Filtered Features

The Laplacian of Gaussian filtered features are features from the histogram that are calculated from an image after a Laplacian of Gaussian filter has been applied to it. The Laplacian of Gaussian filter smooths the image and then highlights the edges that remain. This technique has been used in radiomics studies to potentially reveal structure in the tumor<sup>27,51,52</sup>. Different filter scales (fine, medium, coarse) can be used to highlight features of different sizes. Features that were calculated after filtration included the histogram maximum, mean, median, minimum, standard deviation, entropy, skewness, and kurtosis.

#### Shape Features

Shape features are those that use only the ROI mask to calculate a quantitative metric that describes the tumor. The shape features used in this study were volume, surface area, surface area density, compactness1, compactness2, convex, convex hull volume, convex hull volume 3D, mass, maximum 3D diameter, mean breadth, number of objects, orientation, roundness, spherical disproportion, and sphericity. These are defined and described in the literature<sup>23,55</sup> as well as in the help documentation of IBEX.

#### Feature Naming Conventions

In subsequent tables and figures, features are named for the abbreviated feature category and then the feature name (e.g. contrast from the co-occurrence matrix is listed as COMcontrast). Features with longer names are abbreviated with the abbreviations listed in Table 3.2 Table 3.2: Features used in this study and their abbreviations used in subsequent tables and figures.(e.g. long run emphasis from the run length matrix is listed as RLMIre).



Table 3.2: Features used in this study and their abbreviations used in subsequent tables and figures.

Feature Category	Full Feature Name	Category Name in IBEX	Feature Name in IBEX	Feature Abbreviation
Gray-level run length matrix	Gray level non-uniformity	GrayLevelRunLengthMatrix25	GrayLevelNonuniformity	RLMglnu
Gray-level run length matrix	High gray level run emphasis	GrayLevelRunLengthMatrix25	HighGrayLevelRunEmpha	RLMhglre
Gray-level run length matrix	Long run emphasis	GrayLevelRunLengthMatrix25	LongRunEmphasis	RLMlre
Gray-level run length matrix	Long run high gray level emphasis	GrayLevelRunLengthMatrix25	LongRunHighGrayLevelEmphasis	RLMlrhgle
Gray-level run length matrix	Long run low gray level emphasis	GrayLevelRunLengthMatrix25	LongRunLowGrayLevelEmphasis	RLMlrlgle
Gray-level run length matrix	Low gray level run emphasis	GrayLevelRunLengthMatrix25	LowGrayLevelRunEmphasis	RLMlgire
Gray-level run length matrix	Run length non-uniformity	GrayLevelRunLengthMatrix25	RunLengthNonuniformity	RLMrlnu
Gray-level run length matrix	Run percentage	GrayLevelRunLengthMatrix25	RunPercentage	RLMrunperc
Gray-level run length matrix	Short run emphasis	GrayLevelRunLengthMatrix25	ShortRunEmphasis	RLMsre
Gray-level run length matrix	Short run high gray level emphasis	GrayLevelRunLengthMatrix25	ShortRunHighGrayLevelEmphasis	RLMsrhgle
Gray-level run length matrix	Short run low gray level emphasis	GrayLevelRunLengthMatrix25	ShortRunLowGrayLevelEmphasis	RLMsrlgle
Neighborhood gray-tone difference matrix	Busyness	NeighborIntensityDifference25	Busyness	NDMbusy

Neighborhood gray-tone difference matrix	Coarseness	NeighborIntensityDifference25	Coarseness	NDMcoarse
Neighborhood gray-tone difference matrix	Complexity	NeighborIntensityDifference25	Complexity	NDMcomp
Neighborhood gray-tone difference matrix	Contrast	NeighborIntensityDifference25	Contrast	NDMcontrast
Neighborhood gray-tone difference matrix	Texture strength	NeighborIntensityDifference25	TextureStrength	NDMtextr
Intensity Histogram	Energy	IntensityDirect	EnergyNorm	HISTenergy
Intensity Histogram	Entropy	IntensityDirect	GlobalEntropy	HISTentropy
Intensity Histogram	Kurtosis	IntensityDirect	Kurtosis	HISTkurt
Intensity Histogram	Maximum	IntensityDirect	GlobalMax	HISTmax
Intensity Histogram	Mean	IntensityDirect	GlobalMean	HISTmean
Intensity Histogram	Median	IntensityDirect	GlobalMedian	HISTmed
Intensity Histogram	Minimum	IntensityDirect	GlobalMin	HISTmin
Intensity Histogram	Skewness	IntensityDirect	Skewness	HISTskew
Intensity Histogram	Standard deviation	IntensityDirect	GlobalStd	HISTstd
Intensity Histogram	Uniformity	IntensityDirect	GlobalUniformity	HISTunif

Intensity Histogram	Variance	IntensityDirect	Variance	HISTvar
Gray-level co-occurrence matrix	Auto-correlation	GrayLevelCooccurrenceMatrix25	AutoCorrelation	COMautocorrel
Gray-level co-occurrence matrix	Cluster prominence	GrayLevelCooccurrenceMatrix25	ClusterProminence	COMclusprom
Gray-level co-occurrence matrix	Cluster shade	GrayLevelCooccurrenceMatrix25	ClusterShade	COMclusshade
Gray-level co-occurrence matrix	Cluster tendency	GrayLevelCooccurrenceMatrix25	ClusterTendency	COMclustend
Gray-level co-occurrence matrix	Contrast	GrayLevelCooccurrenceMatrix25	Contrast	COMcontrast
Gray-level co-occurrence matrix	Correlation	GrayLevelCooccurrenceMatrix25	Correlation	COMcorrel
Gray-level co-occurrence matrix	Difference entropy	GrayLevelCooccurrenceMatrix25	DifferenceEntropy	COMdifent
Gray-level co-occurrence matrix	Dissimilarity	GrayLevelCooccurrenceMatrix25	Dissimilarity	COMdissim
Gray-level co-occurrence matrix	Energy	GrayLevelCooccurrenceMatrix25	EnergyNorm	COMenergy
Gray-level co-occurrence matrix	Entropy	GrayLevelCooccurrenceMatrix25	Entropy	COMentropy

Gray-level co-occurrence matrix	Homogeneity	GrayLevelCooccurrenceMatrix25	Homogeneity	COMhomog
Gray-level co-occurrence matrix	Homogeneity 2	GrayLevelCooccurrenceMatrix25	Homogeneity2	COMhomog2
Gray-level co-occurrence matrix	Information measure correlation 1	GrayLevelCooccurrenceMatrix25	InformationMeasureCorr1	COMinfomc2
Gray-level co-occurrence matrix	Information measure correlation 2	GrayLevelCooccurrenceMatrix25	InformationMeasureCorr2	COMinfomc2
Gray-level co-occurrence matrix	Inverse difference moment norm	GrayLevelCooccurrenceMatrix25	InverseDiffMomentNorm	COMinvdifmn
Gray-level co-occurrence matrix	Inverse difference norm	GrayLevelCooccurrenceMatrix25	InverseDiffNorm	COMinvdifn
Gray-level co-occurrence matrix	Inverse variance	GrayLevelCooccurrenceMatrix25	InverseVariance	COMinvvar
Gray-level co-occurrence matrix	Maximum probability	GrayLevelCooccurrenceMatrix25	MaxProbability	COMmaxprob
Gray-level co-occurrence matrix	Sum average	GrayLevelCooccurrenceMatrix25	SumAverage	COMsumavg
Gray-level co-occurrence matrix	Sum entropy	GrayLevelCooccurrenceMatrix25	SumEntropy	COMsument
Gray-level co-occurrence matrix	Sum variance	GrayLevelCooccurrenceMatrix25	SumVariance	COMsumvar

Gray-level co-occurrence matrix	Variance	GrayLevelCooccurrenceMatrix25	Variance	COMvar
Shape	Compactness 1	Shape	Compactness1	SHAPEcompact
Shape	Compactness 2	Shape	Compactness2	SHAPEcompact2
Shape	Convex	Shape	Convex	SHAPEconv
Shape	Convex hull volume	Shape	ConvexHullVolume	SHAPEconvhull
Shape	Convex hull volume 3D	Shape	ConvexHullVolume3D	SHAPEconvhull3d
Shape	Mass	Shape	Mass	SHAPEmass
Shape	Maximum 3D diameter	Shape	Max3DDiameter	SHAPEmaxdiam
Shape	Mean breadth	Shape	MeanBreadth	SHAPEmeanbre
Shape	Number of objects	Shape	NumberOfObjects	SHAPEnumobj
Shape	Orientation	Shape	Orientation	SHAPEorien
Shape	Roundness	Shape	Roundness	SHAPERound
Shape	Spherical disproportion	Shape	SphericalDisproportion	SHAPEspheredis
Shape	Sphericity	Shape	Sphericity	SHAPEsphericity
Shape	Surface area	Shape	SurfaceArea	SHAPEsurfarea
Shape	Surface area density	Shape	SurfaceAreaDensity	SHAPEsurfareaden
Shape	Volume	Shape	Volume	Volume
Laplacian of Gaussian Fine Filter (size=5, sigma=1)	Entropy	IntensityDirect	GlobalEntropy	LOGFFentropy
Laplacian of Gaussian Fine Filter (size=5, sigma=1)	Mean	IntensityDirect	GlobalMean	LOGFFmean

Laplacian of Gaussian Fine Filter (size=5, sigma=1)	Standard deviation	IntensityDirect	GlobalStd	LOGFFstd
Laplacian of Gaussian Fine Filter (size=5, sigma=1)	Uniformity	IntensityDirect	GlobalUniformity	LOGFFunif
Laplacian of Gaussian Fine Filter (size=5, sigma=1)	Kurtosis	IntensityDirect	Kurtosis	LOGFFkurt
Laplacian of Gaussian Fine Filter (size=5, sigma=1)	Skewness	IntensityDirect	Skewness	LOGFFskew
Laplacian of Gaussian Medium Filter (size=7, sigma=1.5)	Entropy	IntensityDirect	GlobalEntropy	LOGMFentropy
Laplacian of Gaussian Medium Filter (size=7, sigma=1.5)	Mean	IntensityDirect	GlobalMean	LOGMFmean

Laplacian of Gaussian Medium Filter (size=7, sigma=1.5)	Standard deviation	IntensityDirect	GlobalStd	LOGMFstd
Laplacian of Gaussian Medium Filter (size=7, sigma=1.5)	Uniformity	IntensityDirect	GlobalUniformity	LOGMFunif
Laplacian of Gaussian Medium Filter (size=7, sigma=1.5)	Kurtosis	IntensityDirect	Kurtosis	LOGMFkurt
Laplacian of Gaussian Medium Filter (size=7, sigma=1.5)	Skewness	IntensityDirect	Skewness	LOGMFskew

---

## Statistical Software

The analyses throughout this dissertation were performed using the R programming environment<sup>56</sup> and the packages: “survival”<sup>57</sup>, “survcomp”<sup>58</sup>, and “lme4”<sup>59</sup>. Figures were also created using the R programming environment and the packages: “ggplot”<sup>60</sup>, “RColorBrewer”<sup>61</sup>, “ggkm”<sup>62</sup>, and “gridExtra”<sup>63</sup>.

## Chapter 4 : Calculating Radiomics Features

A substantial portion of this chapter is written or based on the following publications:

Fave X, Zhang L, Yang J, Mackin D, Balter P, Gomez D, Followill D, Jones AK, Stingo F, Court LE.

Impact of image preprocessing on the volume dependence and prognostic potential of radiomics features in non-small cell lung cancer. Translational Cancer Research. Doi: 10.21037/tcr.2016.07.11. Volume 5, Issue 4, pages 349-363.(c) AME Publishing Company.

Fave X, Zhang L, Yang J, Mackin D, Balter P, Gomez D, Followill D, Jones AK, Stingo F, Liao Z, Mohan R, and Court L. Delta-radiomics features for the prediction of patient outcomes in non-small cell lung cancer. Scientific Reports. Doi: 10.1038/s41598-017-00665-z. Volume 7. © Nature Publishing Group. Licensed under CC BY 4.0 available at <https://creativecommons.org/licenses/by/4.0/legalcode>.

The permissions for reuse of these materials were obtained from AME Publishing Company © and Nature Publishing Group © respectively.

In this chapter we describe the results for Specific Aim 1: Analyze the effect of image preprocessing on a set of radiomics features measured from CT images and identify the optimal preprocessing for each. Our working hypothesis for this aim was that the image preprocessing technique used would have a significant impact on the prognostic ability and volume independence of radiomics features.

### Introduction

The body of literature suggesting that radiomics features may be prognostic in patients with non-small cell lung cancer (NSCLC) has been steadily growing over the past few years. Although these results are intriguing, the rush to determine whether radiomics features have a useful role in tumor analysis has left many of the fundamental questions surrounding them overlooked or only partially answered. Chief among these is how to determine whether a feature is being calculated correctly. In radiomics, no ground truth exists for the features themselves, and as a consequence most studies have settled for selecting features with high

reproducibility in patient test-retest sets and then using statistical tests or a machine-learning algorithm to determine which features are useful for a particular research question<sup>42,64–67</sup>.

However, this approach can lead to high false-positive rates<sup>68</sup>, and has resulted in variability in both the features that are used and how they are calculated (e.g., feature parameters and image preprocessing).

Further, while a feature should be reproducible, reproducibility itself does not guarantee that a feature is informative. For example, a highly smoothed image is much more likely to return the same value for a feature on a retest, but it may also have lost the original spatial differences that the feature was selected to identify. Also, because most of the quantitative imaging features used in radiomics today were initially developed to analyze aerial photographs<sup>12–14</sup>, in which only two-dimensional rectangular photographic images of the same pixel dimensions were compared, normalization for area or volume differences was not originally necessary. However, in tumor analysis, the regions of interest (ROIs) are the irregular contours of three-dimensional tumors. As a result, the volumes of tumor ROIs have substantial inter-patient variability. A feature that is correlated with volume would be likely to have high reproducibility when tested and retested in a set of patients with a wide range of ROI volumes. This correlation can dominate the useful spatial distribution or intensity information in the feature that we hope to measure. The impact of these volume differences on features measured from computed tomography (CT) images has never been systematically investigated. However, two recent studies have demonstrated the effects of volume on features in fluorodeoxyglucose-positron emission tomography (FDG-PET) images in which the total number of voxels per tumor was much smaller, and concluded that radiomics features offer complementary information only above volume thresholds as large as 10 or 45 cm<sup>3</sup> respectively<sup>69,70</sup>.

Radiomics features have also been shown to demonstrate wide variability when the imaging parameters used for acquisition of CT images are changed<sup>41</sup> or when a different CT

scanner is used with the same imaging protocol<sup>40</sup>. The gray-level intensities in CT images are carefully calibrated to the Hounsfield Unit scale. Thus this variability in the features suggests the features may be susceptible to noise in the image, differences between scanners due to differences in the reconstruction kernel and algorithm, or differences in beam quality. In order for features to be useful, the difference between CT scanners must be negated. One option would be to standardize the imaging parameters, especially the tube current, between machines before acquiring the images so that the results are more similar. However when working with retrospective data it would be more convenient to apply smoothing filters to the image or use a bit depth resample that can reduce the impact of noise in the image and potentially help homogenize images acquired on different scanners. Reducing the image noise would also increase the likelihood that the feature values are representative of the actual tumor heterogeneity and not image noise.

In order to determine how best to calculate features, it is first necessary to determine which features are susceptible to changes in image preprocessing or wide variations in tumor volume. This chapter investigates this issue by assessing changes in the correlations between features and volume, the univariate prognostic potential of features, and the CT model dependence as a function of different image preprocessing techniques.

## **Methods**

### Images

The pre-treatment 4DCT images acquired for treatment planning of the patient cohort described in Chapter 3 were used in this analysis and followed all of the exclusion criteria described in Chapter 3 which left 107 images for analysis. The characteristics for the resulting study population are summarized in Table 4.1.

Table 4.1: Summary of the clinical characteristics of the study population, n=107.

Characteristic	No. (%)
Median age (range)	66 years (47-80 years)
Median gross tumor volume (range)	39.6 cm <sup>3</sup> (5.4-567 cm <sup>3</sup> )
Sex	
Male	62 (58)
Female	45 (42)
Tumor stage	
II	12 (11)
III	93 (87)
IV	2 (2)
Tumor histologic findings	
Squamous cell carcinoma	46 (43)
Adenocarcinoma/other	61 (57)

#### Exploratory Analysis of Image Pre-processing Techniques

Initially, three different types of image pre-processing techniques were considered for this study: resampling the image bit depth, smoothing the images, or both smoothing and resampling the image. As part of an exploratory analysis, different filters under each of these categories were examined for their relative ability to change the feature values. A default filter in each category was chosen for comparison to the other filters of the same category then each of these default filters was compared to using no image preprocessing at all.

Resampling the image bit depth was evaluated because it had been used in previously published works<sup>26,70–72</sup> and because it allows for a simple way to control the bin size of the radiomics matrices (co-occurrence matrix, run-length matrix, and neighborhood difference matrix). For example resampling the bit depth from 12 bit to 8 bit, reduces the number of possible intensity values from  $2^{12} = 4096$  to  $2^8 = 256$ , and thus the pixels have been binned by bins of size  $4096/256 = 16$ . Because several of the radiomics features are calculated from matrices that track how often pixels of one intensity are next to each of the other intensities, this resample removes the need to select an appropriate bin size for these matrices, and

instead bins of 1 can be used when calculating the matrix from an image that has already been bit depth resampled. Continuing with the example, the original 12 bit image would have had a co-occurrence matrix of 4096 by 4096 while an 8 bit image would have a co-occurrence matrix of 256 by 256. The range of values in NSCLC tumors is typically much less than the range of values in the entire image. Thus for a hypothetical tumor with values from only 1 to 100 HU, only a 100 by 100 subsection of the 12 bit co-occurrence matrix would be used to calculate the feature since the rest of the co-occurrence matrix would be filled with zeros. For the 8 bit image, the 1-100 HU range would be resampled to 1-7, and the informative subsection of the co-occurrence matrix would thus be a 7 by 7 matrix. By using resampling, the 7 by 7 co-occurrence matrix would be less likely to be sparsely populated than the 100 by 100 co-occurrence matrix especially if the tumor is small and thus may better represent the spatial patterns in the image and be more informative. The CT images used in this analysis were stored as 12 bit images so images were resampled for bit depths ranging from 11 to 6 bits. Then the values for each feature at each bit depth were compared to the original 12 bit image using a Pearson correlation test. This was done to evaluate if the specific choice of bit depth had an impact on the relative order of the patients.

The second category of filters that was examined in this exploratory analysis was the use of smoothing filters. Smoothing filters are designed to reduce image noise and thus were expected to improve the ability of the radiomics features to measure the tumor heterogeneity. Three different types of smoothing filters were examined: a Gaussian filter of size 3 with a sigma of 0.5, a 3D edge preserving smoothing filter with a kappa of 70, and four Butterworth filters with orders of 2 and cutoff frequencies of 75, 100, 125, and 150 respectively. The Gaussian filter is an isotropic filter that acts in the spatial domain. It has been used in other radiomics studies<sup>43,51,70,73</sup> and works by smoothing evenly across the entire image. However this technique can lead to blurring along edges. The edge preserving smoothing filter also acts in the spatial domain but is an anisotropic diffusion filter which means it can smooth along

different directions when it detects an edge so as not to blur the edge<sup>74</sup>. The parameter, kappa, determines the strength of the edges that it can detect<sup>74</sup>. Larger values of kappa mean the filter is more likely to smooth over an edge than preserve it. The Butterworth smoothing filter acts as a low pass filter to remove high frequency noise. This filter has the advantage of acting in the frequency domain, so it is not limited by the size of the filter matrix. Additionally, Butterworth filters have the benefits of reduced ringing and gradual attenuation of higher frequencies. The larger the value for the order, the steeper the cutoff will be for the Butterworth filter. Research within our own group has shown that using Gaussian filters with a sigma between 0.5 and 1.1 result in a significantly lower value for uniformity in tumors with necrosis versus those without<sup>75</sup>. In the same study it was also determined that using the 3D edge preserving smoothing filter with a  $\kappa \geq 60$  resulted in a significantly lower value for skewness in the tumors with necrosis<sup>75</sup>. A separate analysis found that using a Butterworth filter with a cutoff frequency of 125 reduced the dependence of texture features from CT images on the reconstruction FOV<sup>76</sup>. The Pearson correlation coefficient was calculated between the radiomics features calculated from images preprocessed with each smoothing filter and the features calculated from the images preprocessed with the Butterworth filter that had a cutoff frequency of 125. This was done to determine whether the choice of smoothing filter had a substantial impact on the relative order of the patients.

Then using both smoothing and bit depth resampling before feature calculation was examined to combine the advantages of both of these techniques. For this set, images were preprocessed with each of the examined smoothing filters (Gaussian, edge preserving smoothing, and Butterworth with cutoff frequencies of 75, 100, 125, and 150) and an 8 bit depth resample. The choice of 8 bit was selected because the effect of noise in CT for soft tissue and tumor should be less than 16 HU and because this value had been used in other radiomics analyses<sup>26,70–72</sup>. The Pearson correlation coefficient was calculated between each filter and bit depth combination and the Butterworth filter with a cutoff frequency of 125 and 8 bit depth

resample. The Butterworth filter was chosen as the default filter to be used with the 8 bit depth resample in order to stay consistent with the different filter comparisons. This test was done to determine whether the choice of smoothing filter when combined with an 8 bit depth resample had a substantial impact on the relative order of the patients.

The goal of this exploratory analysis was to evaluate the impact of specific preprocessing parameters on the resulting radiomics feature values with the intention of selecting one image preprocessing technique of each type for full investigation of the features' volume dependence, CT model dependence, and prognostic potential. The Pearson correlation coefficient was used to compare different combinations of preprocessed features to the default versions selected in each group. The Pearson correlation coefficient was used because if feature values changed in scale but not rank (e.g. patients with high contrast still had high contrast but all of the values are 1/10 of the values without image preprocessing) then their impact on eventual prognostic models would remain the same. The different bit depth resamples were compared to using no bit depth resample. The different smoothing filters were compared to the Butterworth smoothing filter with a cutoff frequency of 125. The different smoothing and bit depth combinations were compared to the combination of a Butterworth filter with a cutoff frequency of 125 and 8 bit depth resample. Finally an 8 bit depth resample, Butterworth filter with a cutoff frequency of 125, and combination of Butterworth filter with a cutoff frequency of 125 and an 8 bit depth resample were compared to using no extra image preprocessing to evaluate whether these different combinations yielded radiomics features that were different from their original non-preprocessed versions and thus potentially more informative.

A lower intensity threshold of -100 and upper intensity threshold of 200 Hounsfield Units were applied to all images before feature calculation to ensure that no voxels of the surrounding normal lung tissue or bone were included in the GTV. These thresholds have been used in previous studies of radiomics features extracted from CT images of lung cancer<sup>26,43</sup>.

This threshold was performed after smoothing but before a bit depth resample. The threshold was performed after smoothing because the radiomics software IBEX applies the smoothing over the entire bounding box and not just the pixels within the contour. Thus if the image had been thresholded and then smoothed, the smoothing would blur pixels with values of 0 into the contour GTV and artificially decrease the values. Thus images were either (i) thresholded and then resampled, (ii) smoothed and then thresholded, or (iii) smoothed, thresholded, and then resampled.

### Image Pre-processing

After the exploratory analysis, four of the examined image preprocessing techniques were selected for in-depth analysis of their impact on the radiomics features. As in the exploratory analysis, a lower intensity threshold of -100 and upper intensity threshold of 200 Hounsfield Units were applied to all images before feature calculation to ensure that no voxels of the surrounding normal lung tissue or bone were included in the GTV. Images were then further pre-processed with either (i) 8 bit depth resampling, (ii) a Butterworth smoothing filter with an order of 2 and a cutoff frequency of 125, (iii) both Butterworth smoothing (order of 2 and cutoff frequency of 125) and 8 bit depth resampling, or (iv) no additional pre-processing. When both Butterworth smoothing and 8 bit depth resample were used, the Butterworth smoothing was performed first. The radiomics features were calculated from the tumor ROIs after each of these pre-processing techniques had been applied. The general effect of each of these techniques should be to reduce noise in the image and thus improve the overall signal to noise ratio of any radiomics feature. These particular preprocessing options were selected based on the results of the exploratory analysis. Figure 4.1 shows the visible result of each of these pre-processing techniques on a sample tumor ROI. The impact of these four techniques on the co-occurrence matrix for the tumor in Figure 4.1 is illustrated in Figure 4.2.

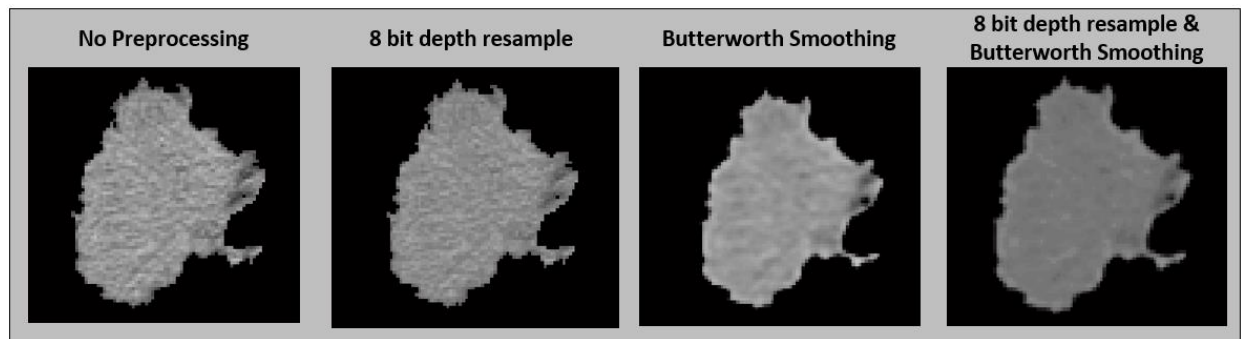


Figure 4.1: Sample image of a patient tumor ROI using each preprocessing technique.

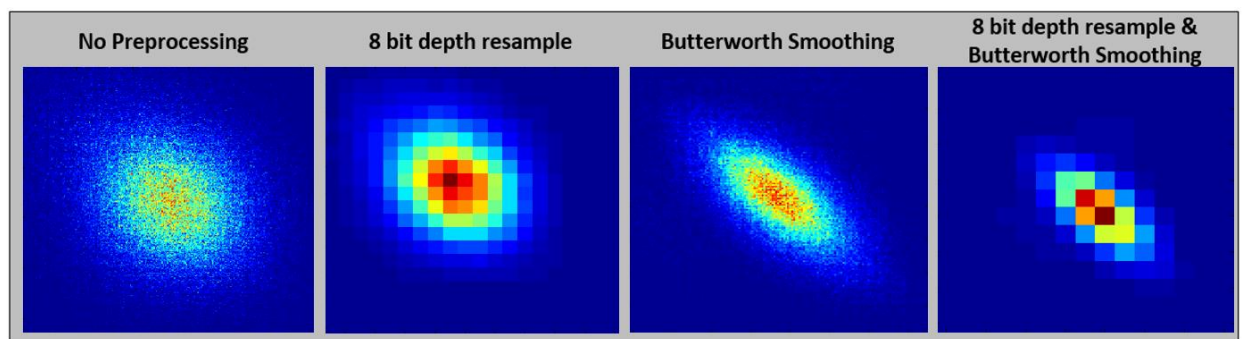


Figure 4.2: Visualization of the co-occurrence matrix for the same tumor ROI after four different image preprocessing techniques. The axes in these figures are the ranges of intensity values. Thus for the No Preprocessing and Butterworth Smoothing co-occurrence matrices, the values range from 1 to 4096 for both the x and y axis as these are 12 bit images. For the 8 bit depth resample and 8 bit depth resample and Butterworth Smoothing images the axes range from 1 to 256. Additionally all four of the images are zoomed in on the useful part of the matrix since in all 4 cases, most of the co-occurrence matrix is filled with zeros due to the image threshold used. The dark blue represents values of zero in the matrix while dark red is scaled to the highest frequency in the matrix. Note that using Butterworth smoothing tends to make the co-occurrence matrix distribution more linear.

## Features

For both the exploratory analysis and in-depth analysis, 45 radiomics features from the histogram, co-occurrence matrix, neighborhood difference matrix, and run-length matrix were calculated and are summarized in Table 4.2. The histogram features summarize characteristics of the intensity distribution for each tumor. The co-occurrence matrix, neighborhood difference matrix, and run-length matrix features all contain information about the spatial distribution of the pixel intensities within a tumor. Full descriptions of these feature categories is available in Chapter 3. Also, all features were calculated using the open-source Imaging Biomarker Explorer (IBEX) software<sup>50</sup> as discussed in Chapter 3.

Table 4.2: The radiomics features used in the analysis of image preprocessing techniques.

Histogram	Co-occurrence Matrix	Run Length Matrix	Neighborhood Difference Matrix
Energy	Autocorrelation	Gray-level	Busyness
Entropy	Cluster	non-	Coarseness
Kurtosis	prominence	uniformity	Complexity
Maximum	Cluster shade	High gray-level	Contrast
Mean	Cluster	run emphasis	Texture Strength
Median	tendency	Long run	
Minimum	Contrast	emphasis	
Skewness	Correlation	Long run high	
Standard deviation	Difference	gray level	
Uniformity	entropy	emphasis	
Variance	Dissimilarity	Long run low	
	Energy	gray level	
	Entropy	emphasis	
	Homogeneity	Low gray-level	
	Homogeneity 2	run emphasis	
	Information	Run	
	measure	percentage	
	correlation	Run-length	
	Information	non-	
	measure	uniformity	
	correlation 2	Short run	
	Inverse	emphasis	
	difference	Short run low	
	moment norm	gray-level	
	Inverse	emphasis	
	difference	Short run high	
	norm	gray-level	
	Inverse	emphasis	
	variance		
	Max probability		
	Sum average		
	Sum entropy		
	Sum variance		
	Variance		

## Volume-Dependence

To determine whether the features became more or less correlated with volume as a result of image pre-processing, we used the spearman rank correlation coefficient ( $r_s$ ) to calculate the correlation with volume of each feature after each preprocessing technique. The Spearman rank correlation coefficient ranges from -1 to 1 and evaluates whether a value decreases or increases monotonically; 1 and -1 represent a perfect correlation and 0 represents no correlation.

Because the feature algorithms used for tumor analysis in current radiomics studies were originally designed for comparing equally sized photographs<sup>12</sup>, it was possible that some algorithms might be inherently dependent on volume and may require correction for the number of voxels in the image. Features with extremely high values ( $r_s > 0.95$ ) for all four preprocessing techniques were identified and new normalized versions of the algorithms for these features were created and added to the feature set for analysis. The Spearman correlation coefficient for these normalized features were then calculated for each preprocessing technique. For completeness, we did not remove the features that exhibited the extremely strong correlations with volume from the remainder of the analysis in this chapter.

To evaluate the corrected versions of the volume-dependent features, two datasets of spherical digital phantoms were created using MATLAB (The Mathworks, Inc., Natick, MA) and saved as DICOM images. Digital phantom images had an in-plane resolution of 0.98 mm and slice thickness of 2.5 mm. Axial images of the digital phantoms can be seen in Figure 4.3. The first phantom dataset represented a texture of pure noise. Three sets of five spheres (12, 90, 175, 320, 445 cm<sup>3</sup>) were used. Each set of five spheres was filled with intensity values from a Gaussian distribution with a mean of 1025 and standard deviation of 25, 50, or 75 respectively. This dataset was used to check that the volume-corrected features did not change with increasing volume and more importantly confirm that the corrected feature values increased or decreased in the expected order as the standard deviation in the digital phantoms increased.

For example, when measuring busyness, the values from the spheres with the larger standard deviation should have higher values for busyness. The second phantom dataset represented a defined texture of pure signal (i.e., no noise). Three sets of five spheres were used again but were filled with a checkerboard pattern. Each set had checkerboard cubes with a side length of 2, 5, or 10 voxels where alternating cubes had an intensity of either 1 or 2. Thus, when viewed in either the axial, sagittal, or coronal plane, a checkerboard pattern was apparent. This second pattern was used to confirm that the corrected features were independent from volume and able to correctly order the phantoms with increasing checkerboard square size.

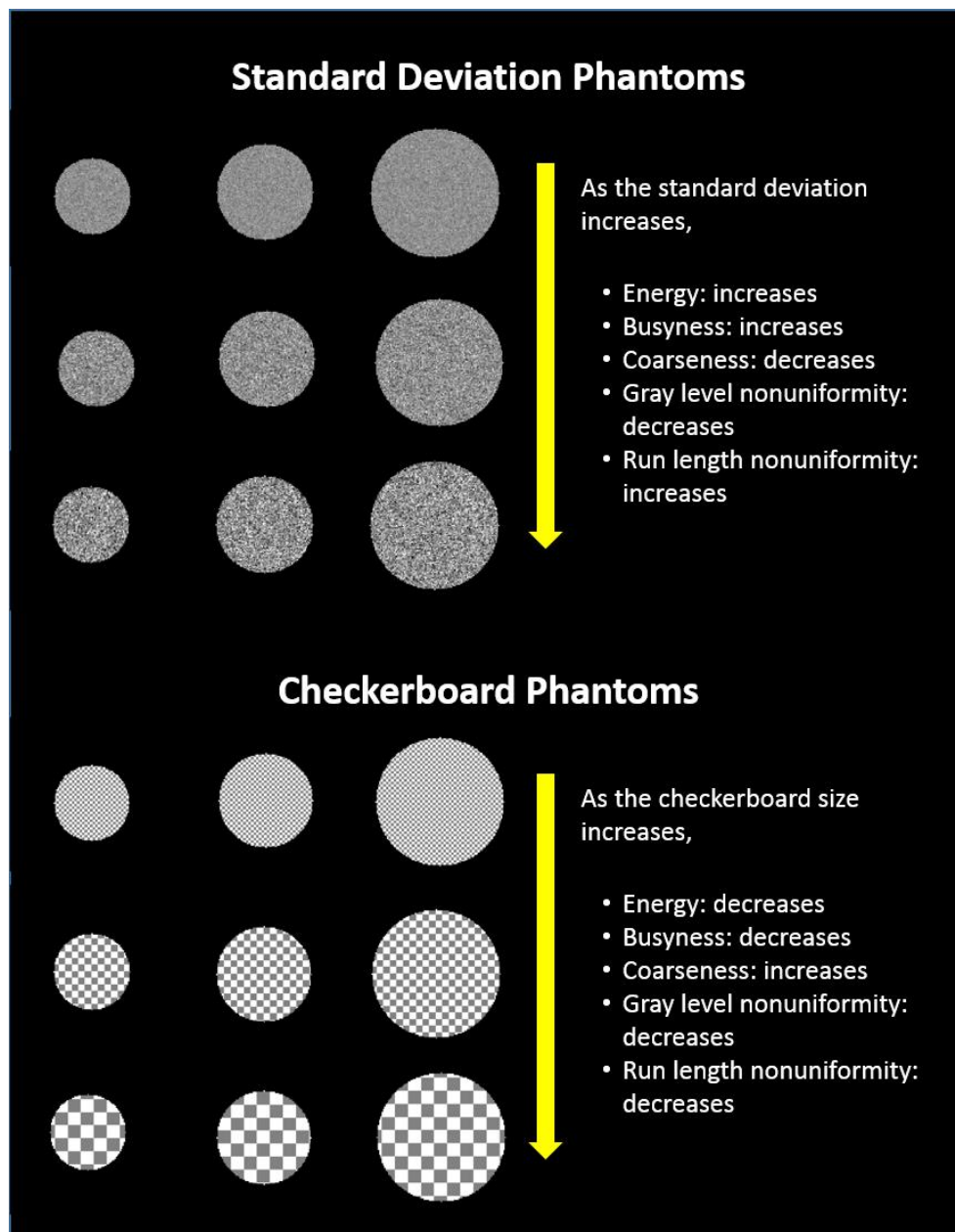


Figure 4.3: Axial views of the 3D spherical digital phantoms created for this study. Two patterns were used: (i) spheres filled with Gaussian noise with increasing standard deviation and (ii) spheres filled with a checkerboard pattern with increasing checkerboard square sizes. For each pattern five spheres of increasing radii were created. These phantoms could then be used to check any volume dependencies were removed when corrected feature algorithms were used and that the new form of the feature was informative (e.g. feature values increased or decreased as the pattern was changed in the expected order).

### CT Model Dependence

To determine whether the feature values significantly varied with the CT model used to acquire the images, a Wilcoxon rank sum test between two scanners was performed for each feature and pre-processing combination. The overwhelming majority of images had been acquired on either a GE Discovery ST or a GE Lightspeed RT16 (755 of the 785 images). The remaining images were acquired on either a Philips Brilliance 64 (n=21) or Philips Brilliance Big Bore (n=9). This test was performed because CT scanner model has been demonstrated to be an important factor in feature reproducibility<sup>40</sup>. A patient subset that had images available from the first week of treatment (n=81) was used for this test because at this time point the patients were roughly split between the two CT scanners (37 patients imaged with the GE Discovery ST and 44 patients imaged with the GE Lightspeed RT16) and their tumors were not expected to have already shown any therapy-induced changes. The Wilcoxon rank sum test evaluates the null hypothesis that two populations have the same distribution of values. P-values were corrected for multiplicity testing using the Benjamini-Hochberg method<sup>77</sup> and were considered significant if the value was less than 0.05 after correction.

### Prognostic Potential

To determine the impact of preprocessing on the usefulness of radiomics features, univariate Cox proportional hazards models were fitted for overall survival. P-values for the fit using the likelihood-ratio test were calculated for each model. P-values were corrected using the Benjamini-Hochberg process to control for the false discovery rate (type 1 error). Corrected p-values<0.05 were considered significant.

Each univariate model was then recalculated using leave-one-out cross validation to generate predictions for each patient in each model. In this framework, each patient's prediction is calculated using a model in which that patient was left out of the coefficient fitting process. Harrell's concordance index (c-index) was then calculated using the predicted risks. The c-index is similar to the area under the curve and evaluates, for each combination of two

patient predictions, how often the patient with the higher predicted risk actually experiences the event (death) before the patient with the lower predicted risk. A c-index value of  $\leq 0.5$  indicates that the model does not perform better than random chance and a value of 1.0 indicates a perfect model. The c-index for a model with volume as the only covariate was also calculated for comparison.

Lastly, to determine whether features actually outperformed volume alone, the log-likelihood ratios between Cox proportional hazards models for overall survival fitted with volume only and models fitted with volume and one radiomics feature at a time were calculated. The p-values for the log-likelihood ratios were then determined and corrected using the Benjamini-Hochberg process. A p-value  $< 0.05$  would mean that including that particular feature significantly improved the model's fit to the data when compared with the fit of a model using volume only.

## Results

### Exploratory Analysis of Image Preprocessing Techniques

The goal of the exploratory analysis was to determine if the specific parameters for three types of filters (bit depth resampling, smoothing, and smoothing with bit depth resampling) played a large role in the impact of that filter on the relative radiomics feature values. Results for the Pearson correlation coefficients comparing the different bit depth resampling levels to no bit depth resampling, and thus images with 12 bit depth, are in Figure 4.4. When the bit depth resampling level was decreased, the strength of the Pearson correlation coefficient decreased for approximately half of the features (21 of the 46). Regardless of the specific bit depth level used, most features remained weakly correlated ( $r \geq 0.6$ ) with their values measured from the default images. The features that had the largest decreases in their correlation strength as bit depth decreased were short run high gray level emphasis, long run length gray level emphasis, and gray level non-uniformity from the run-length matrix, and inverse variance and information measure correlation 1 from the co-

occurrence matrix. Of the features whose correlation decreased below 0.9 at one or more bit depths: 2 demonstrated a decrease at 11 bit depth, 4 at 10 bit depth, 3 at 9 bit depth, 3 at 9 bit depth, 2 at 8 bit depths, 4 at 7 bit depths, and 10 at only 6 bit depths.

Results for the Pearson correlation coefficients comparing the different smoothing filters are in Figure 4.5. In this analysis the Pearson correlation coefficient was calculated between the radiomics feature values after each filter and the feature values obtained after the image was preprocessed with a Butterworth filter with an order of 2 and a cutoff frequency of 125. Changing the cutoff frequency to 150 or 100 resulted in features that were all still highly correlated ( $r \geq 0.9$ ) to the default Butterworth filter. Reducing the cutoff frequency to 75 also resulted in the majority of features still being highly correlated to the default values but 14 of the correlations decreased to be between 0.8 and 0.89 and 2 of the features had correlations between 0.7 and 0.79. Using a Gaussian filter instead of a Butterworth filter resulted in very highly correlated features as well with only one feature (information measure correlation 1 from the co-occurrence matrix) having the lowest value of 0.77 but most features were above 0.9. Similarly for the edge preserve smoothing filter, only 6 features did not have a correlation higher than 0.9. The lowest correlation was still a moderate value of 0.62. Thus in almost all cases the specific choice of smoothing filter does not strongly affect the relative feature values.

Results for the Pearson correlation coefficients comparing the combination of different smoothing filters with an 8 bit depth resample to using a Butterworth filter with an order of 2 and a cutoff frequency of 125 with an 8 bit depth resample are in Figure 4.6. As with the comparison of the different smoothing filters, the overwhelming majority of features were highly correlated when the only change was the Butterworth frequency cutoff to 100 or 150. When the cutoff frequency was decreased to 75, some of the features did begin to show a large change, with 3 (gray level nonuniformity from the run-length matrix, complexity from the neighborhood difference matrix, and inverse variance from the co-occurrence matrix) having a coefficient less than 0.7. The feature that showed the most change was gray level non-uniformity from the run-

length matrix. Using a Gaussian filter instead also resulted in highly correlated features with the exception of gray level non-uniformity from the run-length matrix and inverse variance from the co-occurrence matrix. When features were calculated with an edge preserve smoothing filter, the correlation coefficients decreased below 0.9 for 19 features but the majority were still between 0.7 and 0.9 with only 2 features (inverse variance and correlation from the co-occurrence matrix) having coefficients below 0.7. There was a lot of overlap between the features whose correlation coefficients decreased with a cutoff frequency of 75 and those that decreased when the edge preserve smoothing filter was used.

One of the filters of each type (8 bit depth resample, Butterworth smoothing with a cutoff frequency of 125, and both Butterworth smoothing with a cutoff frequency of 125 and an 8 bit depth resample) was then compared to using no image preprocessing and the results are shown in Figure 4.7. A few features were always highly correlated regardless of which image preprocessing was used; these included long gray level run emphasis, high gray level run emphasis, and gray level non-uniformity from the run-length matrix, the mean and median from the histogram, and 8 features from the co-occurrence matrix. For the remainder of the features, if the correlation decreased with bit depth resampling or smoothing then it also decreased below 0.9 when both bit depth and smoothing were used.

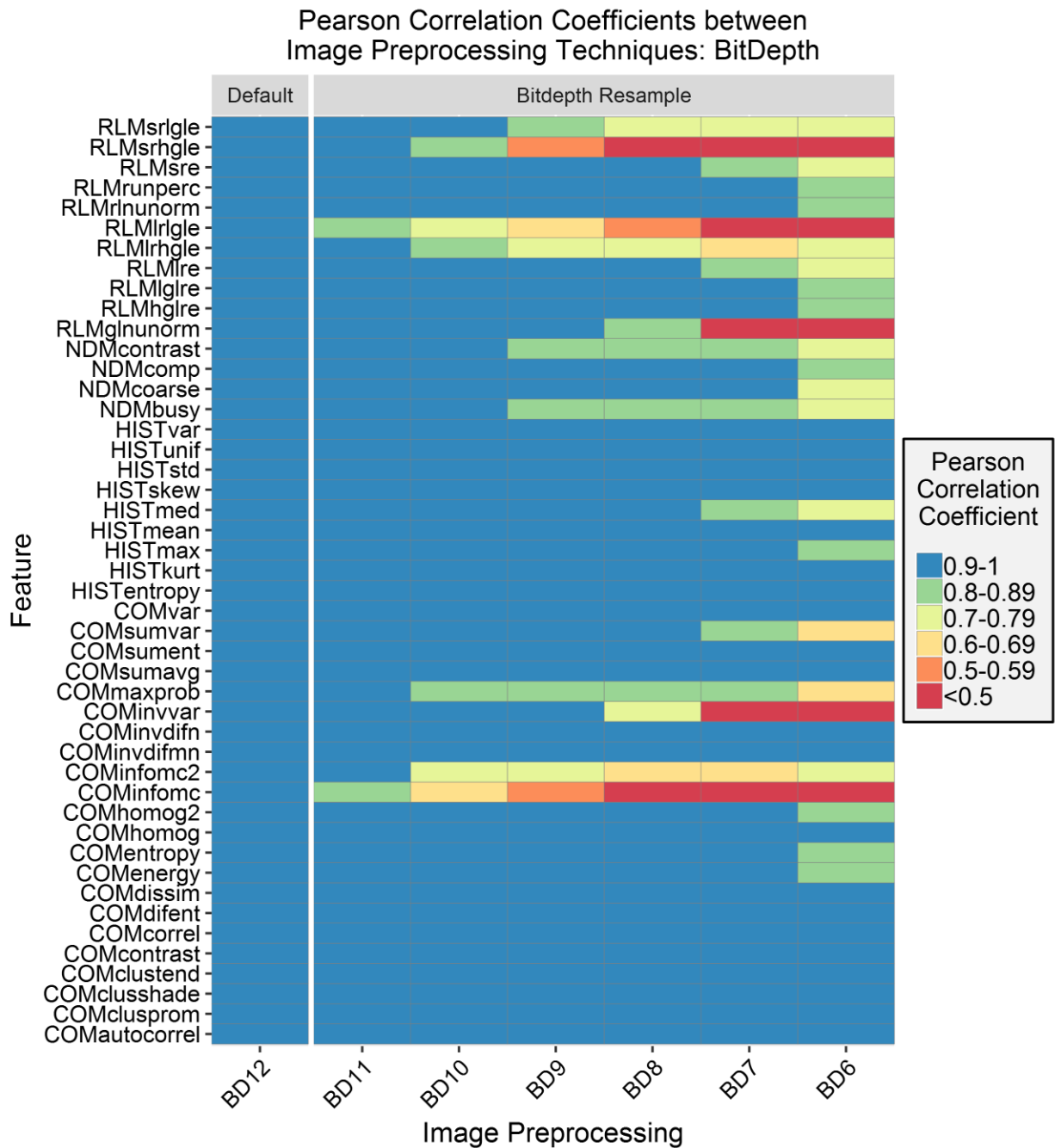


Figure 4.4: Correlation between features calculated after bit depth resampling between 11 and 6 bits and their default values without any bit depth resampling (BD12). Approximately half of the features (21/46) remained highly correlated to their original values even as the bit depth was resampled to 6 bits. For the remaining features, their correlation decreased as the bit depth was decreased with two features showing changes even with only a minor resampling to

11 bit depths and 10 features remaining correlated until a major resampling to 6 bit depths.

Abbreviations: BDXX=Bit depth resampling to XX, (e.g. BD8= 8 bit depth resample).

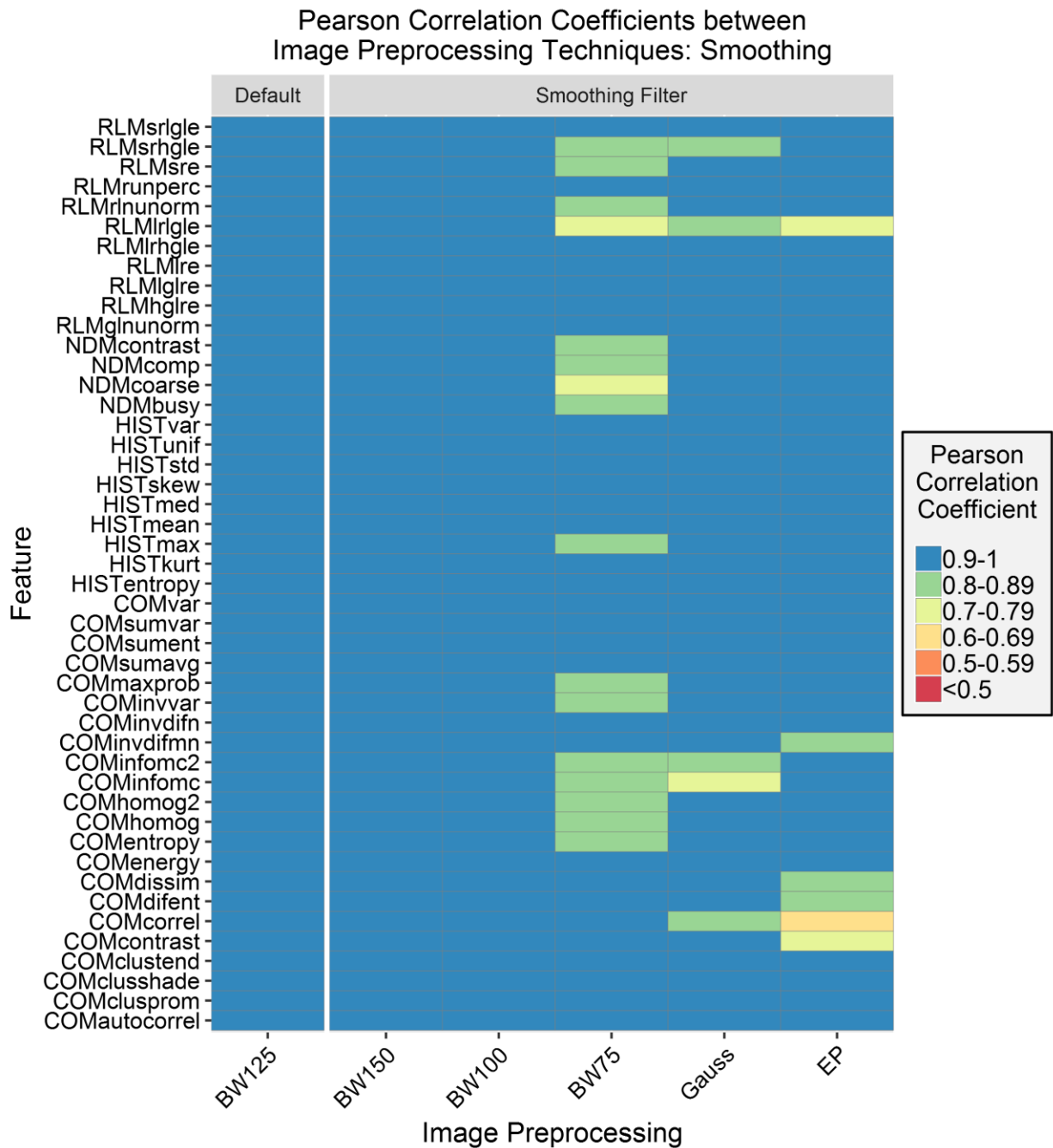


Figure 4.5: Correlation between features calculated after different smoothing filters to their default values from images smoothed with a Butterworth filter with a cutoff of 125. The majority of radiomics features were highly correlated ( $r > 0.9$ ) to their values after a Butterworth filter with a cutoff of 125 was used to when they were calculated with different smoothing filters for comparison. Thus the choice of cutoff or specific smoothing filter does not play a substantial

role in the feature values. Abbreviations: BW125=Butterworth filter with a cutoff value of 125, Gauss=Gaussian filter with size 3 and sigma of 0.5, EP=Edge preserve smoothing filter with a kappa of 70.

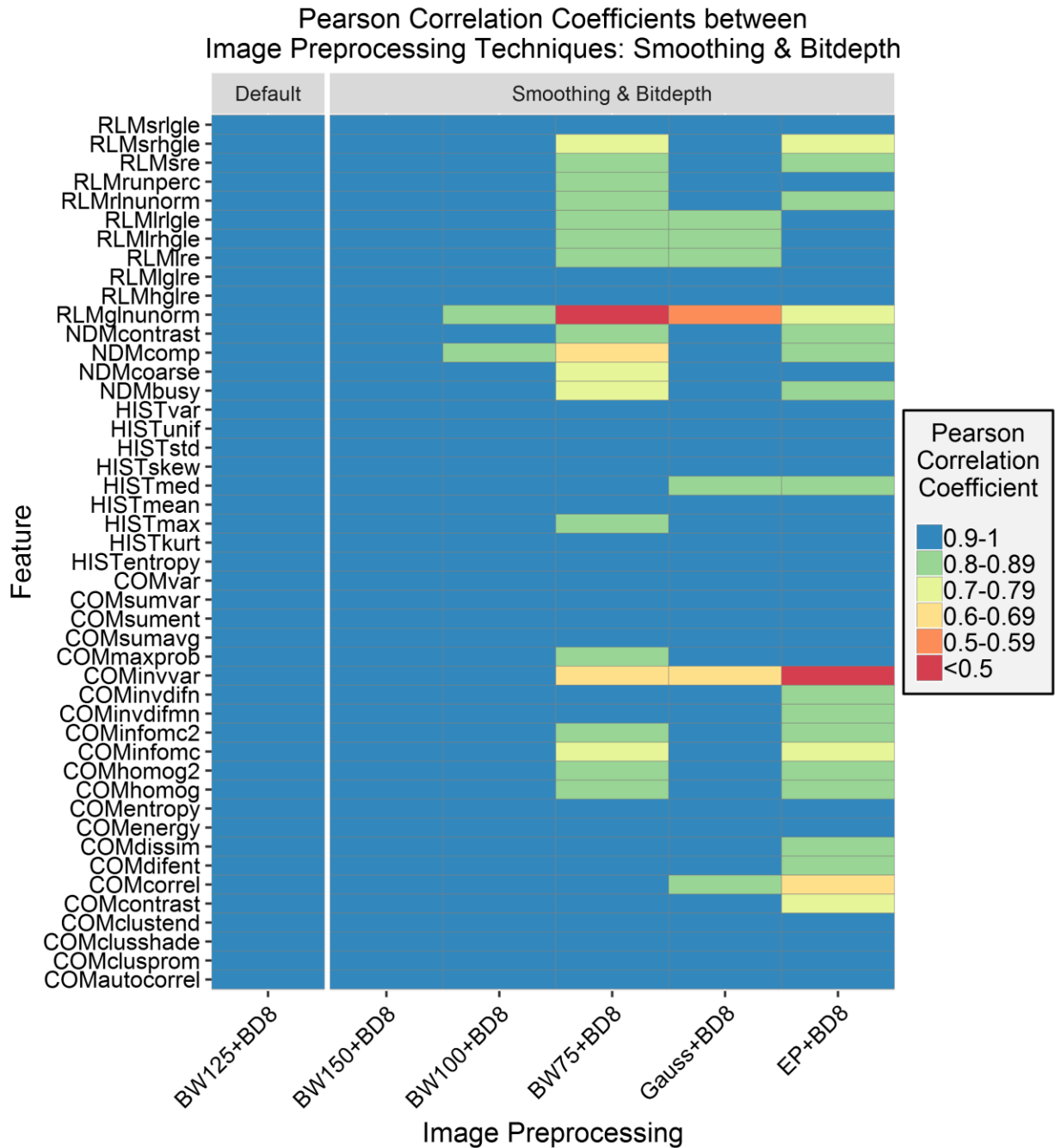


Figure 4.6: Correlation between features calculated after Butterworth smoothing filters with different cutoffs (75,100,150) and an 8 bit depth resample to their default values from images smoothed with a Butterworth filter with a cutoff of 125 and an 8 bit depth resample. The majority of radiomics features were highly correlated ( $r>0.9$ ) to their default values when calculated with different frequency cutoffs. Even with a much lower frequency cutoff of 75, the

overwhelming majority of features had Pearson correlation coefficients  $>0.8$ . Thus the choice a specific cutoff frequency does not play a substantial role in the relative patients' feature values. Abbreviations: BWXX+BD8=Images filtered with a Butterworth filter with a cutoff value of XX and an 8 bit depth resample, (e.g. BW125+BD8, is a Butterworth filter with a cutoff frequency of 125 and 8 bit depth resample).

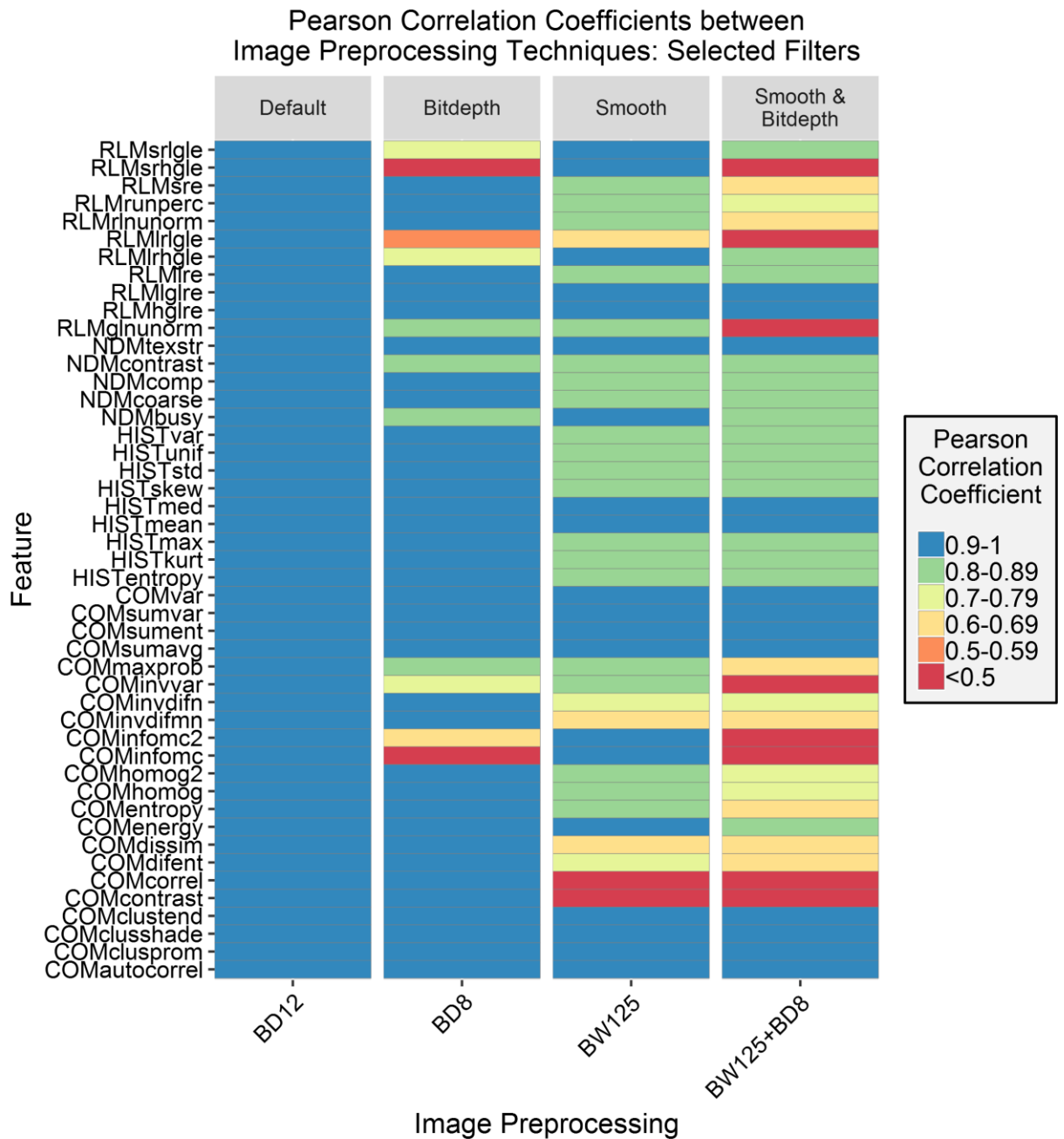


Figure 4.7: Correlation between features calculated after three filters to their default values from images without any image preprocessing (BD12). Features were calculated after either an 8 bit depth resample (BD8), smoothing with a Butterworth filter that had a cutoff frequency of 125 (BW124), or both smoothing with a Butterworth filter with a cutoff frequency of 125 and an 8 bit depth resample (BW125+BD8). Only 13 of the radiomics features were highly correlated

( $r > 0.9$ ) regardless of the image processing. The combination of smoothing and resampling was most likely to substantially decrease the correlation.

### Volume Dependence

The absolute values of the Spearman correlation coefficients for each feature after each tested preprocessing technique are plotted in Figure 4.8. In general, features were more correlated with volume after either Butterworth smoothing or both Butterworth smoothing and bit depth resampling, and were less correlated with volume after bit depth resampling, Figure 4.9. A few features demonstrated strong ( $r_s > 0.85$ ) correlations with volume for only one or two pre-processing techniques. Both information measure correlation and information measure correlation 2 from the co-occurrence matrix had high correlations with volume after no preprocessing or Butterworth smoothing ( $> 0.90$ ), but were not correlated with volume when bit depth resampling was used, either alone or with Butterworth smoothing ( $r_s < 0.5$ ). Inverse difference moment norm from the co-occurrence matrix had a correlation of 0.88 with volume after Butterworth smoothing or after Butterworth smoothing and bit depth resampling. Texture strength from the neighborhood difference matrix had correlation coefficients of -0.94, -0.87, -0.90, for Butterworth smoothing, bit depth resampling, and both Butterworth smoothing and bit depth resampling respectively, but when no pre-processing was used the coefficient was less than 0.5.

Five features demonstrated a very strong volume correlation, with Spearman correlation coefficients absolute values  $> 0.95$  regardless of which preprocessing technique was used. These features included energy from the histogram, coarseness and busyness from the neighborhood difference matrix, and grey-level non-uniformity and run-length non-uniformity from the run-length matrix. After close investigation, we found that these high levels of volume correlation were due to the feature algorithms, which did not normalize for the number of voxels or matrix elements summed. This means that if these features were measured from two ROIs of different sizes but with pixels of only one intensity, two different values would be obtained. Factors that mitigate this source of volume dependence were introduced for each of these feature algorithms (Table 4.3), and new Spearman correlation coefficients were calculated. The

algorithms for energy, grey-level non-uniformity, and run-length non-uniformity were edited by dividing their values by the total number of voxels in the ROI. The algorithms for busyness and coarseness were changed by normalizing the sums of the average difference around each intensity (the neighborhood difference matrix values,  $s(i)$ ) by the number of voxels of that intensity. The maximum of the Spearman correlation coefficients for the normalized features was 0.79, Figure 4.8 . All of the normalized features preprocessed with resampling had correlations  $<0.5$ .

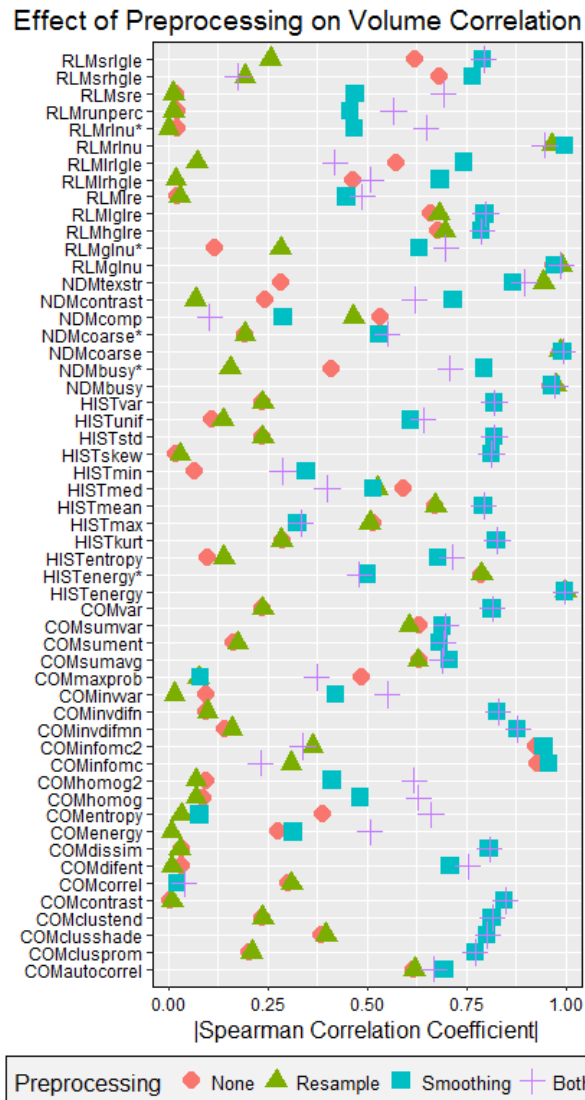


Figure 4.8: The volume correlation for the features measured using the spearman rank correlation coefficient. The absolute value of the coefficients for each feature and preprocessing technique are plotted here. The volume correlation for most of the features was substantially changed with different image pre-processing. Five features were extremely correlated with volume regardless of the preprocessing technique used and were recalculated with a normalizing factor. The normalized version of these algorithms are noted with an asterisk.

Table 4.3: The original and normalized algorithms for the volume-dependent features. The original algorithms for the volume-dependent features from the literature were changed by introducing a normalization term for the number of voxels of each intensity,  $i$ , in the image,  $N(i)$ , or the total number of voxels  $N_v$ . Other terms are:  $p_i$ -probability of intensity  $i$  in the image;  $s(i)$ -sum of the average difference value around voxels of intensity  $i$ ;  $G_h$ -Highest gray-level intensity;  $N_g$ -Number of gray levels;  $N_r$ -Number of run levels;  $p(i,j)$ - probability of gray-level  $i$  having a run of length  $j$ ;  $X(i)$ -the intensity of the  $i^{th}$  voxel in the image,  $X$ .

Feature	Original Algorithm	Normalized Algorithm
Busyness <sup>14</sup>	$\frac{[\sum_{i=0}^{G_h} p_i s(i)]}{[\sum_{i=0}^{G_h} \sum_{j=0}^{G_h} i p_i - j p_j]} \quad (1)$	$\frac{[\sum_{i=0}^{G_h} p_i \frac{s(i)}{N(i)}]}{[\sum_{i=0}^{G_h} \sum_{j=0}^{G_h} i p_i - j p_j]} \quad (2)$
Coarseness <sup>14</sup>	$[\epsilon + \sum_{i=0}^{G_h} p_i s(i)]^{-1} \quad (3)$	$[\epsilon + \sum_{i=0}^{G_h} p_i \frac{s(i)}{N(i)}]^{-1} \quad (4)$
Gray-level non-uniformity <sup>13</sup>	$\frac{\sum_{i=1}^{N_g} \left( \sum_{j=1}^{N_r} p(i,j) \right)^2}{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} p(i,j)} \quad (5)$	$\frac{1}{N_v} \frac{\sum_{i=1}^{N_g} \left( \sum_{j=1}^{N_r} p(i,j) \right)^2}{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} p(i,j)} \quad (6)$
Run-length non-uniformity <sup>13</sup>	$\frac{\sum_{j=1}^{N_r} \left( \sum_{i=1}^{N_g} p(i,j) \right)^2}{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} p(i,j)} \quad (7)$	$\frac{1}{N_v} \frac{\sum_{j=1}^{N_r} \left( \sum_{i=1}^{N_g} p(i,j) \right)^2}{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} p(i,j)} \quad (8)$
Energy <sup>23</sup>	$\sum_i^{N_v} X(i)^2 \quad (9)$	$\frac{1}{N_v} \sum_i^{N_v} X(i)^2 \quad (10)$

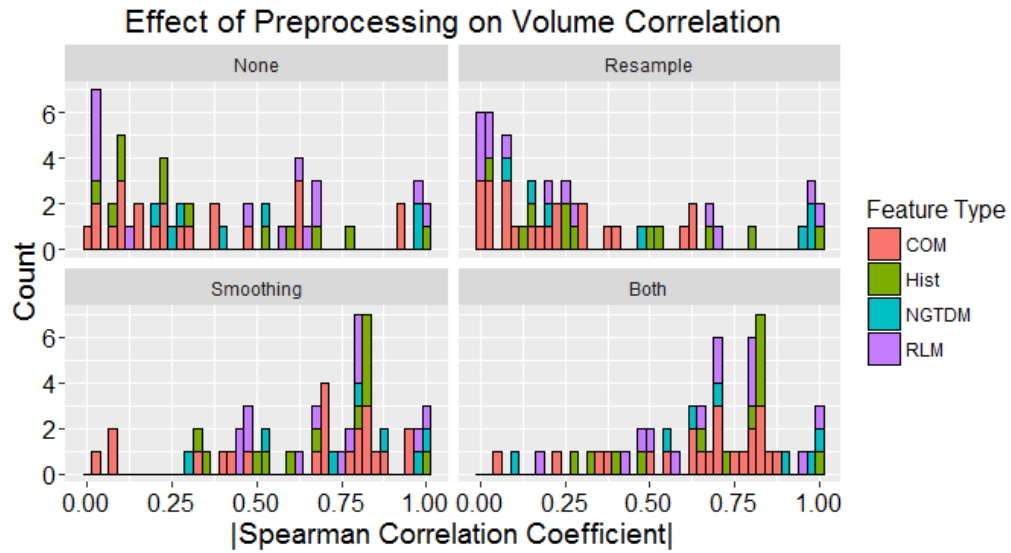


Figure 4.9: The absolute value of Spearman correlation coefficients plotted as a histogram for each preprocessing technique. Butterworth smoothing either alone or combined with bit depth resample increased the strength of the correlation between most of the features and volume. When only bit depth resample was used, the overall volume correlation decreased for the features.

To evaluate whether the corrected feature equations were still informative, the features were calculated from a series of digital phantoms. Each of the corrected features was able to correctly order both of the digital phantom patterns (Gaussian noise with increasing standard deviation and a checkerboard pattern with increasing size of the checkerboard squares). The variation between the values for the three Gaussian noise patterns was substantially less than for the three checkerboard patterns. The values calculated from these phantoms for both the original and corrected version of the five volume dependent features are plotted in Figure 4.10.

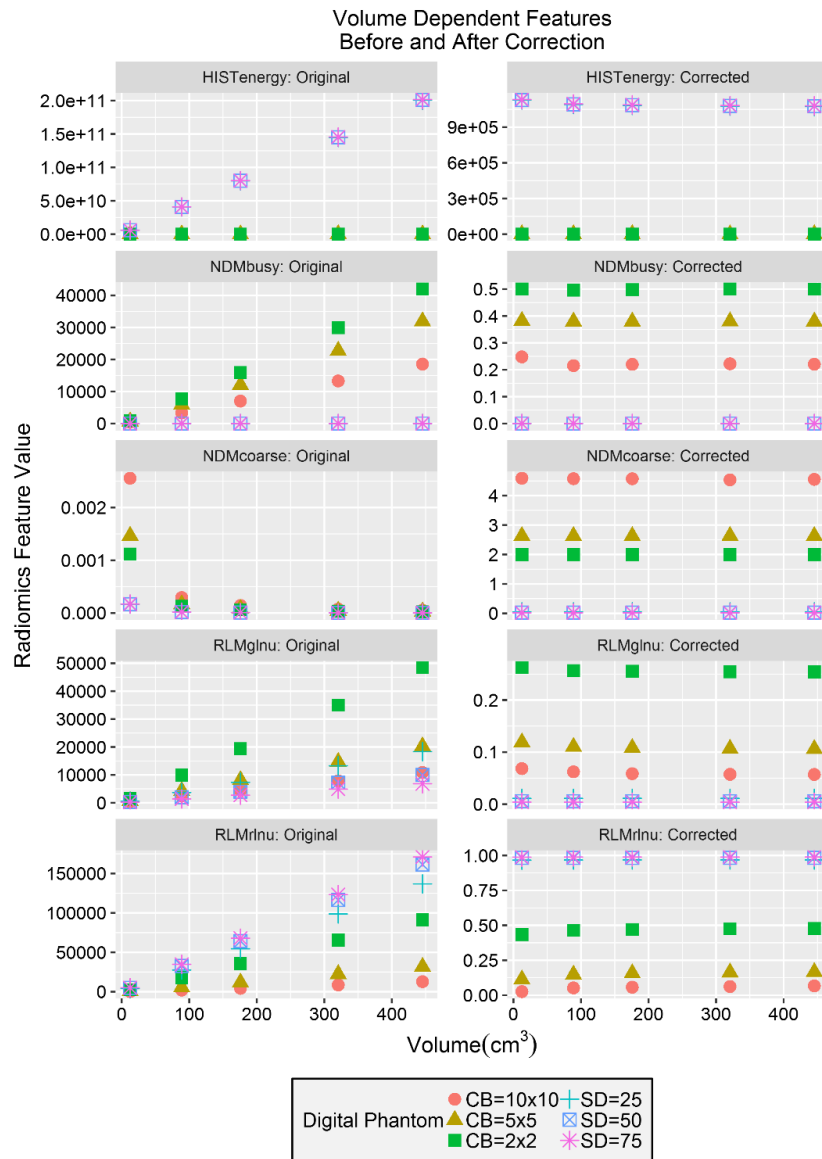


Figure 4.10: Radiomics feature values measured from digital phantoms before and after volume normalization of the feature algorithms. Digital phantoms were filled with either a checkerboard (CB) pattern with squares of 10x10, 5x5, or 2x2 voxels or by Gaussian noise with a standard deviation (SD) of 25, 50, or 75 units. Prior to volume normalization, the values from both phantoms demonstrated a strong volume dependence as seen in the plots on the left side of the figure. After normalization, the values measured from the digital phantoms were independent of volume. Additionally, the feature values were correctly ordered for the digital phantoms.

### CT Model Dependence

The results of the Wilcoxon rank sum test to determine if features were significantly associated with the CT model used to acquire the images before and after image preprocessing are shown in Figure 4.11. No features were always significantly associated with the CT model used to acquire the images regardless of which image preprocessing was used. 24 features were never significantly associated with CT model regardless of which image preprocessing was used. These features included 7 features from the histogram, 8 from the co-occurrence matrix, 4 from the neighborhood difference matrix, and 5 from the run-length matrix. For the remaining 31 features, 2 were not significantly associated with CT Model when they were calculated after bit depth resampling was used, 28 when smoothing was used, 27 when smoothing and bit depth resampling were used together, and 6 when no preprocessing was used. So in general smoothing used either alone or in conjunction with bit depth resampling reduced the likelihood that a feature would be dependent on CT model.

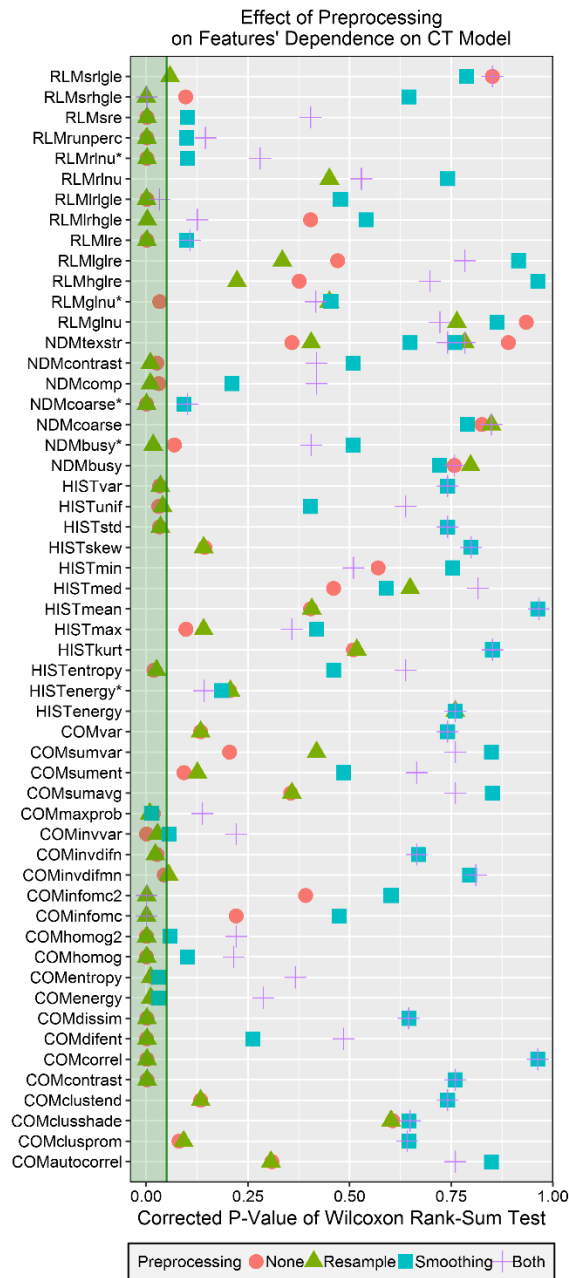


Figure 4.11: The Benjamini-hochberg corrected p-values for the Wilcoxon rank-sum test comparing values measured on two different CT models. Approximately half of the features were always independent of the CT model used to acquire the image. The remaining features were most likely to not be significantly different between CT models when smoothing or smoothing with bit depth resampling was used. The green area highlights features whose p-value was  $< .05$  and thus significant.

## Prognostic Potential

The p-values for the Cox proportional hazard models are plotted in Figure 4.12 for each feature and pre-processing combination, as well as for volume. Almost every feature (39/55) had at least one preprocessing technique that resulted in statistically significant stratification (p-value < 0.05 after Benjamini-Hochberg correction). A few features from each category were never significant in this univariate analysis: normalized busyness, the original (volume-dependent) busyness, complexity, and contrast from the neighborhood difference matrix; maximum, minimum, and the original energy from the histogram; long-run emphasis, run percentage, and the original forms of grey-level non-uniformity and run-length non-uniformity from the run-length matrix; correlation, energy, information measure correlation 2, and max probability from the co-occurrence matrix; and volume. Features that were always significant regardless of the preprocessing technique were high and low gray-level run emphasis from the run-length matrix, mean from the histogram, and the original algorithm for coarseness. In general features were more likely to have a significant p-value after Butterworth smoothing or both Butterworth smoothing and 8 bit depth resampling were used.

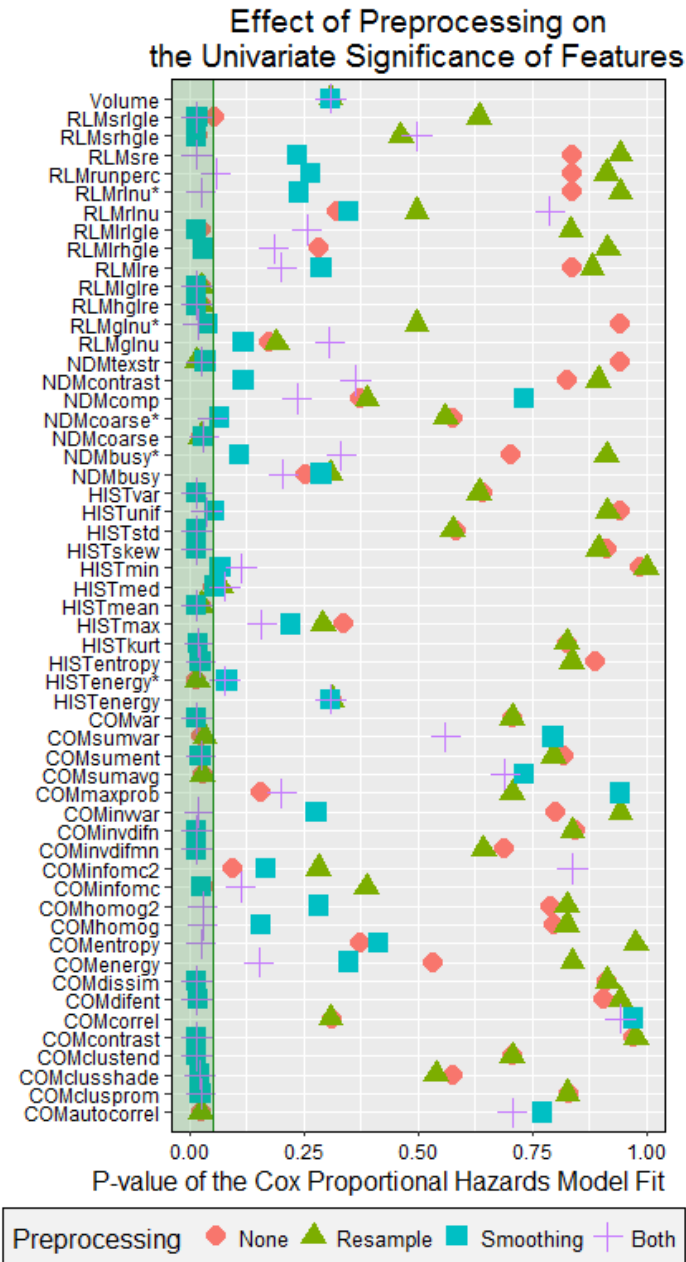


Figure 4.12: The p-values after Benjamini-Hochberg correction for the univariate cox proportional hazards model for each feature and preprocessing combination. The green region of the plot indicates significant values:  $p\text{-value} < 0.05$ . Volume was included in the feature set for comparison.

The c-indices calculated from the predicted values for each univariate model are plotted in Figure 4.13. The c-index for volume was 0.56. The largest calculated c-index was 0.65 for the median from the histogram after both Butterworth smoothing and 8 bit depth resampling. The next highest c-index was 0.60 for both high gray-level run emphasis and short run high gray-level run emphasis from the run-length matrix. In general, using Butterworth smoothing either alone or with 8 bit depth resampling resulted in c-indices close to or slightly larger than the c-index for volume, whereas using 8 bit depth resampling on its own or not using any image pre-processing resulted in c-indices  $< 0.5$ . With the exception of minimum intensity, for every feature at least one preprocessing technique resulted in a c-index greater than 0.5.

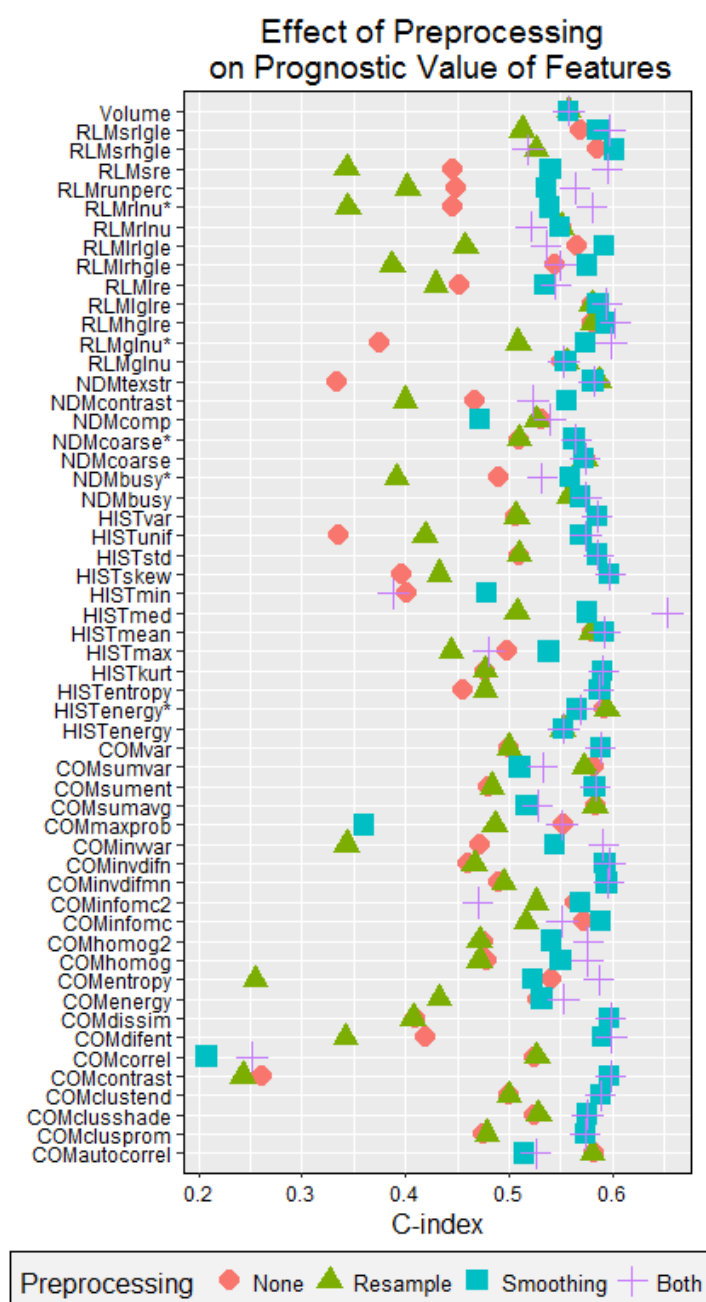


Figure 4.13: Harrell's concordance index (c-index) for each feature and preprocessing combinations. Predictions were generated for each patient during leave-one out cross validation. With the exception of minimum intensity, for every feature at least one preprocessing technique resulted in a c-index value greater than 0.5, though only one was larger than 0.6.

The Benjamini-Hochberg corrected p-values for the log-likelihood ratios comparing Cox proportional hazards models for overall survival fitted with volume only and fitted with volume and one of the radiomics features are plotted in Figure 4.14. Of the 54 features, 25 had at least one significant p-value from this test. Most of the significant features were calculated with either Butterworth smoothing or Butterworth smoothing and 8 bit depth resampling. Short run high gray-level emphasis energy added significant value to the model when no preprocessing was used, texture strength from the neighborhood difference matrix added significant value when 8 bit depth resampling was used, and the volume-corrected version of energy from the histogram was significant when either no preprocessing or 8 bit depth resampling was used. Approximately half of the features (29/54) were not significant regardless of which preprocessing technique was used. This subset included at least one feature from each of the feature categories and all 5 of the uncorrected, volume correlated features identified in the previous section.



### Prognostic Potential versus Volume Dependence

The corrected p-values for the log-likelihood ratio between Cox proportional hazards models fit with volume as their only covariate and models fit with both volume and one radiomics feature are plotted against the volume correlation for each feature after each preprocessing technique, Figure 4.15. All but one of the features with a significant p-value for the log-likelihood ratio had at least a slight correlation with volume ( $r_s > 0.5$ ). However, many features with equally high or higher correlations with volume did not have significant p-values. Thus, features with significant p-values and some correlation with volume are likely providing complementary information. The only feature with a significant p-value and a correlation coefficient less than 0.5 was the minimum of the histogram after Butterworth smoothing. Features with very high volume ( $r_s > 0.95$ ) correlations were not able to add significant value to models built using volume. These features included all of the preprocessing versions of the five original, volume correlated algorithms from the first section of the results, as well as the unprocessed versions of the information measure correlation and information measure correlation 2 from the co-occurrence matrix and the smoothed version of the information measure correlation 2.

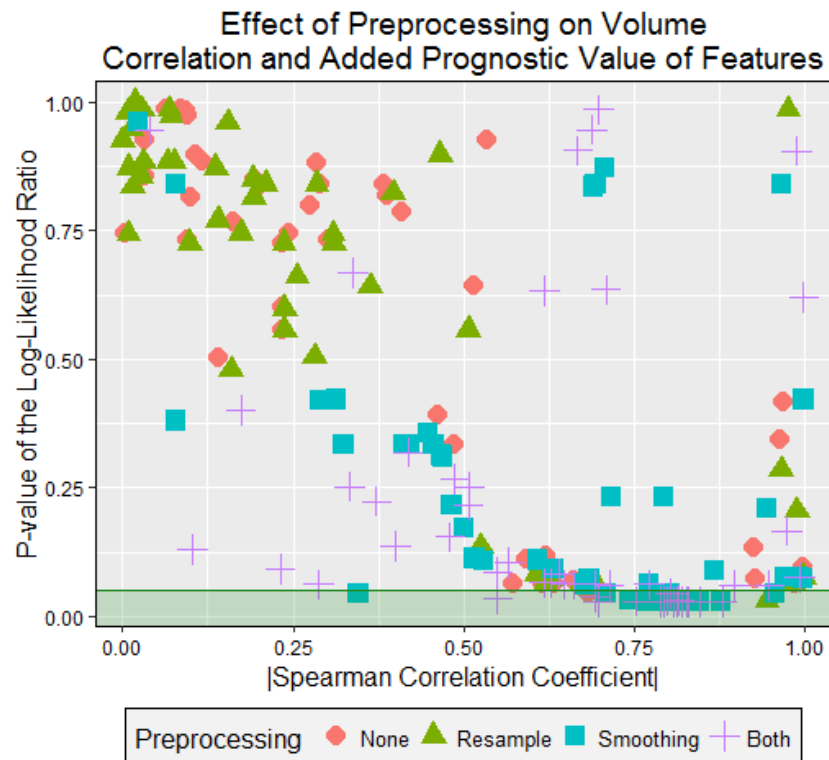


Figure 4.15: Effect of preprocessing on the volume correlation versus the added prognostic value of the features. Volume dependence was measured with the spearman correlation coefficient and the added prognostic value of the features was measured with the log-likelihood ratio between cox proportional hazards models for overall survival using only volume and models using volume and one radiomics feature. The green area highlights features whose p-value was  $< .05$  and thus significant.

## Discussion

This analysis demonstrated that preprocessing can have a strong impact on the volume dependence and univariate significance of many radiomics features. Specifically, Butterworth smoothing increased the likelihood that a feature was significant in univariate Cox proportional hazards models and significantly improved the model fit in Cox proportional hazards models that included volume as a second covariate. This may be because smoothing removes some of the noise in the image and thus allows the measured features to better represent the tumor's

relative heterogeneity and thus its likelihood of responding to treatment. However, this preprocessing technique also increased the correlation with volume of a feature, suggesting that preprocessing techniques must be chosen carefully.

Although various studies have identified features that on their own or as part of a model may yield prognostic information, very little research has been done on the physical basis for high or low feature values. As observed in this study, one tumor characteristic that can influence feature values is tumor volume. We identified 5 features that were highly correlated with volume owing to terms in the feature algorithms that are directly affected by the number of voxels in the entire image. This dependence would not have been an issue in the original design of these features, which were used only to compare aerial photographs that were the same size. However, in tumor analysis, patients with the same relative heterogeneity can have substantially different tumor sizes and thus widely different values for a radiomics feature if it is dependent on the number of voxels. In our analysis, simple normalizations of the original algorithms were able to lower these correlations. Additionally, we showed that the original versions of the algorithms for these 5 features were not able to add significant value to outcome models that already had volume as a covariate. While for two of these features (energy and grey-level non-uniformity), normalizing them did result in significant p-values. Because the direct dependencies we discovered were inherent to the texture equations and not the images, the same relationships are likely to exist in images of different types of cancer, especially those that span a large range of volumes. Similarly, although we used three-dimensional ROIs to capture the full heterogeneity of the tumor, several previous studies have used only the largest axial image slice when determining their ROI<sup>21,37</sup>. The strong dependencies we found for these five features will also apply to two-dimensional slice studies because the algorithms are inherently volume-dependent. Thus, we recommend that future studies consider including these modified algorithms in their future feature sets in place of the original volume-dependent features.

This work also presented a simple but highly versatile technique for studying the merit of individual texture features independently of the confounding variables inherent to patient data. Digital phantoms with either a defined pattern or random values from a Gaussian distribution were useful for demonstrating the possible range of values and volume dependencies of a particular texture. The digital phantoms could also allow users to assess whether their understanding of the image characteristics that lead to relatively high or low texture values is correct. It may be possible for users to extend that understanding to predict whether a tumor will have a relatively high or low texture value. As demonstrated in these results, this methodology could also be used to identify other weaknesses in features, test the ability of features to differentiate data, and potentially establish the most useful parameters for the calculation of features.

A large fraction of the features studied in this work both with and without image preprocessing were at least slightly ( $r_s > 0.5$ ) correlated with volume. These relationships are not necessarily problematic, as the features may still provide information that is complementary to volume. For example, surface area is known to be correlated with volume, yet provides important new information. This idea was supported by the fact that almost all of the features with a significant p-value for the log-likelihood test comparing models with volume as a covariate to models with volume and a radiomics feature were at least slightly correlated to volume ( $r_s > 0.5$ ). These correlations may be due to actual differences in the heterogeneities of large versus small tumors on average which the features are designed to measure. To reiterate, a feature correlated with volume should not necessarily be excluded from a dataset, but a feature calculated with an algorithm that is inherently dependent on the number of voxels should be changed or removed. Otherwise, that feature would return two different values when measured from two ROIs of different sizes even if both have the same intensity in each pixel, e.g. two circles filled with pixels of intensity 20 but one has a radius of 5 pixels and one has a radius of 10 pixels.

In this study we also examined the impact of different preprocessing techniques on both the correlation with volume and prognostic significance of each radiomics feature. For some features an increase in the correlation with volume due to preprocessing may represent the amount of information lost in the image. For example, an image that has been overly smoothed eventually has only one intensity value in all of its voxels. Then, because all of the texture information captured by radiomics features has been erased, the feature could represent only the volume information, which is not affected by image preprocessing. However, using no preprocessing at all can also result in meaningless feature values because the values can be dominated by noise in the image. The ideal preprocessing technique for a particular feature would reduce this image noise while maintaining the tumor's actual relative heterogeneity to generate useful information for modeling. Because a ground truth is not known for radiomics features, we used the significance of the features in univariate analysis to evaluate the usefulness of each feature. If a feature was significant in the univariate analysis, then the preprocessing was concluded to have helped it. We found that, in general, using a Butterworth smoothing filter, either on its own or in conjunction with 8 bit depth resampling, resulted in the ability to extract statistically significant features from tumor ROIs. However, the specific trends were feature- dependent. Thus, feature-specific image preprocessing may be required to maximize the usefulness of each radiomics feature. This is perhaps not surprising considering the differences in specific features. For example, the mean intensity from the histogram would change less with smoothing than a feature from the co-occurrence matrix, which could benefit from appropriate bin sizes in the calculation of the matrix and thus the right choice for bit depth rescale.

One limitation of the current study is that only 3 different preprocessing techniques were evaluated in-depth. It is possible that superior preprocessing techniques could exist, such as using voxel size resampling or edge-detection filtering, or that fine-tuning the parameters could improve these techniques. However we did perform an exploratory analysis of the impact of

fine tuning certain filter parameters prior to the in-depth analysis. Each of the eventually selected preprocessing techniques was compared to various related filter iterations. For example the impact of the 8 bit depth resample on features was compared to using bit depths between 6 and 11 instead and the impact of different Butterworth frequency cutoffs was investigated both with and without the added 8 bit depth resample. This exploratory analysis suggested that using any smoothing filter had a bigger impact on the relative feature values than the difference between two smoothing filters or two different Butterworth frequency cutoffs. Similarly resampling the bit depth affected the feature values but the difference between two specific bit depth levels such as 10 bit and 8 bit were minor. Thus the preprocessing techniques that were finally selected for use in the in-depth analysis portion of this study did not comprise an exhaustive set but instead were selected to demonstrate the large changes in a feature's univariate significance that can occur by using different methods for noise-reduction before feature calculation. Because studies have been published with the same features but different preprocessing techniques and parameters for their feature matrices (co-occurrence matrix, neighborhood difference matrix, and run-length matrix) this is an important result that must be investigated in order for these features to eventually be standardized and used clinically.

Another limitation of this study is that it is likely that many of the specific trends described here will be different for other image modalities or tumor sites. However the overall conclusion that image preprocessing can substantially affect the overall usefulness of a feature should apply in any case. Thus, we highly recommend that future studies examine the most appropriate features to be used for a particular patient population and the calculation parameters accompanying those features before including the features in prognostic models.

## **Conclusions**

Radiomics features calculated from a variety of imaging modalities are being widely studied for potential to help predict patient outcomes or aid physicians in diagnosis. However, so far studies have calculated features using a wide range of software, parameters, and pre-

processing techniques. The goal of this aim was to demonstrate the effect that different pre-processing techniques can have on the usefulness of radiomics features by measuring the volume-dependence and prognostic value of each feature in univariate models. We proposed normalization factors for five features that were highly volume-dependent regardless of the preprocessing technique used. Additionally, we found that most features benefited from image smoothing using a Butterworth filter, either alone or in conjunction with 8 bit depth resampling. While smoothing was more likely to make a feature statistically significant in a univariate model, smoothing also tends to increase the volume dependence of the feature. It is important to balance these two effects in order to determine the optimal preprocessing technique for each feature.

## Chapter 5 : Prognostic Potential of Radiomics Features

A substantial portion of this chapter is written or based on the following publication:

Fave X, Zhang L, Yang J, Mackin D, Balter P, Gomez D, Followill D, Jones AK, Stingo F, Liao Z, Mohan R, and Court L. Delta-radiomics features for the prediction of patient outcomes in non-small cell lung cancer. Scientific Reports. Doi: 10.1038/s41598-017-00665-z. Volume 7. © Nature Publishing Group. Licensed under CC BY 4.0 available at <https://creativecommons.org/licenses/by/4.0/legalcode>.

The permission for reuse of this material was obtained from Nature Publishing Group ©.

In this chapter we describe the results for Specific Aim 2: Determine which radiomics features, measured from CT images, change significantly during the course of radiation therapy and the relationship between these changes and outcome. Our working hypothesis for this aim was that radiomics features can be identified that change during the course of treatment and are predictive for patient outcome.

### Introduction

Most NSCLC radiomics studies have focused on identifying heterogeneous tumors before treatment to identify high-risk patients with potentially more aggressive tumors. However, a challenge that remains to be addressed is the fact that tumors of the same phenotype can respond very differently to treatment<sup>34</sup>. A model that could effectively identify patients whose tumors are not responding to treatment would be beneficial and could be used to recommend patients for adjuvant chemotherapy or a radiation boost. Measuring radiomics features over the course of radiation therapy may allow for the identification and quantification of therapy-induced changes in the tumor. These changes in radiomics features, defined here as delta-radiomics features, may indicate which patients are responding to treatment. A large change in the value of a radiomics feature could indicate better response and potentially better long-term outcomes, while the value of the feature at the end of treatment may be indicative of the treatment's success.

This study had several goals. The first was to identify radiomics features that demonstrate significant changes during radiation therapy. Next, the prognostic value of univariate models using radiomics features measured at the beginning of treatment, end of treatment, or the net changes or slopes during treatment were compared for the outcomes of overall survival, freedom from distant metastases, and local-regional control. Finally multivariate models were built to determine whether pretreatment or delta-radiomics features could improve the prognostic ability of models built using only clinical factors.

## **Methods**

### Patient Data

For this study, we retrospectively reviewed the images and medical records for the patient cohort discussed in Chapter 3. This cohort was comprised of 107 NSCLC patients that were treated at the University of Texas MD Anderson Cancer center as part of an IRB approved clinical trial<sup>45</sup>. The retrospective analysis performed for this study was approved by our IRB with a waiver of informed consent. The patients were treated with radiation therapy and concurrent chemotherapy. They had been randomized to receive treatment with either photons or protons to 66 or 74 Gy. Because of their participation on this trial the patients were imaged weekly during treatment with a four-dimensional CT (4DCT)<sup>45</sup>. For this retrospective analysis, the medical records of these patients were reviewed to determine their clinical factors: sex, age, smoking status, pack years, tumor histology, overall disease stage, T stage, N stage, Karnofsky performance status (KPS), and total prescribed radiation dose. In this analysis, treatment modality was not considered for classification purposes although the impact of modality on radiomics features was investigated separately in Chapter 7.

The primary endpoints for this analysis were overall survival, freedom from distant metastases, and local-regional control. For this analysis, a local-regional recurrence was defined as evidence of disease inside the treatment field, adjacent to the treatment field, or in

the regional lymph nodes. A distant metastasis was defined as evidence of disease anywhere else in the body including the contralateral lung. Patients were censored at their last date of follow-up if they did not reach the endpoint.

Table 5.1: Clinical characteristics of the NSCLC patient population used for modeling.

Clinical Factors	Number of Patients (n=107)
Sex	
F	45
M	62
Age	
<65	45
≥65	62
T stage	
T1 or T2	49
T3 or T4	58
N stage	
N0 or N1	24
N2 or N3	83
Overall disease stage	
II	12
IIIa	44
IIIb	49
IV	2
Tumor histology	
Squamous cell carcinoma	46
Adenocarcinoma or other	61
Smoking status	
Current	34
Former	64
Never	9
Pack years (continuous)	
0-24	20
25-49	37
50-74	28
75+	22
Karnofsky performance status	
90-100	52
70-80	55
Total radiation dose	
>70 Gy	72
<70 Gy	35

## Landmark Analysis

Survival studies using measures of response that are calculated at multiple time points, such as the radiomics features measured at weekly intervals in this study, require a landmark time point to be used for calculating the time until the endpoint is reached<sup>78,79</sup>. Otherwise a bias can be introduced by responders since they must have already survived to the time of treatment to be classified as responders<sup>78,79</sup>. In this study, patients were classified as high or low risk using multivariate models that included clinical factors (recorded at the time of entrance to the study), pre-treatment radiomics features (measured from the treatment planning images), and delta-radiomics features (measured from the different weekly images through treatment). Because a variable number of days occurred for each patient between when they entered the trial and when their pre-treatment images were acquired and between their pre-treatment images and last weekly images, a landmark time point was required to measure survival. For this study, endpoints were defined from a landmark time point of 90 days from the day the patient was entered on the clinical trial until one of the endpoints of death, presence of distant metastases, or local-regional failure was met. Patients not reaching the endpoint were censored at their last follow-up date. The landmark point was calculated by determining the total number of days from entering the trial to end of treatment for each patient. The maximum interval (by which all measures of response had been determined) was 83 days, which was rounded to 90 days for simplicity. This ensures that the time until the endpoint is reached or the patient is censored is uniformly measured across all patients and not biased by the number of days they have already survived to reach the end of treatment.

## Features

Features included in this analysis were shape features (n=16), intensity histogram features (n=11), co-occurrence matrix features (n=22)<sup>11,12</sup>, neighborhood difference matrix features (n=5)<sup>14</sup>, and run-length matrix features (n=11)<sup>13</sup>, Table 5.2. To determine the best parameters for the non-shape features, each feature was calculated four times: (i) with no

image preprocessing other than thresholding, (ii) with smoothing using a Butterworth filter with an order of 2 and a cutoff of 125 followed by thresholding, (iii) with thresholding followed by an 8-bit depth resample, and (iv) with Butterworth smoothing, thresholding, and an 8-bit depth resample<sup>80</sup>. The Butterworth smoothing acts to remove Gaussian noise from the images, which may obscure the lower frequency biological variations that radiomics features are designed to measure. The 8-bit depth resample is used as an alternative to modifying the binning parameter for the histogram and radiomics matrices. Using 8-bit depth images results in a bin width of 16 HU and thus is more likely to reflect actual density changes in neighboring pixels than bins with a width of 1 HU, which largely reflect image noise. More details on these image processing techniques is available in Chapter 4 where we examined their impact on common radiomics features in detail.

Table 5.2: List of features used in the analysis of univariate and multivariate prognostic potential.

Histogram	Co-occurrence Matrix	Run Length Matrix	Neighborhood Difference Matrix	Shape
Energy	Autocorrelation	Gray-level	Busyness	Compactness1
Entropy	Cluster	non-	Coarseness	Compactness2
Kurtosis	prominence	uniformity	Complexity	Convex
Maximum	Cluster shade	High gray-level	Contrast	Convex hull
Mean	Cluster	run emphasis	Texture Strength	volume
Median	tendency	Long run		Convex hull
Minimum	Contrast	emphasis		volume 3D
Skewness	Correlation	Long run high		Mass
Standard deviation	Difference	gray level		Max 3D
Uniformity	entropy	emphasis		diameter
Variance	Dissimilarity	Long run low		Mean breadth
	Energy	gray level		Number of
	Entropy	emphasis		objects
	Homogeneity	Low gray-level		Orientation
	Homogeneity 2	run emphasis		Roundness
	Information	Run		Spherical
	measure	percentage		disproportion
	correlation	Run-length		Sphericity
	Information	non-		Surface area
	measure	uniformity		Surface area
	correlation 2	Short run		density
	Inverse	emphasis		Volume
	difference	Short run low		
	moment norm	gray-level		
	Inverse	emphasis		
	difference	Short run high		
	norm	gray-level		
	Inverse	emphasis		
	variance			
	Max probability			
	Sum average			
	Sum entropy			
	Sum variance			
	Variance			

Optimal image pre-processing was determined on a feature-specific basis using the following steps, which are also illustrated in Figure 5.1. First a univariate Cox regression model for overall survival was fitted for each preprocessed version of each feature using only the pretreatment images for each patient. The significance of the feature in the model was calculated to determine whether a model built on only this feature was a better fit than the null model. This step identified radiomics features that were predictive and therefore might be useful for calculating delta-radiomics features. Next a Wilcoxon rank sum test was performed for each feature and pre-processing combination to determine if the feature values were significantly different when images were acquired on the GE Discovery ST versus the GE Lightspeed RT16, as CT scanner model has been demonstrated to be an important factor in feature reproducibility<sup>40</sup> and the overwhelming majority of images had been acquired on these two scanners (755 of the 785 images). The remaining images were acquired on either a Philips Brilliance 64 (n=21) or Philips Brilliance Big Bore (n=9). A patient subset that had images available from the first week of treatment was used for this test because at this time point the patients were roughly split between the two CT scanners used in this study (37 patients imaged with the GE Discovery ST and 44 patients imaged with the GE Lightspeed RT16) and their tumors would not yet have shown any therapy-induced changes. Finally, the correlation between each feature and the gross tumor volume was calculated using Spearman's rank correlation coefficient. The feature values and gross tumor volumes were calculated from the pretreatment images for this step. For each feature, the pre-processed version that was significant in univariate analysis for survival (p-value <0.10) and did not have a significant value (p-value >0.05) for the Wilcoxon rank sum test between CT scanners was included in the final feature set. Features that never met these two criteria regardless of the image pre-processing used were excluded from the feature set. If a feature met both criteria for more than one image pre-processing type, the version of the feature that had the smallest correlation with volume was selected. A p-value of 0.10 was used as the threshold for significance in this pre-analysis

because the p-values were used only for feature selection, not hypothesis testing, and thus the filtering need not be overly stringent. This choice was balanced against the need to remain conservative so that the feature dimensionality is decreased during this step. For the same reason, no multiplicity correction was used at this stage.

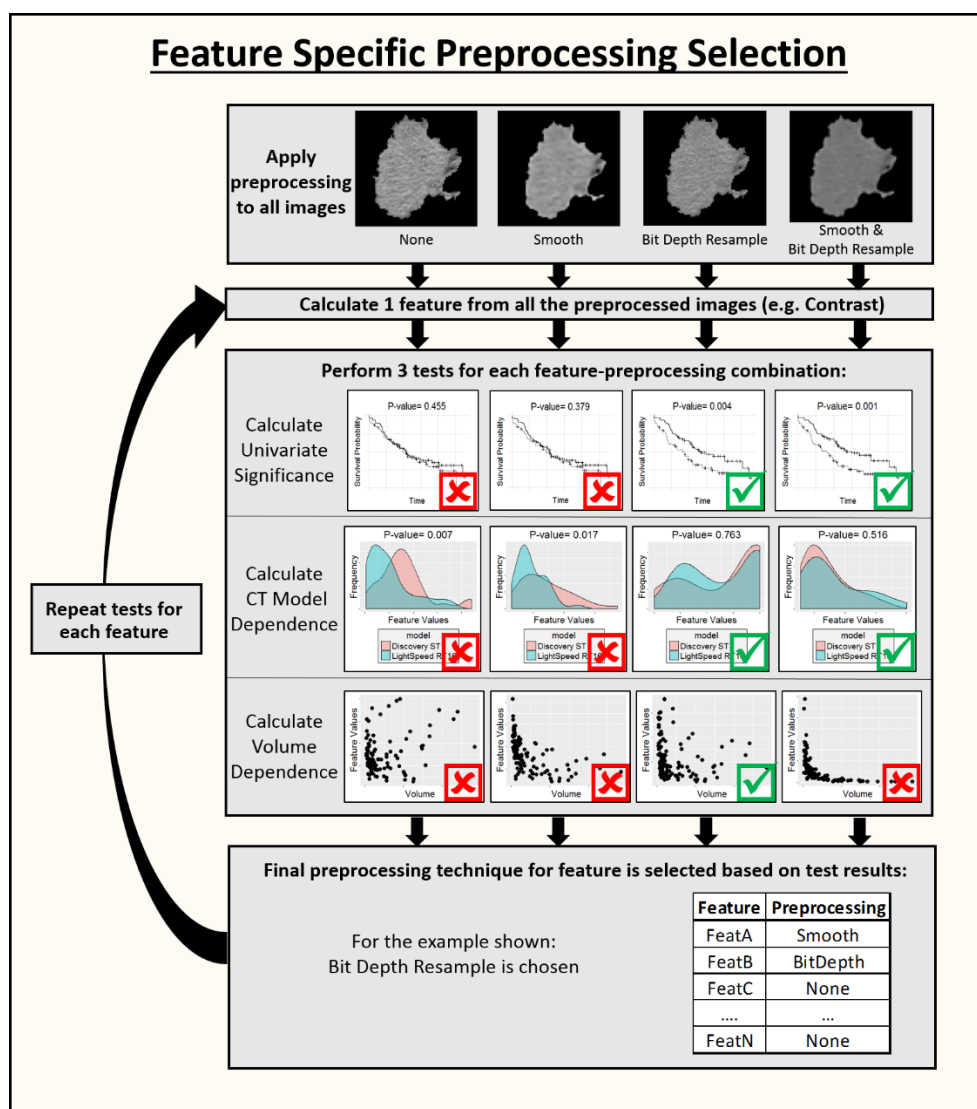


Figure 5.1: Workflow for the selection of feature specific image preprocessing. The images are all processed in four ways: no extra processing, smoothing with a Butterworth filter, resampling to an 8 bit depth, and both smoothing with a Butterworth filter and resampling with an 8 bit depth. Each feature is calculated from the four sets of processed images. Then the best processing is determined on a feature specific basis by evaluating the univariate significance, dependence on the CT model used to acquire the images, and volume dependence. Based on the results of these tests, one image preprocessing is selected for that feature. Then the process is repeated for the next feature. If no image preprocessing for a feature allows it to pass the tests, then the feature is removed entirely.

### Delta-radiomics Features

Two tests were conducted to determine which of the optimized features changed during treatment and thus might be useful indicators of tumor response. First a linear mixed effects model with random intercepts for each patient was built for each feature in the form,

$$\Delta Feature \sim \Delta time + (1|PatID)$$

Equation 5.1

Here,  $\Delta Feature$  was the feature value measured from each weekly 4DCT,  $\Delta time$  was the number of days from the commencement of treatment before the image was acquired, and  $PatID$  was a patient-specific identifier that allows the model to account for the fact that we had multiple, longitudinal measurements of each feature for each patient by assigning each patient their own intercept. The p-value of the log-likelihood ratio for each model was calculated. P-values were corrected for multiple comparisons using the Benjamini-Hochberg method<sup>77</sup>. If the corrected p-value was less than 0.05, the model was considered significant and indicated that the changes in the feature were significantly associated with the time since treatment had begun. For each feature with a significant p-value in this test, simplified measures of the overall change were calculated and defined as delta-radiomics features. The delta-radiomics features were defined as the percent net change, Equation 5.2, the linear regression slope, and the value of the feature at the last week of treatment for each patient.

$$\% Net Change = (Feature_{WeekFinal} - Feature_{Week1}) / Feature_{Week1}$$

Equation 5.2

Here,  $Feature_{WeekFinal}$  was the value of the feature at the end of treatment and  $Feature_{Week1}$  was the value at the first weekly 4DCT for each patient. A one-sample, two-tailed *t*-test was conducted for the percent net change and linear regression slope delta-radiomics features to determine whether the overall changes for the group were significantly different from 0, and

values were again corrected using the Benjamini-Hochberg method. Features that passed both the linear mixed effects and *t*-test analyses (corrected p-value <0.05) were considered to demonstrate significant therapy-induced changes and were included as potential delta-radiomics covariates in both univariate and multivariate model building.

### Univariate Analysis

Covariates examined for the univariate analysis were the clinical factors listed in Table 5.1 and the radiomics and delta-radiomics features that passed the previous tests. For each of the radiomics features, four versions were tested: the values at pre-treatment (pre-TX), the percent net change over the course of treatment (% net change), the linear regression slope (slope) over the course of treatment, and the values at the end of treatment (end-TX).

For each of these covariates, the univariate model fit and predictive performance were assessed for each outcome (overall survival, freedom from distant metastases, and local-regional recurrence) using two tests:

- (i) P-value of the log-likelihood ratio: Univariate cox proportional hazards models were fit to the data using the entire patient set. The log-likelihood ratio of the fit was calculated to determine whether or not the fit was significantly better than the null model. P-values were corrected for multiplicity using the Bonferroni correction. The Bonferroni correction was used because the different versions of the radiomics features are not independent from each other.
- (ii) C-index: Univariate cox proportional hazards models were then generated in a leave-one-out cross validation (LOOCV) loop. On each iteration of the loop, the coefficient for the single covariate was refitted and then a prediction for the left-out patient was calculated using the model. The predictions generated from the LOOCV were used to calculate Harrell's concordance index<sup>81</sup> (c-index). The c-index is analogous to the area under the curve but is designed for survival data

instead of binary data. Values of the c-index can range from 0 to 1 with a value of 1 indicating perfect prediction and a value  $\leq 0.5$  indicating that a model performs no better or worse than a random guess. Thus, the overall c-index is a more likely representation of how the model would perform on new patients.

### Multivariate Analysis

Multivariate Cox regression models were built for each of the primary endpoints using leave-one-out cross validation (LOOCV) and Akaike Information Criterion (AIC) with the following procedure, which is also illustrated in Figure 5.2. First, one patient was removed from the dataset and a Cox proportional hazards model was built using all of the clinical factors and the remaining patients. The covariates were reduced using stepwise AIC in both directions. Next, all of the pretreatment radiomics features were added to this model individually and the log-likelihood ratio was calculated to determine if they had significantly improved the model. Then the subset of features that had significantly improved the model fit (p-value  $< 0.05$ ) were all added together to the clinical model. Stepwise AIC was repeated in both directions with forced nesting of the clinical covariates to select the best pretreatment features. These steps were then repeated with the delta-radiomics features with forced nesting of the clinical and pretreatment radiomics covariates. The delta-radiomics versions of the features were identified by the suffixes “netPercentChange”, “Slope”, or “WeekLast”, while the pretreatment radiomics features are indicated by the suffix “Week0”. This process was repeated with each patient left out in turn so that at the end there were three models for each left-out patient: one with only clinical factors, one with clinical factors and pretreatment radiomics features, and one with clinical factors, pretreatment radiomics features, and delta-radiomics features. The total number of times each covariate was selected for the three models over all of the LOOCV iterations was calculated. Covariates that were selected in more than half of the iterations were retained and considered high-performing. Final versions of the three models using only these frequently selected covariates were then calculated and compared using the log-likelihood ratio

to determine whether the radiomics and/or delta-radiomics features significantly ( $p\text{-value} < 0.05$ ) improved the fit of the model to the data. If no feature was selected in more than half of the iterations for a particular model, then the null model or the nested model from the previous iteration was used.

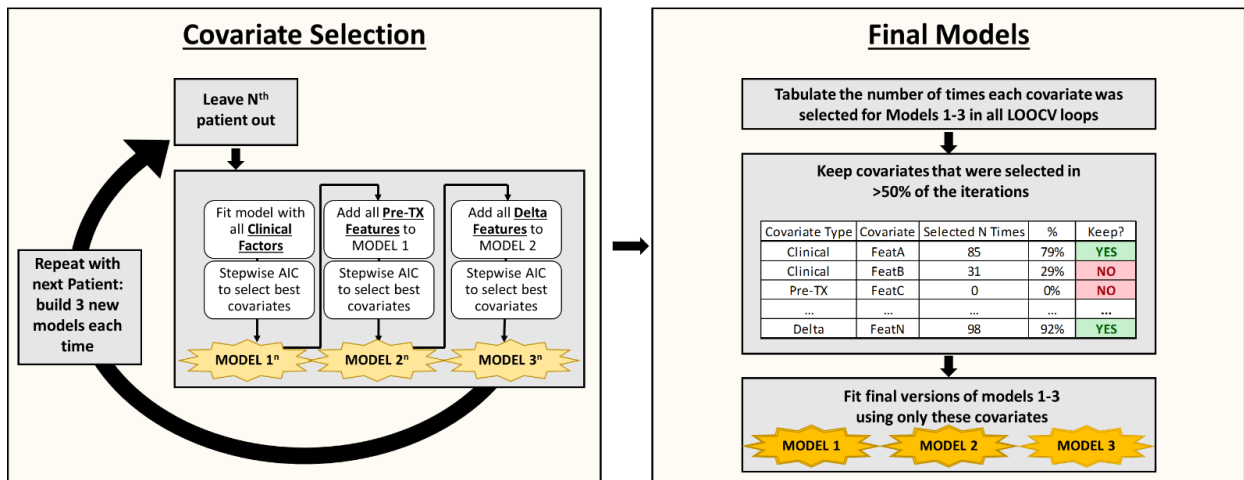


Figure 5.2: Workflow for building of multivariate models. A LOOCV loop is used to generate 3 models on each iteration: (1) Only clinical factors, (2) clinical factors and pre-treatment (TX) features, and (3) clinical factors, pre-TX features, and delta-radiomics features. After the three models have been built with each patient left out once, the number of times each covariate was selected is tabulated. Then those covariates that are selected in greater than 50% of the iterations are kept. These are then used to fit final versions of the 3 models.

To evaluate the prognostic potential of these features, a new LOOCV was performed. For this analysis, the three models were built on each iteration using only the high-performing clinical, radiomics, and delta-radiomics covariates from the original LOOCV. No covariate reduction was performed, but the coefficients were refit on each iteration. On each iteration of the loop, a prediction for the left-out patient was calculated using each of the three models. Because the patient was left out of the coefficient fitting process, predictions generated for the left out patient were unbiased. Once the loop was complete, and each patient had a prediction for each model, the c-index was calculated for each model. The c-indices allowed for the comparison of the predictive accuracy of models that included radiomics and delta-radiomics features to models incorporating only clinical factors. Finally, patients were stratified as high or low risk based on whether their prediction was above or below the median prediction for each model. Kaplan-Meier curves were plotted using this patient stratification, and the log-rank test was used to determine whether the stratifications were significant (p-value <0.05).

All statistical analyses were performed in R language<sup>56</sup> using the survival<sup>57</sup>, lme4<sup>59</sup>, MASS<sup>82</sup>, and ggplot2<sup>60</sup> analysis packages.

## **Results**

### Feature Selection

The initial feature set had 49 texture features measured before treatment, with four different image preprocessing types and 16 shape features, for a total of 212 feature and preprocessing combinations. Of these, 75 were significant in univariate analysis ( $p < 0.10$ ), and 123 were not significantly different between different CT scanners ( $p < 0.05$ ). These results are shown in Figure 5.3-Figure 5.6. Using the feature selection process described in the methods, this feature set was reduced to 31 features. Of these, 9 were calculated with no extra preprocessing, 15 were calculated with Butterworth smoothing, and 7 were calculated with Butterworth smoothing and 8-bit depth resampling, Figure 5.6. No features were calculated

using just 8 bit depth resample. At least one feature from every feature category was represented in this final feature set.

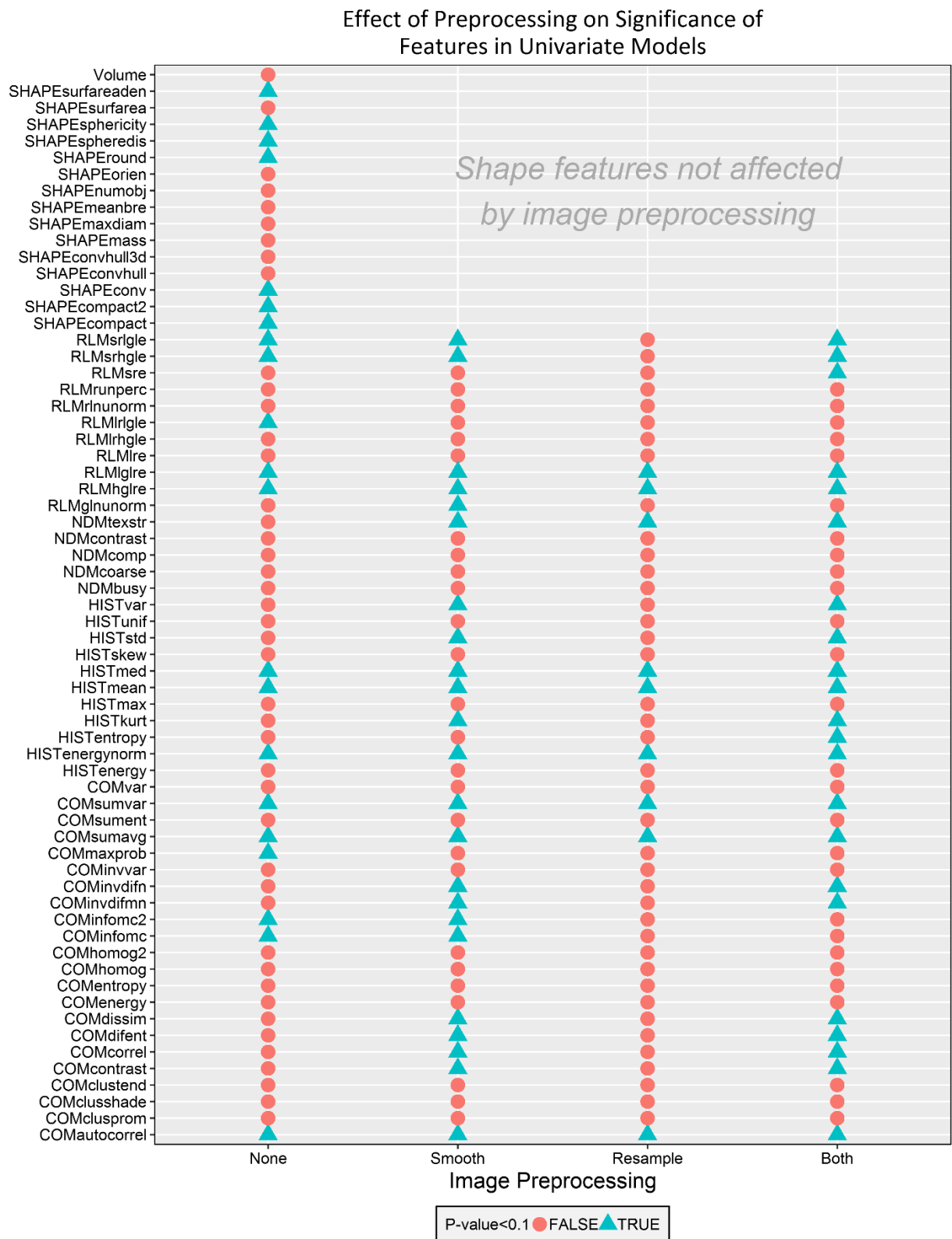


Figure 5.3: Impact of image preprocessing on the univariate significance of radiomics features.

This figure plots each radiomics feature versus the image preprocessing styles used to

calculate it. Blue triangles signify that the Feature-Preprocessing combination resulted in a significant fit for a univariate Cox regression using only the features measured at pretreatment (p-value <0.10). Red circles indicate that the univariate fit was not significant and thus that the feature should not be measured with that preprocessing style. Note that the shape features do not change with image preprocessing and so were calculated only with the basic thresholding step. For the image preprocessing styles: None means that the feature was calculated with only a simple thresholding step, Smooth that the feature was calculated with Butterworth smoothing and thresholding, Resample that the feature was calculated with thresholding and 8-bit depth resampling, and Both that the feature was calculated with Butterworth smoothing, thresholding, and 8-bit depth resampling.

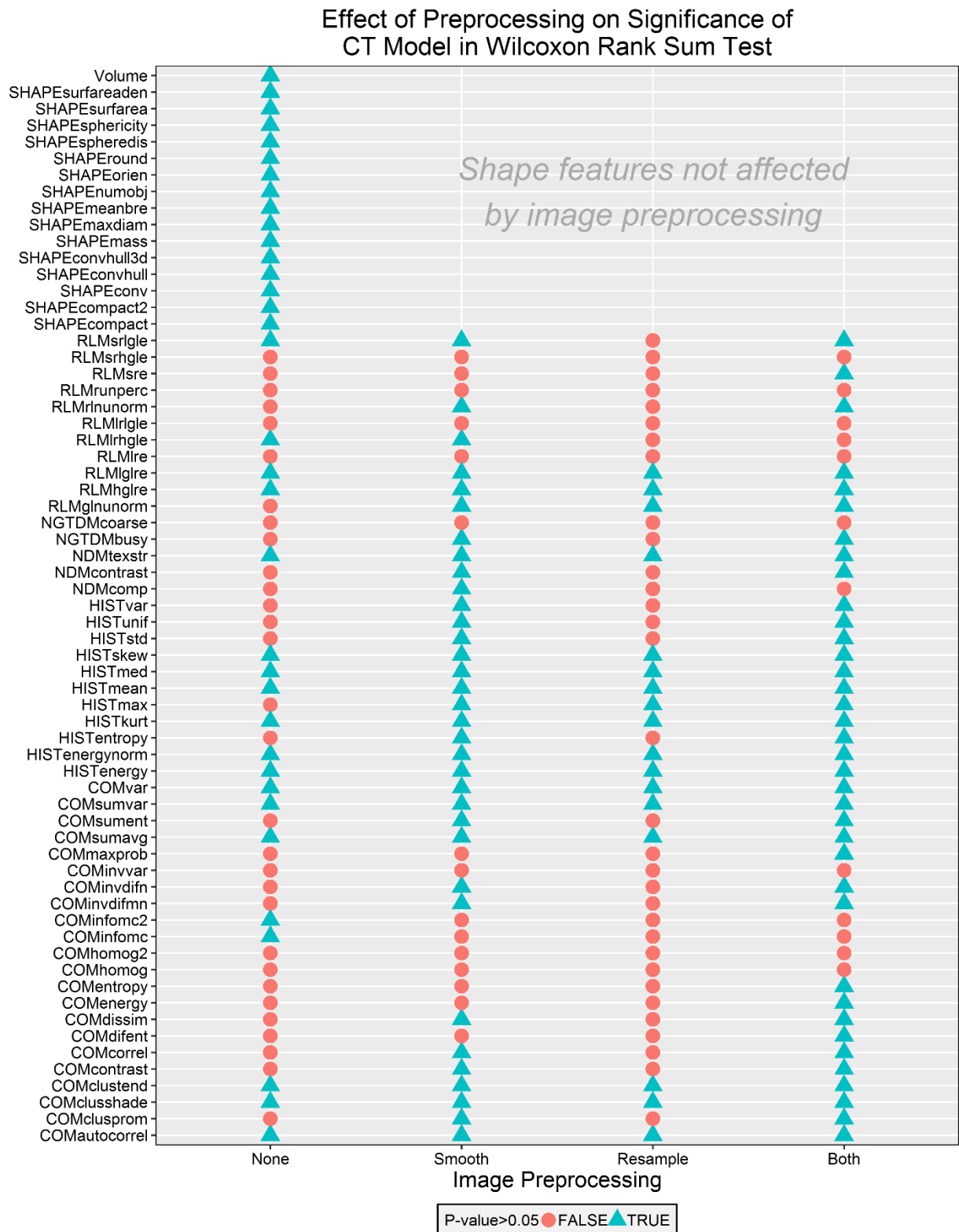


Figure 5.4: Impact of image preprocessing on the significance of CT Model in a Wilcoxon rank sum test for each radiomics feature. This figure plots each radiomics feature versus the image

preprocessing styles used to calculate it. Blue triangles signify that the Feature-Preprocessing combination did not have a significant p-value (i.e.,  $<0.05$ ) in the Wilcoxon rank sum test for the impact of the CT Model (GE Lightspeed RT16 vs GE Discovery ST). Red circles indicate that the p-value was significant and thus that the feature calculated with that preprocessing style was significantly affected by the CT scanner model with which the images were acquired. Note that the shape features do not change with image preprocessing and so were calculated only with the basic thresholding step. None means that the feature was calculated with only a simple thresholding step, Smooth that the feature was calculated with Butterworth smoothing and thresholding, Resample that the feature was calculated with thresholding and 8-bit depth resampling, and Both that the feature was calculated with Butterworth smoothing, thresholding, and 8-bit depth resampling.

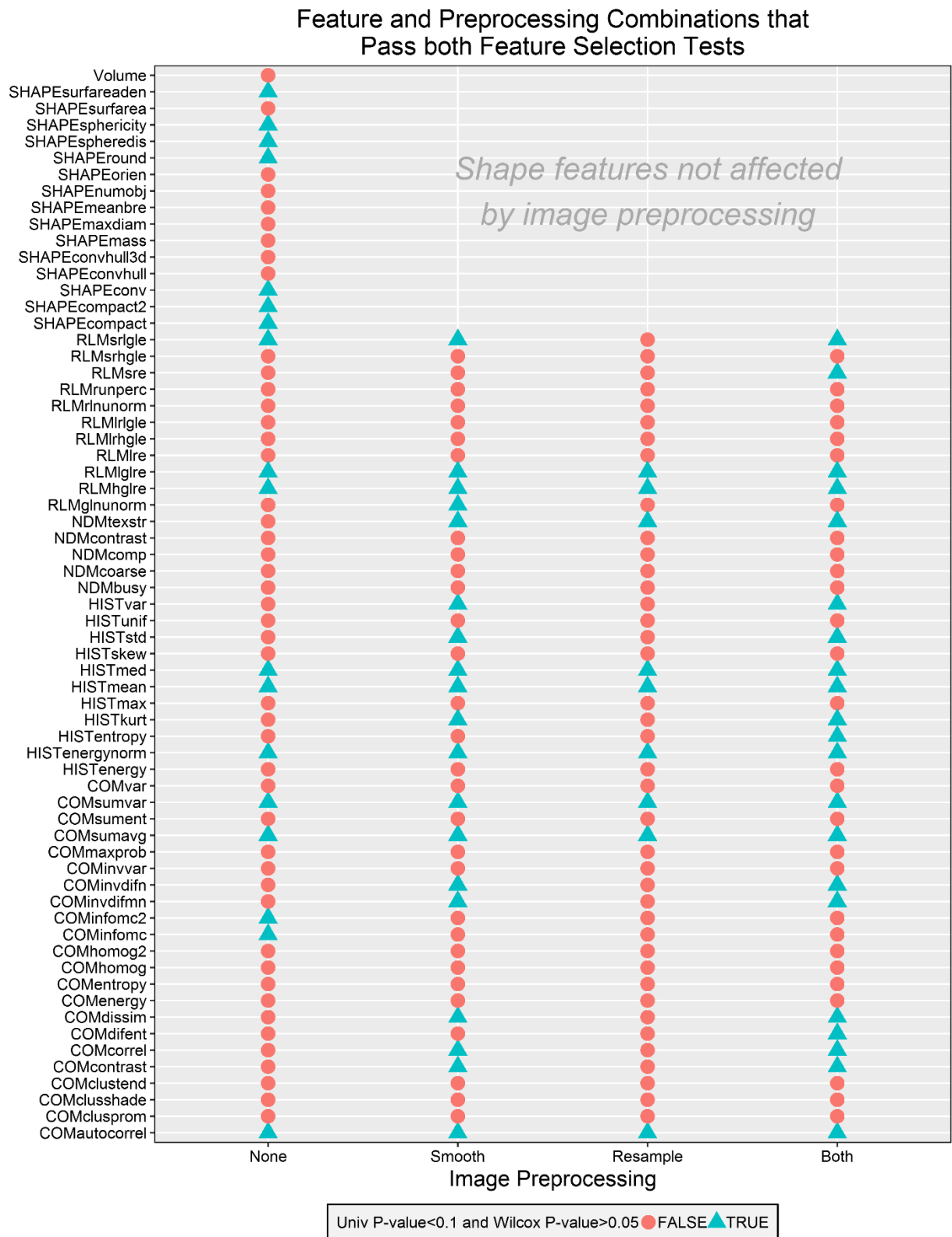


Figure 5.5: Impact of image preprocessing on the univariate significance and significance of the CT Model in a Wilcoxon rank sum test for each radiomics feature. This figure plots each

radiomics feature versus the image preprocessing styles used to calculate it. Blue triangles signify that the Feature-Preprocessing combination passed both tests (had a significant p-value [ $<0.1$ ] in the univariate analysis and did not have a significant p-value [ $<0.05$ ] in the Wilcoxon rank sum test analyzing the impact of CT scanner model). Red circles indicate that the feature failed at least one test. Note that the shape features do not change with image preprocessing and so were calculated only with the basic thresholding step. None means that the feature was calculated with only a simple thresholding step, Smooth that the feature was calculated with Butterworth smoothing and thresholding, Resample that the feature was calculated with thresholding and 8-bit depth resampling, and Both that the feature was calculated with Butterworth smoothing, thresholding, and 8-bit depth resampling.

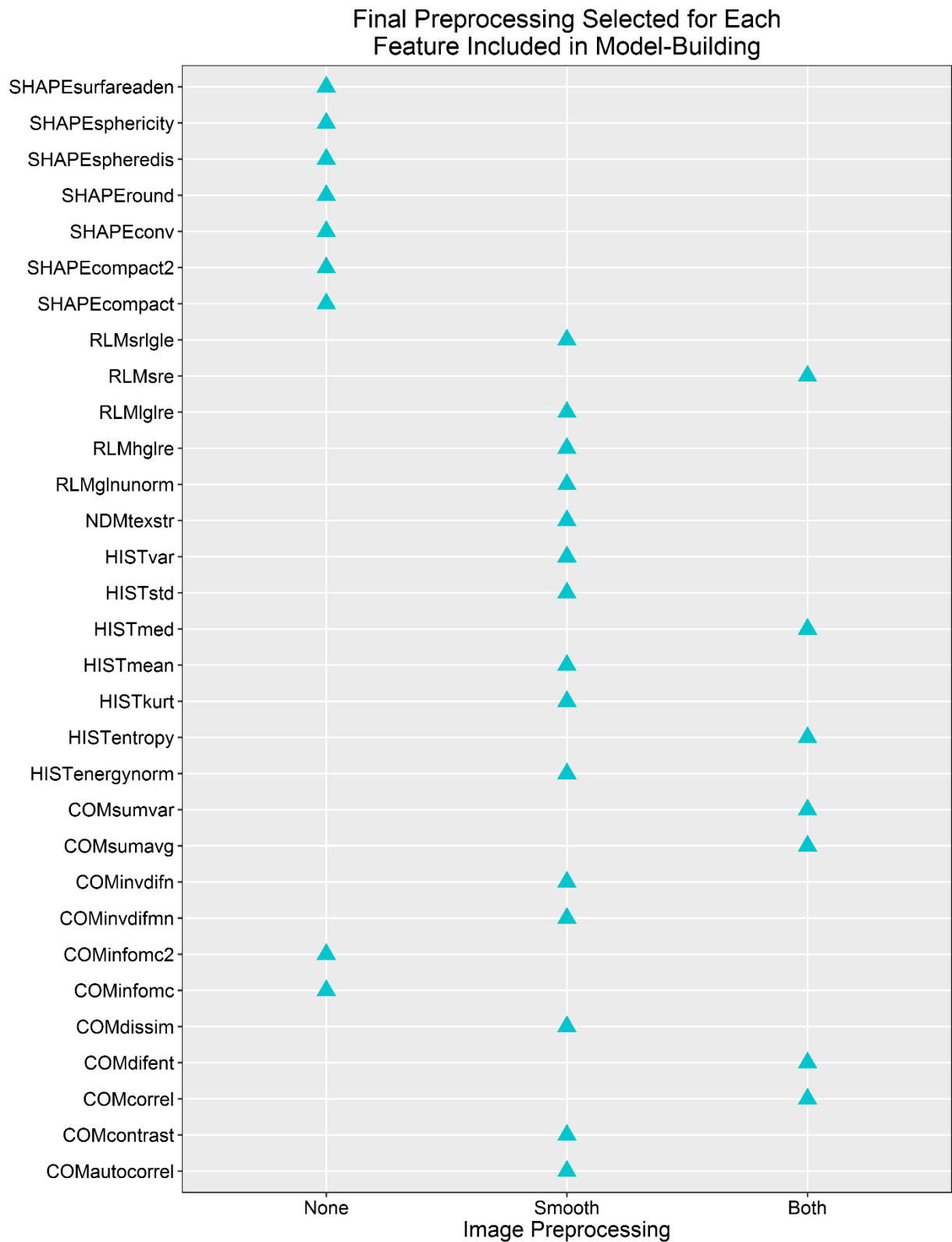


Figure 5.6: Final image preprocessing that was selected for each radiomics feature used in the prognostic analysis. This figure shows the final set of features that remained after our entire

feature selection process along with the feature-specific image preprocessing chosen for each. None means that the feature was calculated with only a simple thresholding step, Smooth that the feature was calculated with Butterworth smoothing and thresholding, Resample that the feature was calculated with thresholding and 8-bit depth resampling, and Both that the feature was calculated with Butterworth smoothing, thresholding, and 8-bit depth resampling. No features were selected using just an 8-bit depth resample so this column does not appear in the figure.

### Delta-radiomics Features

All 31 features had significant p-values for the log-likelihood ratio of their linear mixed effects model, with time since beginning of treatment as the covariate and random intercepts for each patient, even after Benjamini-Hochberg correction for multiplicity. The net changes and slope in each feature were also significant in *t*-tests comparing their means to 0 after multiplicity correction for every feature. Thus a total of 31 features were available for feature selection in the multivariate model building.

### Univariate Analysis

For overall survival, 67 of the 107 patients reached the endpoint of death. The median survival time was 638 days. 50 patients had a distant metastases with a median time until reaching the endpoint or censoring of 311 days. 23 patients had a local-regional recurrence with a median time until reaching the endpoint or censoring of 420 days.

For the univariate analysis of clinical factors, none of the univariate models using clinical factors had a significantly better fit than the null model after multiplicity correction, Figure 5.7. Then when the c-indices were calculated, the majority of features had a c-index below 0.5, and the highest c-index for any of the outcomes was only 0.56, Figure 5.8. Thus, overall the clinical factors did not appear to be highly significant predictors of outcome when used on their own in cox proportional hazards models.

For the univariate analysis of radiomics features, none of the univariate models had a significantly better fit than the null model after multiplicity correction regardless of outcome, Figure 5.9. However, when the c-indices were calculated, many of them were above 0.5, Figure 5.10. For overall survival, the features measured at pre-treatment were most likely to be prognostic. The lowest c-index for a feature measured at pre-treatment was 0.54 for entropy from the histogram while the maximum was 0.69 for median from the histogram. The highest c-indices for the features' end of treatment values, net percent change, and slope were 0.58,

0.58, and 0.60 respectively for information measure correlation from the co-occurrence matrix, sphericity, and compactness.

For distant metastases, the c-index results were similar. The features measured at pre-treatment were most likely to be prognostic with the highest c-index of 0.72 occurring again for the median from the histogram. The highest c-index for the values at the end of treatment was 0.60 for kurtosis from the histogram; for the net percent changes was 0.57 for information measure correlation 2 from the co-occurrence matrix; and for the slopes was 0.58 for compactness.

For local-regional recurrence, the c-indices were relatively low. The highest c-index was only 0.55 and occurred for the values at the end of treatment of texture strength from the neighborhood difference matrix. The highest c-indices for the other feature types were 0.52 for roundness at pre-treatment, 0.52 for the net percent change in inverse difference moment norm from the co-occurrence matrix, and 0.49 for the slope in roundness.

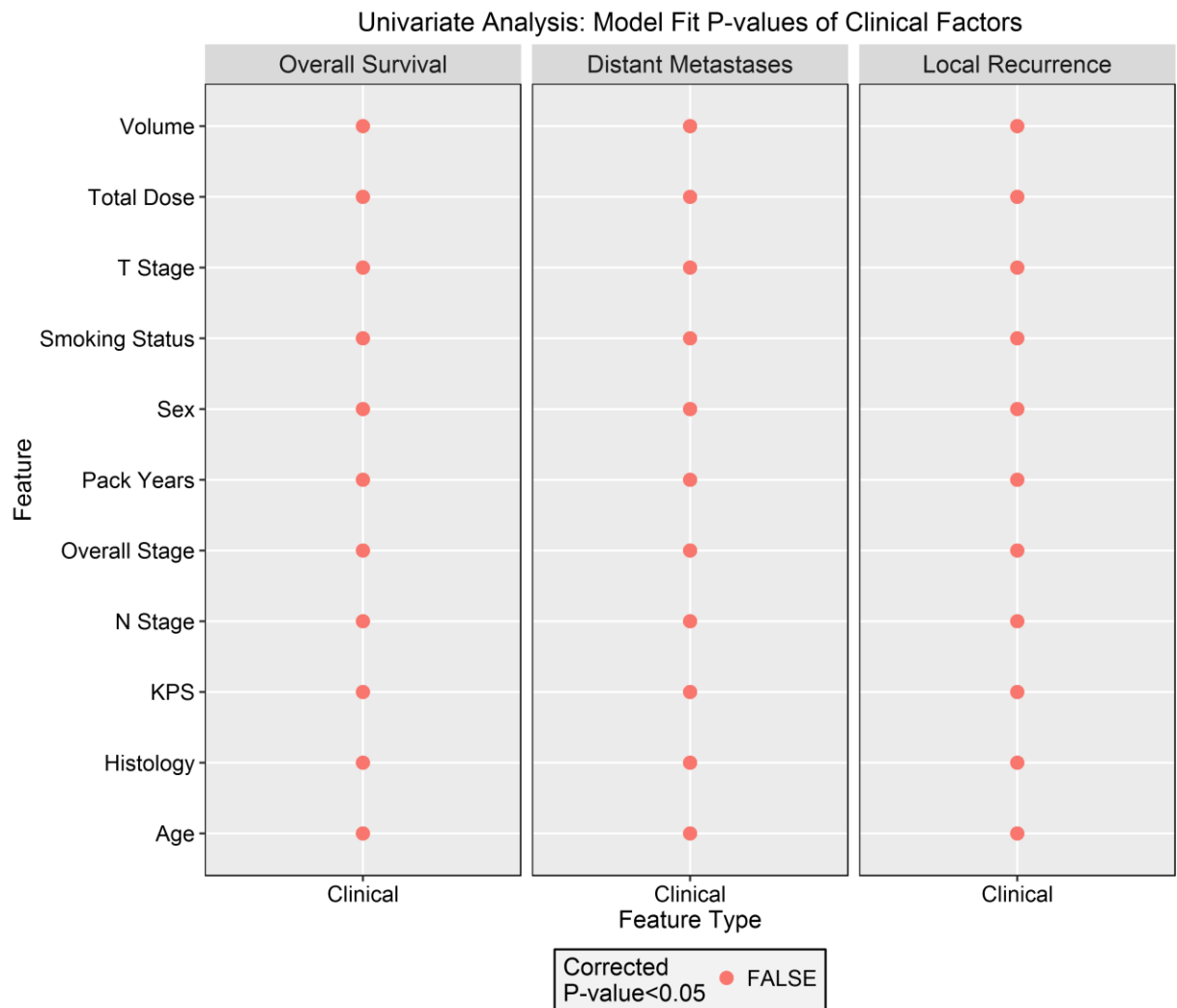


Figure 5.7: The Benjamini-Hochberg corrected p-values for the log-likelihood ratio of each univariate model using one of the clinical factors. None of the univariate models was significantly better than the null model.

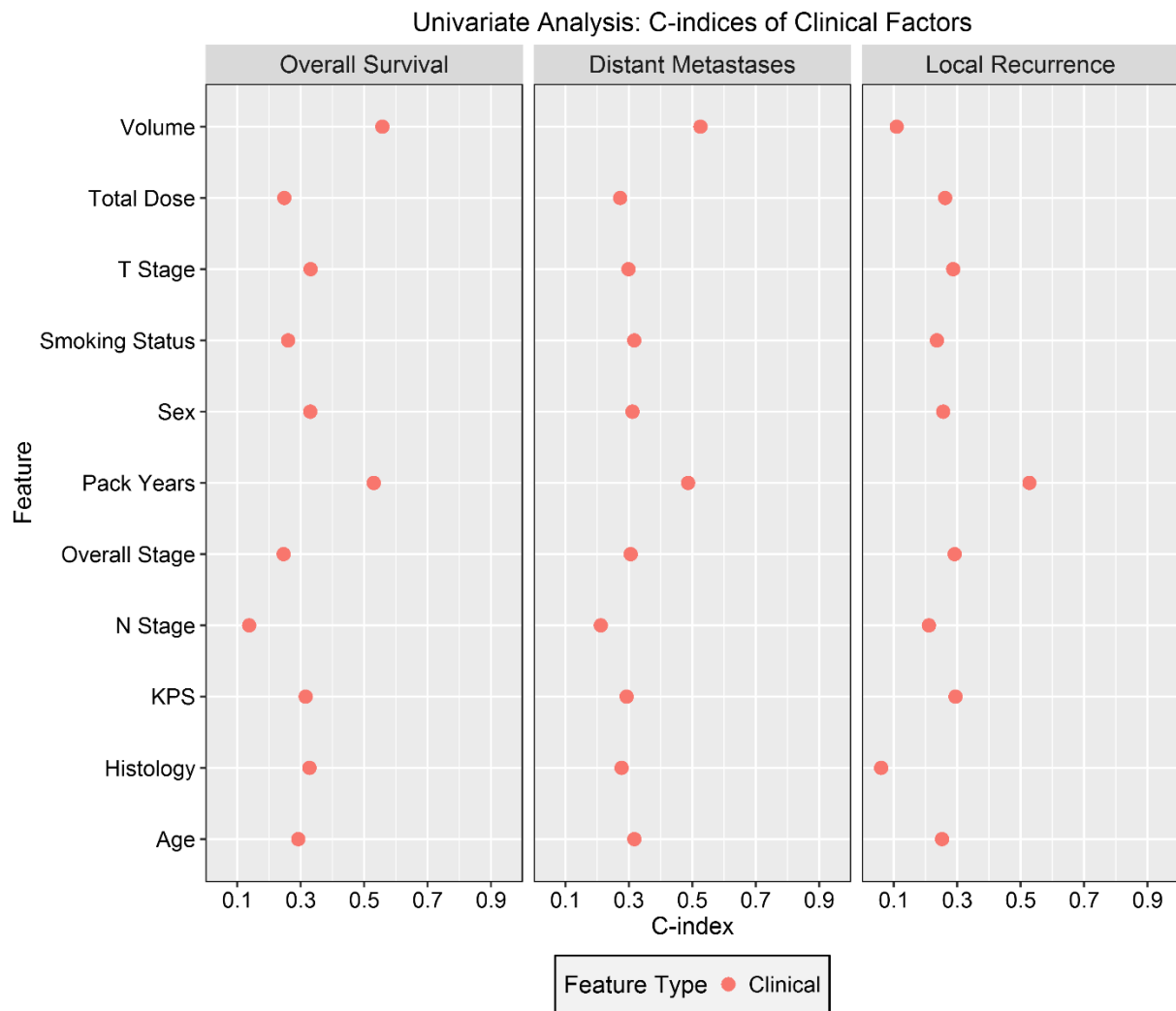


Figure 5.8: C-indices for univariate clinical models calculated using a LOOCV loop to generate patient-specific predictions for three outcomes. The overwhelming majority of the c-indices were below 0.5 and the highest c-index was only 0.56 for overall survival.

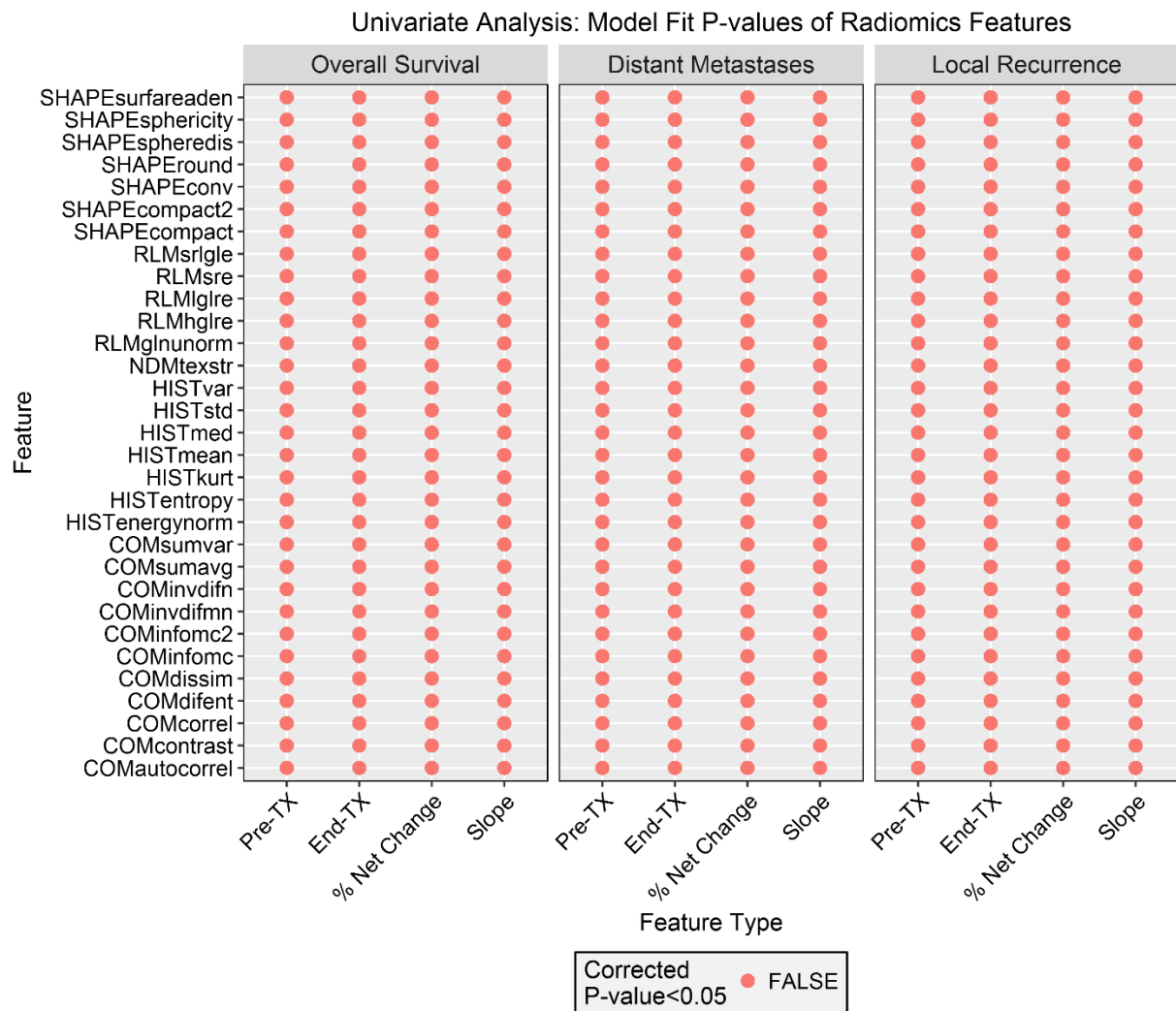


Figure 5.9: The Bonferroni corrected p-values for the log-likelihood ratio of each univariate model using four different versions of each radiomics feature. None of the univariate models was significantly better than the null model after multiplicity correction.

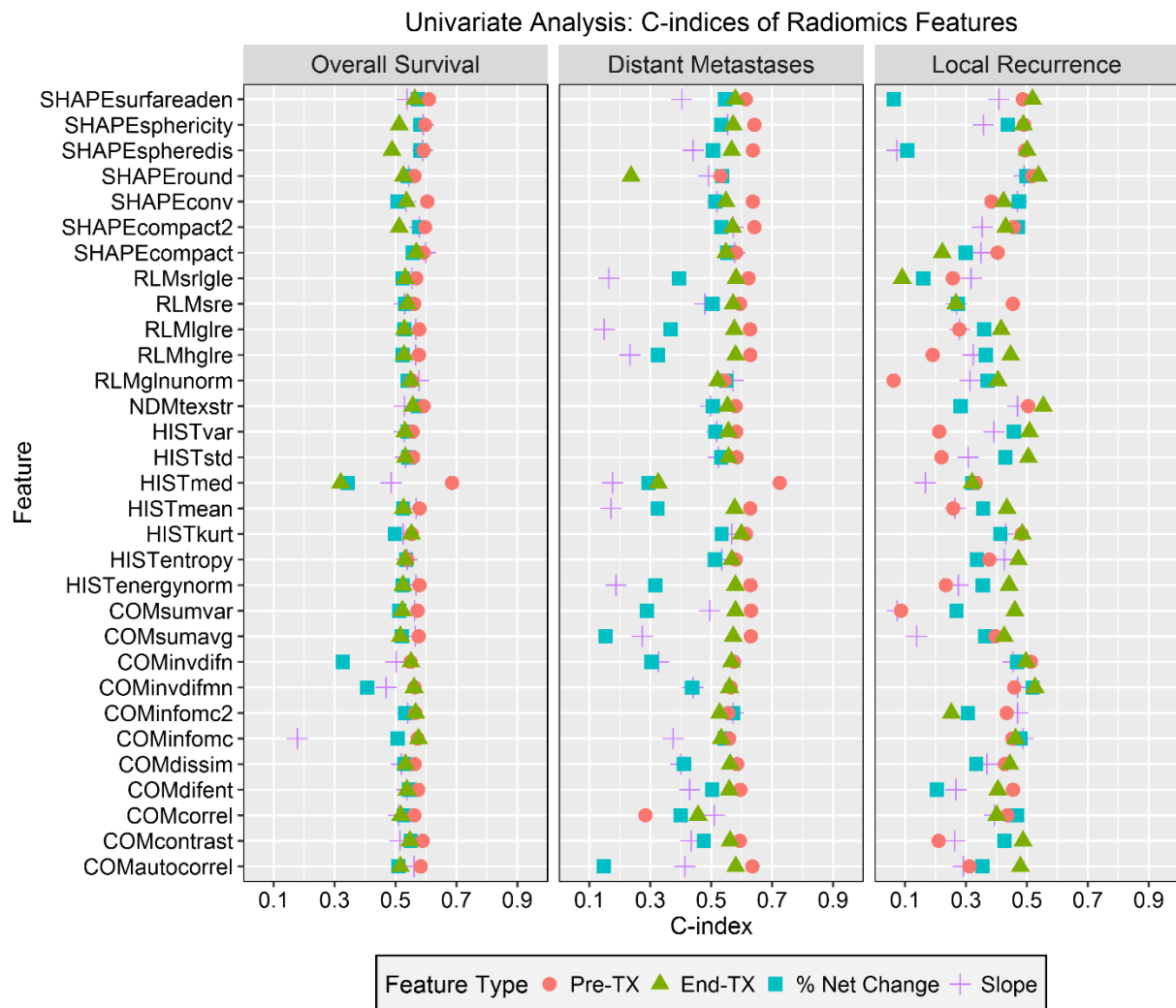


Figure 5.10: C-indices for univariate radiomics models calculated using a LOOCV loop to generate patient-specific predictions for three outcomes. The maximum c-indices for overall survival and distant metastases were 0.69 and 0.72 respectively and were both calculated using the median from the histogram measured at pre-treatment. For local-regional recurrence the highest c-index was 0.55 calculated using texture strength from the neighborhood difference matrix measured at the end of treatment.

## Multivariate analysis

The final results of the multivariate analysis are summarized in Table 5.3 for all three outcomes and all three models. The number of times each clinical factor and radiomics feature was selected in the first LOOCV is described in Table 5.4 for each outcome.

Table 5.3: Final comparison of the three models for each outcome. For the c-indices and log-rank test p-values, a value of NA indicates that no extra covariates were selected for this model and thus the value cannot be evaluated. Similarly, a value of NA for the log-likelihood ratio p-value implies that the two models being compared had the same covariates and thus the log-likelihood ratio cannot be computed. Model 1 is the model with only clinical factors. Model 2 is the model with clinical factors and pretreatment features. Model 3 is the model with clinical factors, pretreatment features, and delta-radiomics features. \*significant at  $p < 0.05$ ; \*\* significant at  $p < 0.005$ ; \*\*\* significant at  $p < 0.001$ .

Outcome	Overall Survival	Distant Metastases	Local-Regional Recurrence
C-index			
Model 1	0.597	0.539	NA
Model 2	0.672	0.632	NA
Model 3	0.675	NA	0.558
Log-likelihood ratio p-value			
Model 1 vs model 2	$4.20 \times 10^{-5***}$	$4.87 \times 10^{-4***}$	NA
Model 1 vs model 3	$2.10 \times 10^{-5***}$	NA	NA
Model 2 vs model 3	0.020*	NA	NA
Log-rank test p-value			
Model 1	$5.27 \times 10^{-3} ***$	0.38	NA
Model 2	$2.40 \times 10^{-6***}$	$1.56 \times 10^{-3***}$	NA
Model 3	$1.30 \times 10^{-5***}$	NA	0.0269*

Table 5.4: The number of times each clinical factor and radiomics feature was selected in the first leave-one-out cross validation for each outcome. Features included in the final model have their model coefficients included in the table. Abbreviations: LOOCV, leave one out cross validation; SCC, squamous cell carcinoma; KPS, Karnofsky Performance Status; WK0, week 0 (i.e., pretreatment)

Model	Feature Type	Covariates	No. of times selected in LOOCV	Included in final model	Model Coefficient
Overall Survival	Clinical Factors	T stage (T stage 3-4)	107	Yes	-0.933
		Sex (male)	107	Yes	0.866
		Histology (SCC)	107	Yes	0.772
		Total radiation dose (>70 Gy)	107	Yes	-0.652
		KPS	1	No	
		Pack years	1	No	
	WK0 Features	SHAPEcompact2.Week0	107	Yes	0.546
	Delta Features	RLMglnu.Slope	96	Yes	-0.240
		NDMtexstr.Slope	67	Yes	-0.330
		HISTkurt.netPercentChange	26	No	
		NDMtexstr.WeekLast	13	No	
		HistKurt.WeekLast	3	No	
Distant Metastases	Clinical Factors	Age ( $\geq 65$ years)	93	Yes	-0.467
		Sex (male)	86	Yes	0.492
		Overall stage ( $\geq$ IIIB)	61	Yes	0.552
		T stage (T stage 3-4)	59	Yes	-0.490
		Smoking status (former)	58	Yes	-0.062
		Smoking status (current)	58	Yes	0.060
		Total radiation dose	46	No	
		KPS	11	No	
	WK0 Features	SHAPEcompact2.Week0	105	Yes	0.585
		COMsumvar.Week0	13	No	
		SHAPEcompact.Week0	1	No	
		HISTkurt.Week0	1	No	

		COMautocorrel.Week0	1	No	
Delta		COMcorrel.netPercentChange	7	No	
Features		COMinfomc2.WeekLast	6	No	
		COMinvdifmn.WeekLast	2	No	
		COMinvdifmn.Slope	1	No	
		COMinvdifmn.netPercentChange	1	No	
Local-regional	Clinical	Smoking status (former)	19	No	
Recurrence	Factors	Smoking status (current)	19	No	
		KPS	15	No	
		Pack years	3	No	
	WK0	NDMtexstr.Week0	17	No	
	Features	SHAPEspheredis.Week0	1	No	
Delta		NDMtexstr.WeekLast	88	Yes	-0.517
Features		SHAPEround.netpercentChange	13	No	
			1	No	
		HISTkurt.WeekLast			

For overall survival, the clinical factors included in the final model were T stage, patient sex, tumor histology, and total radiation dose. The pretreatment feature that was included was compactness2 from the shape category. The delta-radiomics features that were included were the slopes in grey-level non-uniformity from the run-length matrix and texture strength calculated from the neighborhood difference matrix. Adding the single selected pretreatment feature compactness2 increased the c-index from 0.597 to 0.672 for overall survival. The log-likelihood ratio between these two models was significant. However, further addition of delta-radiomics features made a negligible difference to the c-index and did not substantially affect the patient stratification by the Kaplan-Meier curves as can be seen in Figure 5.11, Figure 5.12, and Figure 5.13. The log-likelihood ratio between model 2 (with clinical factors and pretreatment radiomics features) and model 3 (with clinical factors, pretreatment radiomics features, and delta-radiomics features) was significant, indicating an improved fit.

For distant metastases, no delta-radiomics features were included in the final model. The final clinical factors included in the model were tumor T stage, overall disease stage, and patient age, sex, and smoking status. Adding a pretreatment feature, compactness2 from the shape category, did result in an increase in the c-index from 0.539 to 0.632. The log-likelihood ratio between model 1 (clinical factors only) and model 2 (clinical factors and pretreatment radiomics features) was highly significant. Furthermore, patient stratification was significant when the pretreatment radiomics feature was added, while it was not significant for the purely clinical model, see Figure 5.14 and Figure 5.15.

For local-regional recurrence, no clinical factors or pretreatment radiomics features were selected in more than half of the LOOCV iterations. As a result, none were considered high-performing or were available for use in the final models for local-regional recurrence. However, the delta-radiomics feature texture strength from the neighborhood difference matrix measured at the end of treatment was selected in a majority of the LOOCV iterations. As a result, only the model including delta-radiomics features was built. This univariate model

resulted in a low value for the c-index (0.558) but a statistically significant stratification of the patients (p-value = 0.0269) as can be seen in Figure 5.16. In lieu of calculating the log-likelihood ratio between this model and model 1 (clinical features only) or model 3 (clinical, pretreatment radiomics, and delta-radiomics factors), the log-likelihood ratio between this model and the null model was calculated (p-value = 0.0725).

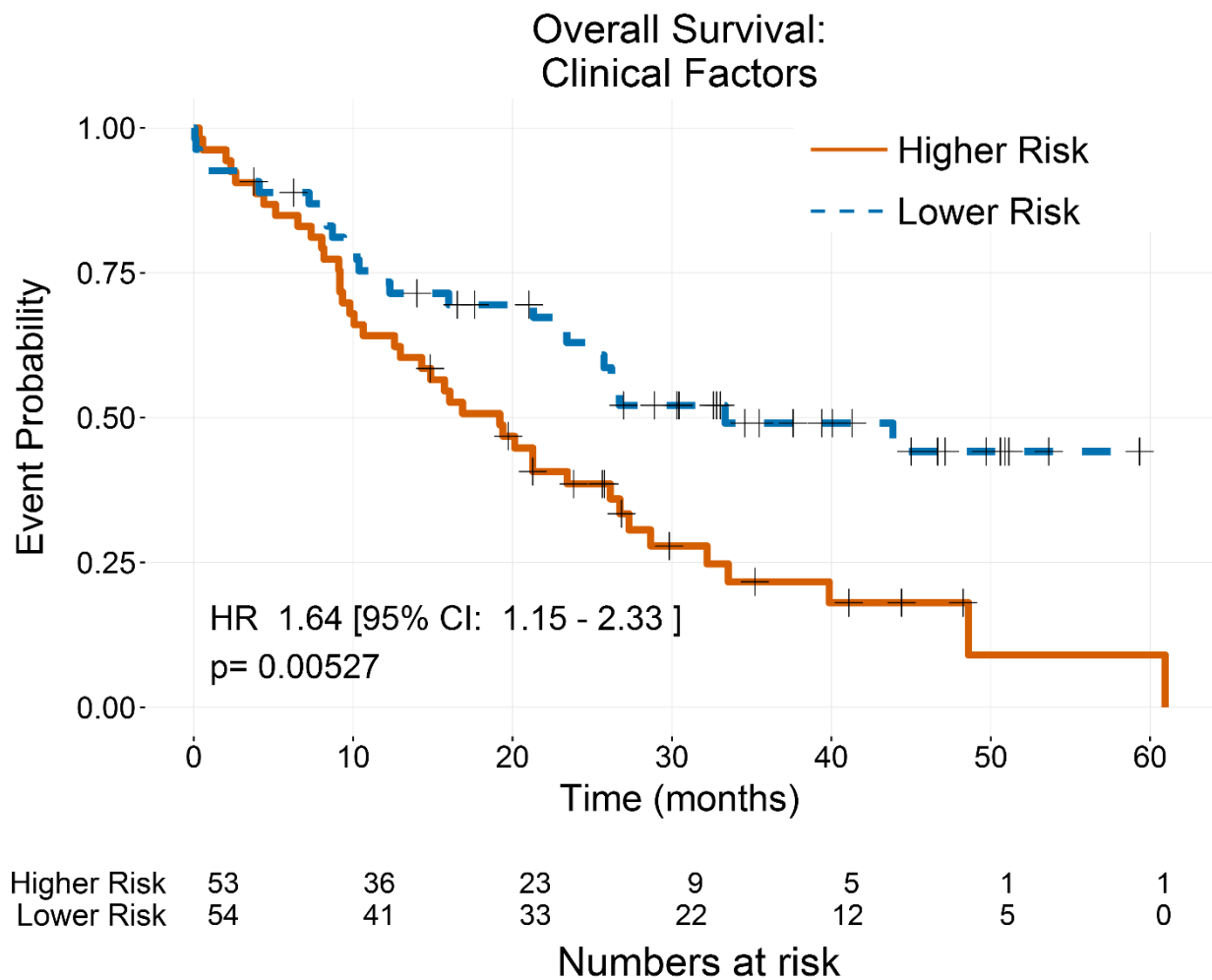


Figure 5.11: Kaplan-Meier curves for overall survival using clinical factors alone to generate Cox proportional hazards models. Models were built within a LOOCV loop, and on each iteration of the loop, the model was used to generate a prediction for the left out patient. Patients were classified as high or low risk based on if their prediction was above or below the median prediction value. The p-value for the log-rank test was less than 0.05 and thus the stratification was significant.

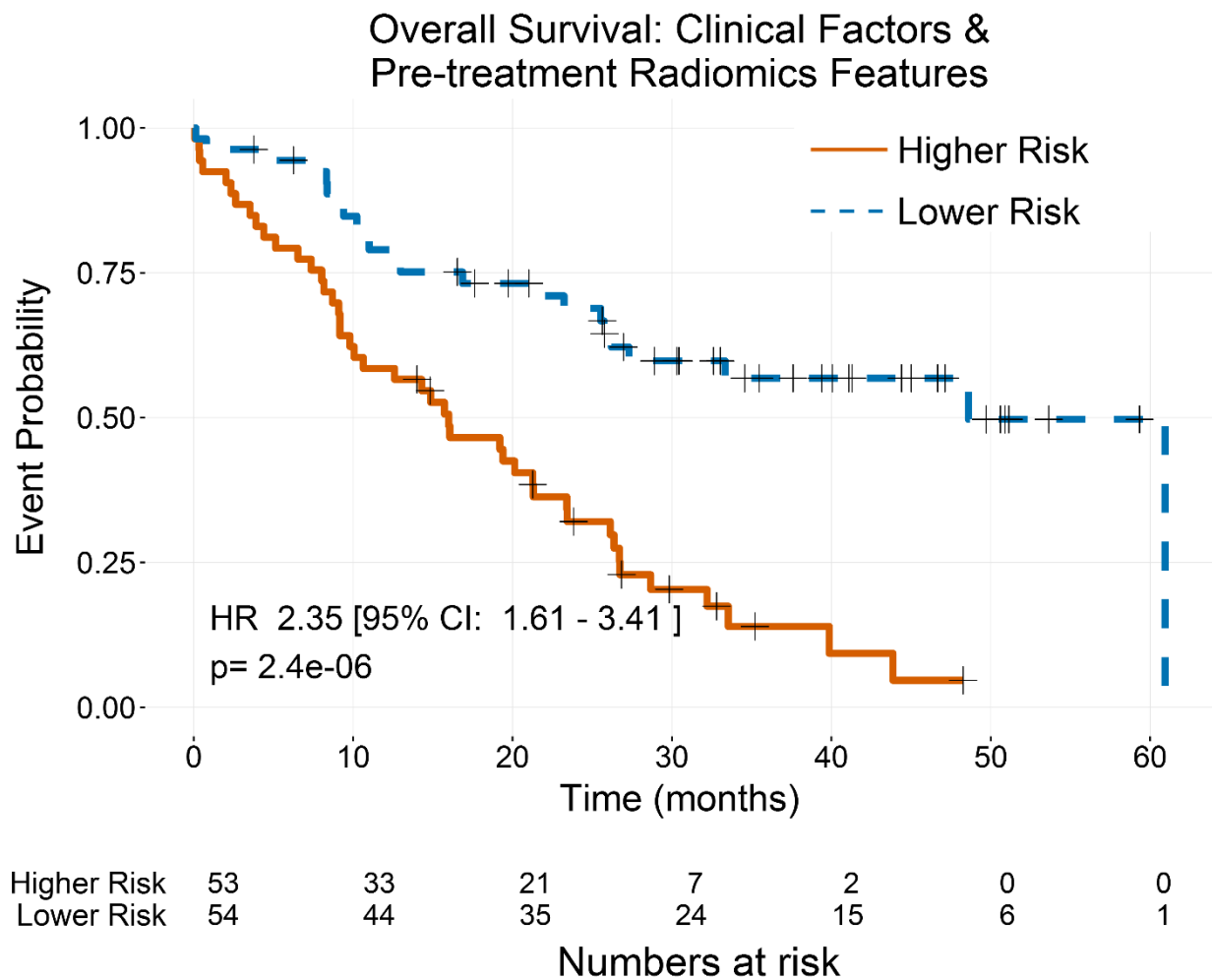


Figure 5.12: Kaplan-Meier curves for overall survival using clinical factors and pre-treatment radiomics features to generate Cox proportional hazards models. Models were built within a LOOCV loop, and on each iteration of the loop, the model was used to generate a prediction for the left out patient. Patients were classified as high or low risk based on if their prediction was above or below the median prediction value. The p-value for the log-rank test was less than 0.05 and thus the stratification was significant.

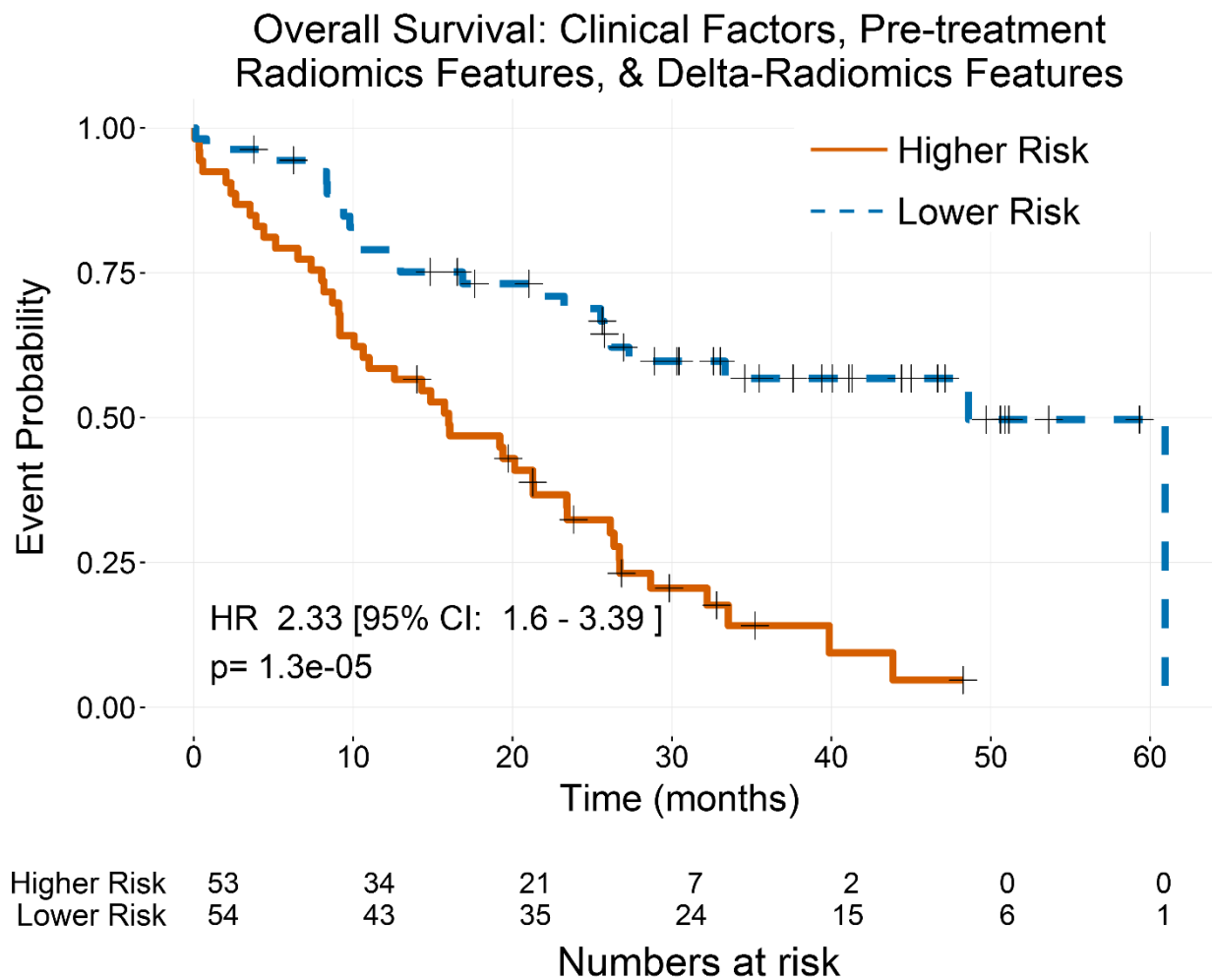


Figure 5.13: Kaplan-Meier curves for overall survival using clinical factors, pre-treatment features, and delta-radiomics features to generate Cox proportional hazards models. Models were built within a LOOCV loop, and on each iteration of the loop, the model was used to generate a prediction for the left out patient. Patients were classified as high or low risk based on if their prediction was above or below the median prediction value. The p-value for the log-rank test was less than 0.05 and thus the stratification was significant.

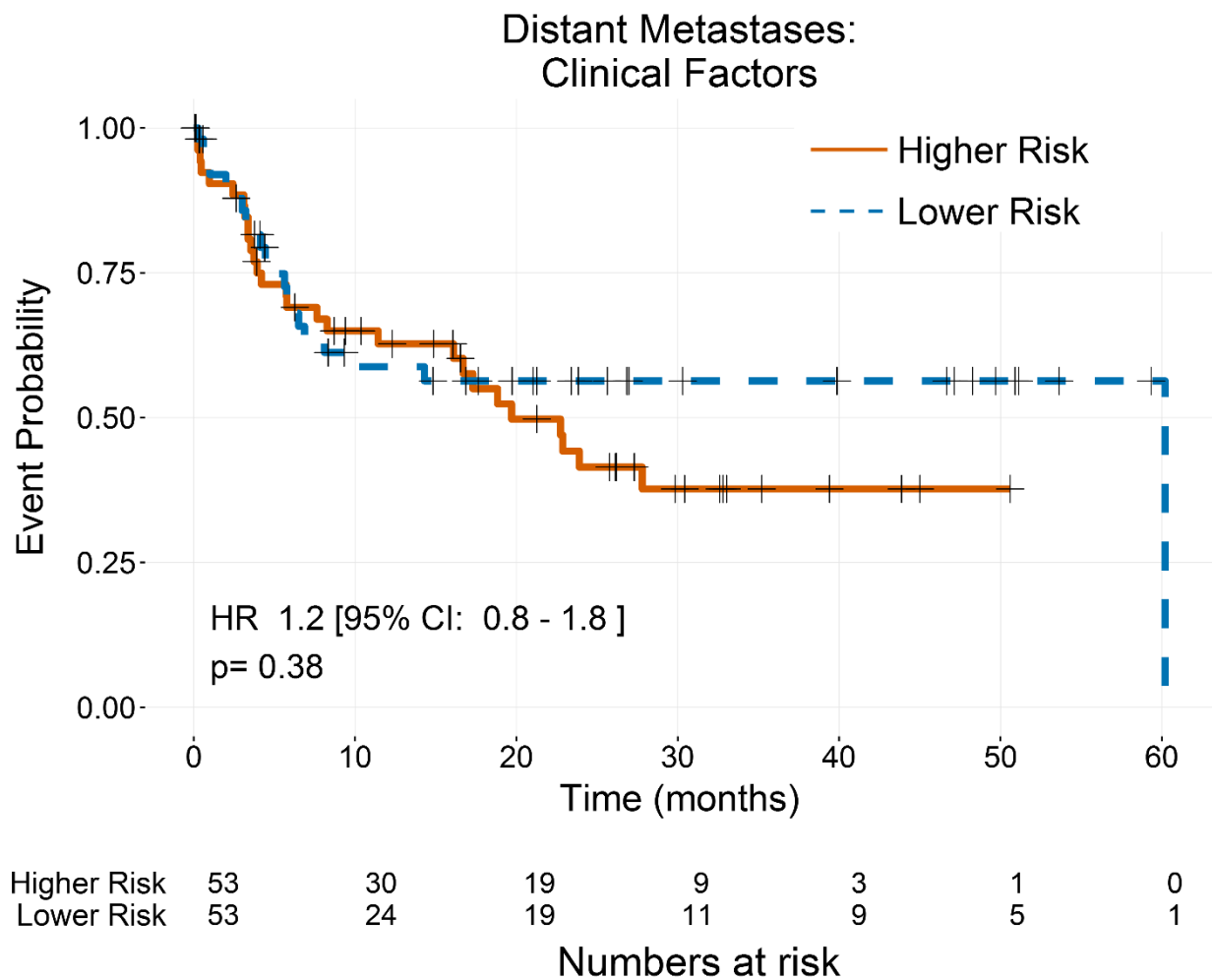


Figure 5.14: Kaplan-Meier curves for distant metastases using clinical factors alone to generate Cox proportional hazards models. Models were built within a LOOCV loop, and on each iteration of the loop, the model was used to generate a prediction for the left out patient. Patients were classified as high or low risk based on if their prediction was above or below the median prediction value. The p-value for the log-rank test was not less than 0.05 and thus the stratification was not significant.

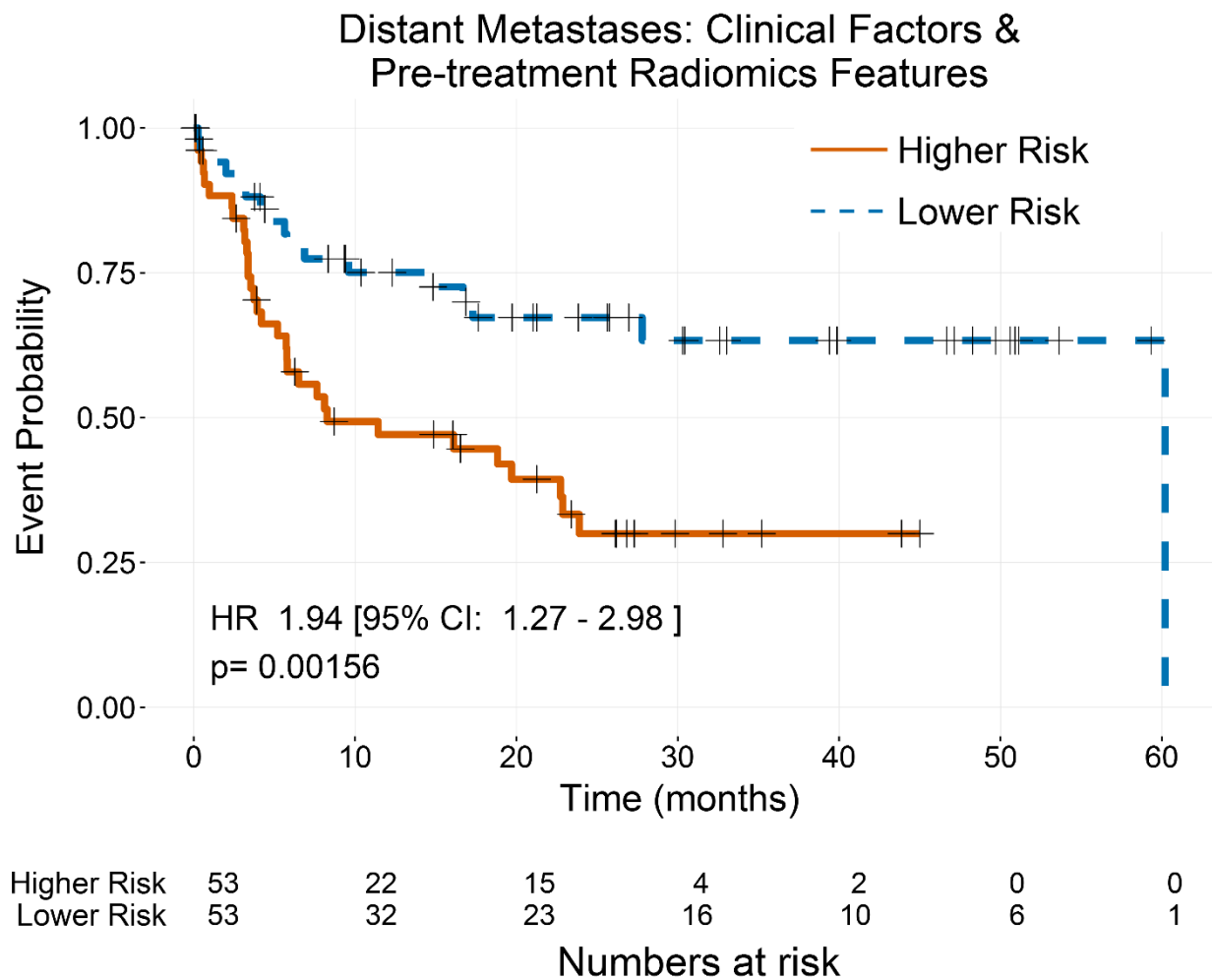


Figure 5.15: Kaplan-Meier curves for distant metastases using clinical factors and pre-treatment features to generate Cox proportional hazards models. Models were built within a LOOCV loop, and on each iteration of the loop, the model was used to generate a prediction for the left out patient. Patients were classified as high or low risk based on if their prediction was above or below the median prediction value. The p-value for the log-rank test was less than 0.05 and thus the stratification was significant.

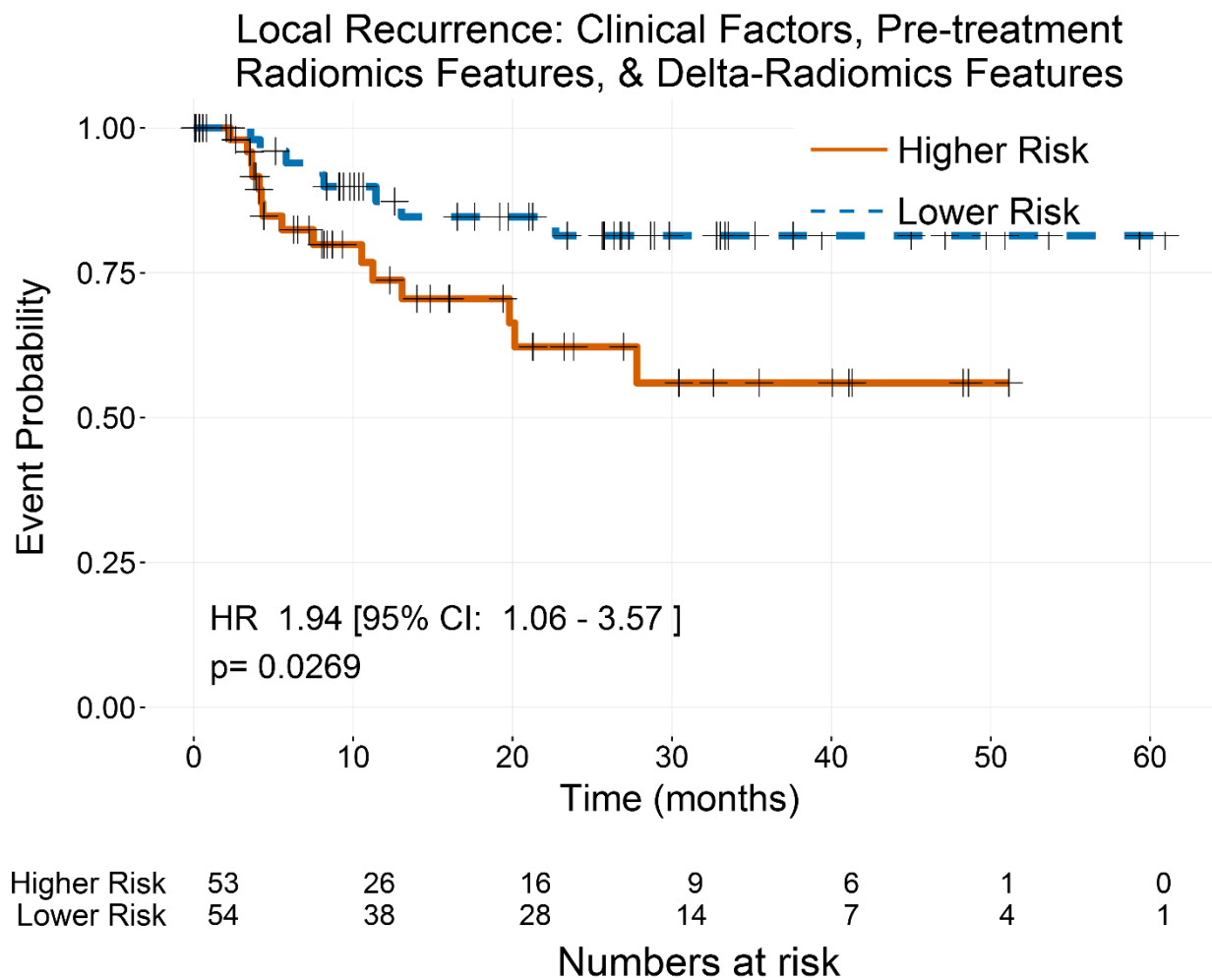


Figure 5.16: Kaplan-Meier curves for local-regional recurrence using clinical factors, pre-treatment features, and delta-radiomics features to generate Cox proportional hazards models. Models were built within a LOOCV loop, and on each iteration of the loop, the model was used to generate a prediction for the left out patient. Patients were classified as high or low risk based on if their prediction was above or below the median prediction value. The p-value for the log-rank test was less than 0.05 and thus the stratification was significant.

## Discussion

The first part of this analysis was to select the best image-preprocessing technique for each feature and to exclude features that could not meet two requirements (significance in univariate models and no significance between CT scanner models). This step resulted in feature-specific image preprocessing such as having certain features from the co-occurrence matrix calculated with 8 bit depth resampling and thus from a co-occurrence matrix with bins of 16 HU while other co-occurrence matrix features were calculated with only smoothing and thus from a co-occurrence matrix with bins of 1 HU. Previous studies using features from texture matrices have always calculated the features from a matrix with the same parameters. We chose to use this feature-specific selection process instead because it is not known which parameters for the texture matrices or for a specific feature are best. Additionally, it is possible that different features are most useful for measuring tumor heterogeneity at different intensity scales which is evaluated by this method. Over half of the features did not pass the minimum requirements for any of the image preprocessing techniques and thus the feature set was also reduced from 65 features to 31. This step was important because in the subsequent univariate analysis a heavy penalization for multiplicity correction was implemented and so reducing the number of features to those that could plausibly be useful aided in reducing the strength of that penalization.

In addition to passing the feature-specific image preprocessing criteria, all 31 of the final features also passed two tests that evaluated whether they changed consistently during treatment. First the p-values of the log rank test comparing univariate linear mixed effects models to the null model for each radiomics feature were evaluated. The fact that all 31 of these features passed this test suggests that the values change in a general positive or negative direction for each feature. This conclusion was supported by the results of the t-tests that compared the mean of the slopes and relative net changes in the features to a normal distribution with a mean at 0. If a feature had failed these tests it would have been either

because it did not change at all for any of the patients (and is probably measuring image noise) or because for some patients it increased while for some it decreased and the net result for the cohort was 0. We wanted to exclude features that changed in both directions because that would indicate the changes were unlikely to be due from treatment and we were specifically looking for therapy-induced changes in the features.

Two different metrics were then evaluated for the univariate analysis: the significance of a univariate cox proportional hazards model fit and the c-index. For the clinical factors, pre-treatment radiomics features, and delta-radiomics features the results were largely negative. No clinical factor or radiomics feature had a significant univariate fit after multiplicity correction. This suggests the overwhelming majority of covariates we tested are not prognostic for any of the endpoints being measured. This was supported by the fact that when the prognostic ability of the features was tested individually using the c-index, most of the clinical factors and radiomics features had c-indices less than 0.5. However for the clinical factors, the number of pack years did have a c-index higher than 0.5 for all three end points, while for the radiomics features the best performing feature was the median from the histogram which had a c-index close to 0.7 for both overall survival and distant metastases when the value at the beginning of treatment was used. In general the best version of any selected radiomics feature was the value at pre-treatment for overall survival or distant metastases. This may have been due to our initial selection criteria for the features which only selected features that were prognostic for overall survival when measured at pre-treatment (p-value<0.1 without multiplicity correction). Interestingly for local recurrence, the best version of the features that had at least one c-index greater than 0.5 was the values at the end of treatment and this was the only feature that got selected in the later multivariate analysis. While the results of the univariate analysis were not strongly supportive of the prognostic potential of the delta-radiomics features, there was a possibility that when used in conjunction with clinical factors or radiomics features measured at

pre-treatment they could add an incremental improvement and so were still included in the multivariate analysis

While the inclusion of delta-radiomics features had a statistically significant impact on the overall likelihood of a model for overall survival compared to a model with only clinical and pretreatment radiomics features, the impact on the model's prognostic abilities was negligible. For distant metastases, no delta-radiomics features were selected in the final round of model building. This suggests that delta-radiomics features do not offer substantially new prognostic information for these outcomes though they were still prognostic for overall survival. The same pretreatment radiomics feature, compactness2, was selected for both the overall survival and time to distant metastases models and improved their prognostic potential. For both overall survival and distant metastases, the coefficient for this feature was positive, meaning a patient had a higher predicted risk of experiencing the outcome if the value for compactness2 from their ROI was relatively large. This feature was related to the volume and shape of the tumor ROI, i.e. how spiculated it may appear. The feature values were also affected by the tumor location, since a tumor attached to the chest wall was contoured with at least one smooth side compared to a tumor surrounded by lung which ranged anywhere between fully smooth or fully spiculated. Compactness 2 was also found to be predictive in a radiomics study by Aerts et al where it was included as part of a four feature radiomics signature<sup>23</sup>. This study is unique in that it demonstrated that compactness 2 added significant new information to a variety of clinical factors already routinely obtained, as opposed to only TNM staging and tumor volume. In this study, the clinical model was built first and then radiomics features were added to it rather than building a purely radiomics model and assessing its capabilities. This is important because the introduction of radiomics features into a routine clinical workflow is unlikely to be accepted unless models built using radiomics features outperform models built using only routinely acquired clinical factors.

Interestingly, in the models for local-regional recurrence, the only covariate that was predictive for outcomes was a radiomics feature, texture strength from the neighborhood difference matrix, measured from images acquired during the last week of treatment. This feature was designed to quantify whether an image has clear, perceivable characteristics that can be considered as texture and the overall strength of that signal<sup>14</sup>. Further work is needed to identify what this feature may represent in the context of NSCLC tumor analysis. This result may be evidence that, although it is not possible to predict local-regional recurrence prior to treatment, the state of the tumor at the end of the treatment can be assessed using radiomics.

One possible cause of the poor selection of delta-radiomics features in the models may be due to the initial feature preselection process. The full feature set was first reduced to features whose pretreatment values were at least prognostic in univariate models for overall survival. It is possible that the results would differ if this requirement was changed to instead select for delta-radiomics features that are significant in univariate models. The original requirement was chosen for two reasons: first, because several publications have shown that pretreatment radiomics features have informative value and thus changes in the features that are already prognostic may reflect actual biological changes in the tumor, and second, if model building was limited to delta-radiomics features that were significant in univariate analyses the results could be biased and overly optimistic.

One limitation of this study was the lack of a dataset for independent model validation due to the fact that patients are not routinely imaged weekly during their treatment. This limitation was mitigated by using cross validation, which has been shown to be an effective method for creating unbiased patient-specific predictions<sup>83,84</sup>. Another limitation of this study was that the median predicted value was used as the cut-off point for high- and low-risk patients. This is not an optimized approach, and it is very likely that a different model-specific value would yield different results. However, testing multiple cutoffs to find the best one without an independent validation dataset to test it in has been repeatedly shown to yield overly

optimistic results<sup>68,85,86</sup>. By using the median, this source of bias is avoided and the conclusions remained conservative. Lastly, because the images used in this analysis were non-contrast CT images, vessels passing through the lesion could not be segmented from the contour. Thus the contours for the tumor ROIs may contain vasculature along with the solid tumor component we are interested in. The inclusion of vasculature in the tumor ROIs may affect the radiomics features and the calculated tumor volume.

Radiomics is in some ways fundamentally limited because the features are not inherently descriptive. This is in contrast to clinical covariates which, when selected in prognostic models, lend themselves to hypotheses, e.g., age is likely to affect survival because a younger person is statistically likelier to live longer than an older person. For radiomics features, this type of reasoning is difficult and instead new studies must be undertaken to correlate feature values with biological characteristics such as genetic mutations. Radiomics features also suffer from lack of robustness, as they have been demonstrated to vary with imaging equipment, ROI contouring, and imaging parameters. Thus the implementation of radiomics features in a clinical setting would require substantial effort to standardize both imaging and measurement parameters. This study identified two features, compactness2 and texture strength, which may be of clinical significance. The first step in determining their robustness will be to examine the impact of segmentation on both features' values and prognostic potentials. This is especially critical for compactness2 since it is a shape based feature and thus could be substantially impacted by segmentation

## **Conclusions**

This study found evidence that radiomics features change during the course of radiation therapy for NSCLC. However, these changes in features did not significantly outperform features measured before treatment in multivariate models for overall survival and distant metastases. Thus it may be more important to focus efforts on improving the standardization of features measured before treatment and identifying a biological or molecular explanation for

their predictive values. One radiomics feature measured at the end of treatment did outperform both clinical factors and pretreatment radiomics features for prediction of local-regional recurrence. This feature, texture-strength, could become an indicator for tumor response since it was only prognostic when measured at the end of treatment. Despite the fact that this study did not find strong evidence supporting the prognostic potential of delta-radiomics features, the results of this study are important because the potential of tracking radiomics features throughout treatment for NSCLC was investigated. Furthermore, while other studies have used delta-radiomics features for other treatment sites or for normal tissue toxicity, they have used only the relative net change in their models<sup>37,39,87</sup>. This study included both the slope of a linear regression for the features of each patient, which may be less susceptible to noise than the relative net change, and the feature values at the end of treatment, which may reflect tumor response.

# Chapter 6 : Reproducibility of Radiomics Features from CBCT Images

A substantial portion of this chapter is written or based on the following publication:

Fave, X, Mackin, D, Yang, J, Zhang, J, Fried, D, Balter, P, Followill, D. , Gomez, D, Jones, AK, Stingo, F, Fontenot, J, and Court, L. Can radiomics features be reproducibly measured from CBCT images for patients with non-small cell lung cancer? Medical Physics doi: 10.1118/1.4934826. Volume 42, Issue 12, pages 6784-6797. 2015. ©John Wiley and Sons.

The permission for reuse of this material was obtained from John Wiley and Sons ©.

In this chapter we describe the results for Specific Aim 3: Identify which radiomics features can be reproducibly measured from CBCT images. Our working hypothesis for this aim was that a subset of radiomics features can be identified that are robust to the increased scatter and motion present during CBCT imaging.

## Introduction

In future studies, it would be valuable to measure delta-radiomics features through treatment for larger patient cohorts. While it may be possible for another IRB-approved study to acquire weekly 4DCT images, it would be far easier to use the low-dose cone-beam CT (CBCT) images that are already routinely acquired for patient setup verification. Using these images would allow for the collection of larger patient cohorts, as well as more frequent imaging because CBCT images are typically acquired before every treatment fraction or at least weekly during treatment. However, CBCT images are known to have worse image quality than regular CT due to their use of a flat-panel detector and larger cone angle which results in larger amounts of scatter being detected. Additionally, CBCT imaging requires a longer scan time than CT images (~1 minute) and patients are not usually asked to hold their breath. As a

result, motion artifacts are more likely in CBCT images. This is of special concern for NSCLC patients whose tumors may move with respiration.

The purpose of this study was to individually investigate sources of error in CBCT images (different imaging protocols, scatter, and motion) on a typical set of radiomics features in order to determine the features' susceptibility to each. Once these different impacts have been characterized, guidelines can be developed for obtaining CBCT images that will render consistent radiomics features for use in future studies.

## Methods

### Patient Test-Retest CBCT Images

CBCT images were acquired from a subset of patients that were part of the main cohort discussed in Chapter 3. We retrospectively searched the imaging history for each of the patients in the cohort for repeat CBCT images. From this search, only ten patients were found who had two sets of CBCT images obtained within 15 minutes of each other using the same imager. This retrospective analysis was approved by the University of Texas M.D. Anderson Cancer Center IRB with a waiver of informed consent. Characteristics for these 10 patients are tabulated in Table 6.1.

Table 6.1: Clinical characteristics for the test-retest CBCT patient population.

Characteristic	Number of Patients	Percent of Patients
n, number of patients	10	NA
Median age (range)	65.5 (49-80)	NA
Median GTV volume (range)	77.5 cm <sup>3</sup> (17-315 cm <sup>3</sup> )	NA
Gender		
Male	3	30%
Female	7	70%
Tumor stage		
II	1	10%
III	9	90%
Tumor histology		
Squamous cell carcinoma	2	20%
Adenocarcinoma/other	8	80%

All CBCT patient images were acquired using the thoracic imaging protocol on a Varian linac: peak tube voltage of 110 kVp, tube current of 20 mA, and exposure time (total pulsed beam-on time) of 7-14 sec. Images were reconstructed as a 512 x 512 grid with pixel dimensions of 0.8 mm and a 2.5 mm slice thickness. For each of these 10 patients, we deformably transferred the GTV contour from the treatment plan to their two CBCT image sets using an in-house deformation image registration software, CT-assisted targeting<sup>47,48</sup>. The images and contours were imported into the IBEX radiomics software to extract the values for the imaging features.

The concordance correlation coefficient (CCC) was calculated for each feature using this test-retest image set. Features whose CCC<0.9 were not considered reproducible and were excluded from the rest of the analysis<sup>43,88</sup>. This test removed features that were not reproducible even when measured in images obtained from the same patient within 15 minutes using the same imager. The cutoff value of 0.9 was chosen based on the recommended criteria of McBride et al that considered a correlation of 0.9 to reflect moderate strength-of-agreement and all correlations<0.9 to be poor. The Spearman correlation coefficient ( $r_s$ ) was also calculated between each feature and the region-of-interest (ROI) volume. The Spearman coefficient was calculated for the test and retest image volumes individually. Any feature with  $r_s>0.85$  in both image sets was excluded from the rest of the analysis in order to remove features whose CCC was high only because of that feature's strong correlation with volume<sup>89,90</sup>. The cutoff value of 0.85 is within the range of values of that has been cited in the literature as representative of strong correlations such as Zou et al<sup>89</sup> who considered any  $r_s>0.8$  to be strongly correlated and Mukaka et al<sup>90</sup> who interpreted only  $r_s>0.9$  to have very high correlations. Because the purpose of this step was to reduce the likelihood of false positives in the following experiments where each feature was independently examined for its reproducibility under different conditions and not to establish an explicit volume-dependence, a relatively high cutoff was selected to remove only strong relationships.

For the remaining features, the mean intra-patient test-retest differences were calculated with Equation 6.1. These values were used as benchmarks for reproducibility in the subsequent phantom studies. This criterion was used because the phantom was expected to change substantially less from scan to scan than a patient. In Equation 6.1,  $N_{pats}$  is the number of patients and  $x_{n,t}$  and  $x_{n,r}$  are the  $n^{th}$  patient's test and retest values respectively.

$$\text{Mean inpatient difference} = \frac{\sum_{n=1}^{N_{pats}} |x_{n,t} - x_{n,r}|}{N_{pats}}$$

Equation 6.1

### Texture Phantom

The Credence Cartridge Radiomics phantom was used to evaluate the impact of different scanners, protocols, scatter levels, and amounts of motion on texture values extracted from CBCT images, Figure 6.1. This phantom was designed at our institution specifically for investigating the reproducibility of texture features<sup>40</sup>. The phantom is a hollow, acrylic, rectangular prism. Inside are cartridges of 10.1 x 10.1 x 3.2 cm<sup>3</sup> made of different materials: wood, dense cork, regular cork, shredded rubber, acrylic, and resin. The phantom also contains four 3D printed cartridges with tessellated hexagons of different sizes. A previous analysis of the phantom imaging characteristics conducted at our institution using CT scans demonstrated that the texture values extracted from the shredded rubber and dense cork cartridges were closest to values obtained from patient tumors<sup>40</sup>. For this reason, only these two materials were used in this analysis. An ROI of 6.0 x 6.0 x 2.0 cm<sup>3</sup> was positioned at the center of these two materials for feature extraction in each image, Figure 6.1. This size was used because it was close to the median size of the patient GTV volumes (77.5 cm<sup>3</sup>) and to avoid including the edges of the phantom in the ROI.

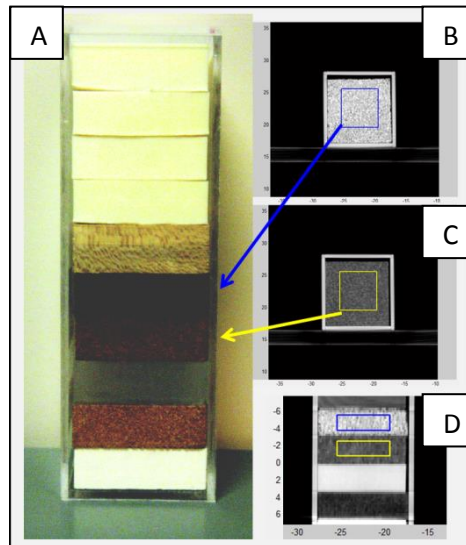


Figure 6.1: The Credence Cartridge Radiomics Phantom. (A) Photograph of the radiomics phantom used in this study and (B-D) CBCT images of the phantom with the ROIs used. Only the (B) shredded rubber and (C) dense cork cartridges were used for the current analysis.

### Texture Features & Pre-Processing

For this study, a comprehensive set of 68 radiomics features were initially selected, Table 6.2. Features were selected to cover the diverse range of features that have been used in previous texture feature studies using CT images of NSCLC<sup>23,26–28,43,91</sup>. Selected features included first-order descriptors from the intensity histogram; second-order features to describe spatial relationships in gray level intensities from the co-occurrence matrix<sup>12,23</sup>, run-length matrix<sup>13</sup>, and neighborhood gray-tone difference matrix<sup>14</sup>; and Laplacian of Gaussian (LoG) filtered features, which can highlight tumor characteristics not visible in the original image<sup>27,51</sup>.

Table 6.2: Features that were included in the analysis of CBCT feature reproducibility.

Histogram	Co-occurrence Matrix	Run Length Matrix	Neighborhood Difference Matrix	LoG Filtered Features
<u>without rescaling</u>	Autocorrelation	Gray-level	Busyness	<u>fine filter</u>
Max	Cluster	nonuniformity	Coarseness	Entropy
Mean	prominence	High gray-level	Complexity	Mean
Median	Cluster shade	run emphasis	Contrast	Standard deviation
Entropy	Cluster tendency	Long run		Uniformity
Energy	Contrast	emphasis		Kurtosis
Standard deviation	Correlation	Long run high		Skewness
Uniformity	Difference	gray-level		
Kurtosis	entropy	emphasis		
Skewness	Dissimilarity	Long run low		
Variance	Energy	gray-level		
	Entropy	emphasis		
<u>with rescaling</u>	Homogeneity	Low gray-level		<u>medium filter</u>
Max	Homogeneity2	run emphasis		Entropy
Mean	Information	Run length		Mean
Median	measure	nonuniformity		Standard deviation
Entropy	correlation	Run percentage		Uniformity
Energy	Information	Short run		Kurtosis
Standard deviation	measure	emphasis		Skewness
Uniformity	correlation2	Short run high		
Kurtosis	Inverse	gray-level		
Skewness	difference	emphasis		
Variance	moment norm			
	Inverse			
	difference norm			
	Inverse variance			
	Max probability			
	Sum average			
	Sum entropy			
	Sum variance			
	Variance			

The patient ROIs were pre-processed with a thresholding step to exclude air, bones, and normal lung tissue. Values less than -150 HU or greater than 200 HU were excluded for the patient images. For the motion phantom a lower threshold of -700 HU was used to ensure none of the surrounding lung-equivalent material was included in the ROI. Thresholds were not used for the texture phantom ROIs in order to ensure that all of the voxels within the ROI would be included.

All the images were also rescaled to 8-bit images before calculating the co-occurrence matrix, run-length matrix, and neighborhood difference matrix features; this was done to reduce the effect of noise on the texture features and prevent sparsely populated matrices from being produced. The histogram features were calculated both with and without 8-bit rescaling. The LoG features were calculated without the rescaling step because the Gaussian filter already acts to smooth the images and reduce noise. The LoG features were calculated at two different scales: a fine filter (fineFilt) with a window size of 5 and sigma of 1 and a medium filter (medFilt) with a window size of 7 and sigma of 1.5. Note this analysis was performed prior to the in-depth analysis of image preprocessing on features in Chapter 4 and thus feature-specific image-preprocessing was not used.

### Effect of Scanners

During the course of treatment, a patient may receive some of his or her dose fractions on a different linac than the one used for the first fraction, and thus the daily or weekly CBCT images would be acquired from separate machines. Additionally, images from different patients that are accumulated for a radiomics study are likely to come from different imagers. To determine whether these differences have an influence on the resulting texture values, the texture phantom was imaged with the CBCT imagers on 19 linacs, including 9 Elekta linacs and 10 Varian linacs. Two scans were acquired per machine: one with the default head protocol and one with the default thoracic protocol for that machine. The standard image reconstruction was used for all scans. The characteristics of these linacs and scans are described in Table

6.3 Each scan was classified as a Varian head scan (V-Head), Varian thorax scan (V-Thorax), Elekta head scan (E-Head), or Elekta thorax scan (E-Thorax). For each texture feature the absolute difference between values measured from every possible pair of scanners was calculated. These differences were then categorized by the types of scans being compared (e.g. V-Thorax scan vs. E-Head). The differences were compared individually to the mean intra-patient difference for each feature. If the difference between two scans was less than the mean intra-patient difference, the comparison passed and the feature was considered reproducible between those two scans. The mean intra-patient difference was used as the criteria, because it was assumed that the phantom would demonstrate substantially less variation than the patients when scanned under different conditions. The overall percentage of passing scans for each comparison category was recorded. A high percentage of passing scans implies that the feature is reproducible between that subset of linacs.

Table 6.3: Scan characteristics for the phantom CBCT images used in this study. \*For Elekta machines the exposure time represents the pulse length, for Varian machines the exposure time represents the full beam-on time.

ID	Manufacturer	Protocol	Image Size (pixels)	Pixel Size (mm)	Slice Thickness (mm)	Tube Voltage (kVp)	Exposure Time* (ms)	Tube Current (mA)
1	Elekta	Head	410	1.0	5.00	120	20	32
2	Elekta	Head	410	1.0	3.00	120	20	32
3	Elekta	Head	410	1.0	4.00	120	20	32
4	Elekta	Head	410	1.0	4.00	120	20	32
5	Elekta	Head	410	1.0	5.00	120	20	32
6	Elekta	Head	410	1.0	4.00	120	20	32
7	Elekta	Head	410	1.0	4.00	120	20	32
8	Elekta	Head	410	1.0	2.00	120	40	40
9	Elekta	Head	410	1.0	2.00	120	40	40
1	Elekta	Thorax	270	1.0	5.00	100	10	10
2	Elekta	Thorax	270	1.0	3.00	100	10	10
3	Elekta	Thorax	270	1.0	4.00	100	10	10
4	Elekta	Thorax	270	1.0	4.00	100	10	10
5	Elekta	Thorax	270	1.0	5.00	100	10	10
6	Elekta	Thorax	270	1.0	4.00	100	10	10
7	Elekta	Thorax	270	1.0	4.00	100	10	10
8	Elekta	Thorax	270	1.0	2.00	100	10	10
9	Elekta	Thorax	270	1.0	2.00	100	10	10
10	Varian	Head	512	0.5	2.50	100	7300	20
11	Varian	Head	512	0.5	2.50	100	7480	20
12	Varian	Head	384	0.7	2.50	100	7160	20
13	Varian	Head	512	0.5	2.50	100	7460	20
14	Varian	Head	512	0.5	2.50	100	7560	20
15	Varian	Head	512	0.5	2.50	100	7300	20
16	Varian	Head	512	0.5	2.50	110	13180	20
17	Varian	Head	512	0.5	1.98	100	7470	20
18	Varian	Head	512	0.5	1.98	100	7350	20
19	Varian	Head	512	0.5	1.98	100	7350	20
10	Varian	Thorax	512	0.9	2.50	110	13160	20
11	Varian	Thorax	512	0.9	2.50	110	13480	20
12	Varian	Thorax	384	1.2	2.50	110	12900	20
13	Varian	Thorax	512	0.9	2.50	110	13600	20
14	Varian	Thorax	512	0.9	2.50	110	13660	20
15	Varian	Thorax	512	0.9	2.50	110	12980	20
16	Varian	Thorax	512	0.9	2.50	110	13180	20
17	Varian	Thorax	512	0.9	1.98	125	13395	20
18	Varian	Thorax	512	0.9	1.98	125	13275	20
19	Varian	Thorax	512	0.9	1.98	125	13275	20

## Effect of Scatter

CBCT image quality is largely limited by the amount of scatter created by the volume being imaged. The impact of different amounts of scatter from different sized patients on radiomics values is unknown. The texture phantom used in this study is relatively small and does not well approximate the amount of scatter created by a patient. To determine whether increased scatter would substantially change texture feature values, the texture phantom was imaged on its own, then with one layer (thickness of 2.5-8 cm on each side) of scatter material (solid water equivalent and sandwich size Ziploc bags of rice), and then with two layers (thickness of 5-11 cm on each side) of scatter material, Figure 6.2. These three setups were imaged with both the head and thoracic protocols on a Varian linac.

The absolute differences in each of the features for both protocols with either one layer or two layers of scatter material versus no surrounding scatter material and the difference between one layer of scatter and two layers of scatter were calculated. The log of the ratio of these differences in phantom values to the mean intra-patient differences was then calculated as the metric for this test, Equation 6.2. A negative value for the log of the ratio implies the phantom differences were less than the mean intra-patient difference and passed while a positive value implies the phantom differences were larger and thus that the feature failed.

$$\log_{10}\left(\frac{\text{phantom diff}}{\text{mean intrapatient diff}}\right)$$

Equation 6.2

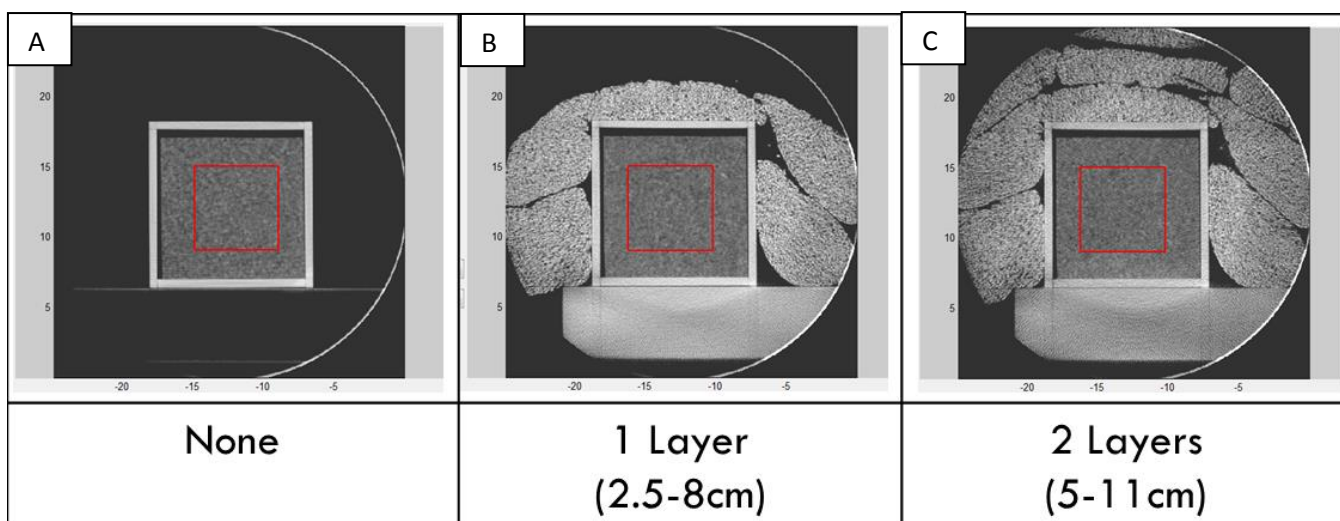


Figure 6.2: Axial images of the radiomics phantom with added scatter material. To measure the effect of scatter, the texture phantom was imaged with and without surrounding scatter material of various thicknesses.

### Effect of Motion

A third source of uncertainty in texture values obtained from CBCT images is the effect of motion. Motion has a larger effect on values measured from CBCT images than conventional CT images because the scans take longer to acquire. To analyze the effect of this motion, a CIRS dynamic motion phantom was used, Figure 6.3 (CIRS, Virginia). This motion phantom has a width of 30 cm, height of 20 cm, and length of 15 cm. These thicknesses provide some scatter and were on par with the overall thicknesses created in the previous section studying only scatter. No extra scatter material was added around the phantom.

This anthropomorphic phantom has a rod made of lung-equivalent material that can be programmed to move cyclically with different respiratory rates through the phantom lungs. This rod has a cavity designed for the placement of different tumor-equivalent or dose measurement inserts. For our analysis, we created a block of the shredded rubber material that had the size and shape of this cavity (height of 4.5 cm and diameter of 4.1 cm). Then the phantom was programmed to move the rod using a  $1-2\cos^4(t)$  waveform with peak-to-peak amplitudes of

either 0, 2, 4, 6, 8, 10, 15, 20, or 25 mm. This waveform has been shown to be representative of respiratory motion<sup>46,92</sup>. A new CBCT scan was acquired using the same thoracic protocol on a Varian linac for each programmed motion.

A 3D ROI encapsulating the shredded rubber material was delineated manually on the images of the phantom acquired with no motion. This ROI was copied to the images with motion. Radiomics values were then calculated for each image set. Equation 6.2 Equation 6.2 was used to determine at which amplitude of motion features ceased being reproducible. Here the absolute difference between the texture value measured from images with motion and the values measured without motion were calculated and used as the phantom difference values in the numerator of Equation 6.2. The values were compared to the mean intra-patient difference values, and as before, negative values implied passing, while positive values implied failing. This test was repeated using only the center slice of the motion phantom's original 3D ROI. The values from only the center slice were expected to be more reproducible because the average density change in the center of the image is less than at the edges, especially as the tumor motion increases. This test was done to determine if texture features could be reliably measured from tumors with large motion if the edges were excluded.

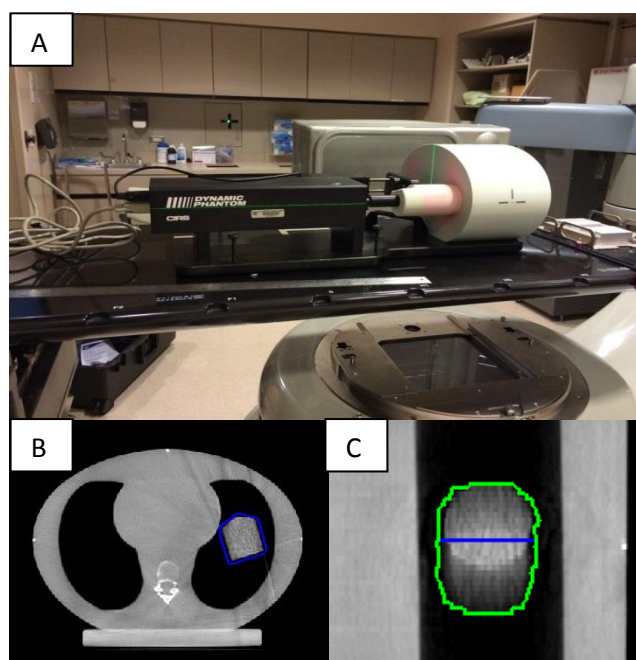


Figure 6.3: Images of the CIRS dynamic motion phantom. (A) Setup for taking CBCT images of the phantom with the shredded rubber insert in place. (B) An axial slice of the CBCT scan of the phantom with the insert visible and (C) a zoomed-in coronal slice of the insert with the largest motion of 25 mm.

## Results

### Patient Test-Retest CBCT Images

In the first part of this study we examined whether any features should be excluded because they were not reproducible even when measured from two images of the same patient acquired on the same scanner within 15 minutes. Of the original 68 features, 23 had a  $CCC < 0.9$  and were excluded from further analysis, Table 6.4. Of the remaining 45 features, 8 were excluded because the absolute value of their  $r_s$  with volume was greater than 0.85 for both the test and retest image sets, and thus might only be reproducible because they are volume-dependent, Table 6.4. Thus, a total of 37 features remained for the subsequent analyses. These included 5 features from the histogram without 8-bit scaling, 5 features from the histogram with 8-bit scaling, 16 features from the co-occurrence matrix, 8 features from the

run-length matrix, 1 feature from the neighborhood difference matrix, and 2 features filtered with the medium LoG filter.

Table 6.4: The results of the CCC and  $r_s$  tests for the patient test-retest data. \* Indicates features that have passed all three of these initial tests and were used in the rest of the analysis. VolDepA is the spearman correlation coefficient with volume using the first group of patient images, VolDepB uses the second group of patient images.

Feature	CCC	VolDepA	VolDepB	Feature	CCC	VolDepA	VolDepB
RLMsrlgle	0.986	-0.964	-0.964	HISTSCmed	0.965	0.841	0.902
* RLMsrhgle	0.968	0.358	0.358	HISTSCmean	0.980	0.842	0.915
RLMsre	0.984	-0.818	-0.891	* HISTSCmax	0.909	0.830	0.834
RLMrunperc	0.991	-0.867	-0.891	HISTSCkurt	0.728	-0.115	0.164
RLMrlnu	0.986	-0.830	-0.891	HISTSCentropy	0.885	0.430	0.467
* RLMrlrgle	0.977	0.782	0.600	HISTSCenergy	0.992	1.000	0.988
RLMrhgle	0.990	0.939	0.939	HISTmed	0.982	0.782	0.903
RLMlre	0.988	0.891	0.842	HISTmean	0.980	0.891	0.915
RLMlgire	0.981	-0.830	-0.915	* HISTmax	0.908	0.689	0.693
RLMhglre	0.982	0.842	0.915	HISTkurt	0.697	-0.176	0.200
RLMglnu	0.900	-0.758	-0.697	* HISTentropy	0.926	0.624	0.539
NDMtexstr	0.765	-0.950	-0.950	HISTenergy	0.992	1.000	0.988
NDMcontrast	0.838	-0.867	-0.867	* COMvar	0.981	0.491	0.406
NDMcomp	0.764	0.433	0.317	* COMsumvar	0.980	0.745	0.745
* NDMcoarse	0.943	0.750	0.667	* COMsument	0.903	0.491	0.503
NDMbusy	0.609	-0.933	-0.933	COMsumavg	0.965	0.782	0.903
LOGMFunif	0.780	-0.095	-0.238	COMmaxprob	0.531	-0.091	-0.103
LOGMFstd	0.118	-0.548	-0.405	* COMinvvar	0.984	0.709	0.673
LOGMFskew	0.895	0.810	0.762	COMinvdifn	0.936	0.952	0.964
LOGMFmean	0.969	-0.857	-0.690	COMinvdifmn	0.885	0.939	0.952
* LOGMFkurt	0.932	0.833	0.786	* COMinfomc2	0.933	0.842	0.818
LOGMFentropy	0.815	0.190	0.167	* COMinfomc	0.944	-0.770	-0.624
LOGFFunif	0.736	-0.033	-0.217	* COMhomog2	0.983	0.733	0.733
LOGFFstd	0.703	-0.667	-0.583	* COMhomog	0.980	0.733	0.733
LOGFFskew	0.602	0.850	0.917	COMentropy	0.874	0.115	0.297
LOGFFmean	0.774	-0.583	-0.567	COMenergy	0.671	0.042	-0.127
LOGFFkurt	0.788	0.867	0.883	* COMdissim	0.977	-0.733	-0.624
LOGFFentropy	0.859	0.150	0.300	* COMdifent	0.976	-0.673	-0.564
HISTunif	0.784	-0.624	-0.588	COMcorrel	0.966	0.939	0.903
* HISTstd	0.963	0.370	0.467	* COMcontrast	0.978	-0.673	-0.636
* HISTskew	0.967	-0.661	-0.721	* COMclustend	0.981	0.491	0.406
HISTSCunif	0.866	-0.394	-0.503	* COMclushade	0.940	-0.370	-0.358
* HISTSCstd	0.963	0.358	0.467	* COMclusprom	0.992	0.479	0.430
* HISTSCskew	0.967	-0.661	-0.721	COMautocorrel	0.977	0.867	0.818



## Effect of Different Scanners

In the second part of this study we examined whether changing the scanner or protocol resulted in changes in the texture features that were larger than the mean intra-patient difference. Features that changed less than the mean intra-patient difference passed a comparison and the overall passing percentages were recorded for each scanner/protocol combination. The results of the inter-scanner analysis for each feature are shown in Table 6.5 for the shredded rubber cartridge and in Table 6.6 for the dense cork cartridge. Features were most likely to be reproducible when scans using the same protocol and the same manufacturer were compared. This result is highlighted in

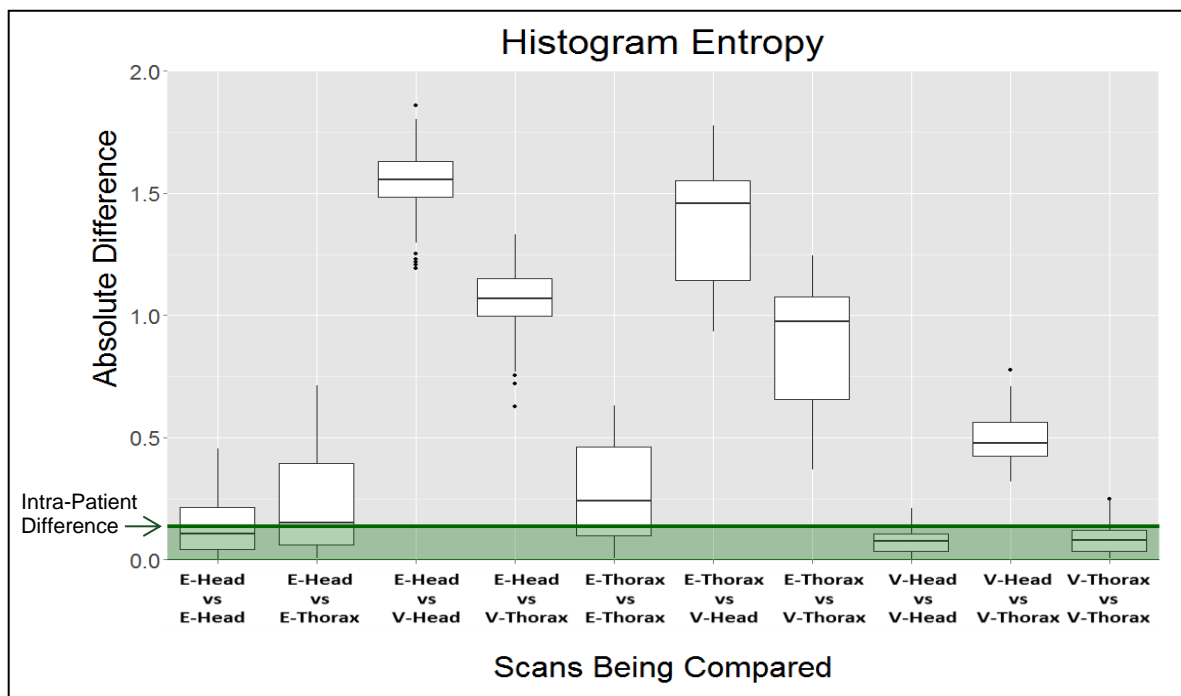


Figure 6.4, where the results for entropy measured from the histogram are shown as an illustrative example.

For shredded rubber, when the same protocol and manufacturer were used, 3 features had a passing percentage of 100%, the average across all features was 31%, and 3 features had a 0% passing percentage. When scans from the same manufacturer but different protocols were compared, the highest passing percentage was 90%, the average was only 19%, and 23

features had a 0% passing percentage. When scans from different manufacturers were compared, the highest passing percentage was only 36%, the average was less than 1%, and 36 features had a 0% passing percentage.

For dense cork, the results were slightly better for each category and 1 feature, cluster shade from the co-occurrence matrix, had a 100% passing percentage for every category. When the same protocol and manufacturer were used, 6 features had a passing percentage of 100%, the average across all features was 43%, and 2 features had a 0% passing percentage. When scans from the same manufacturer but different protocols were compared, 4 features had a passing percentage of 100%, the average was 26%, and 13 features had a 0% passing percentage. When scans from different manufacturers were compared, 2 features had a passing percentage of 100%, the average was 10%, and 33 features had a 0% passing percentage.

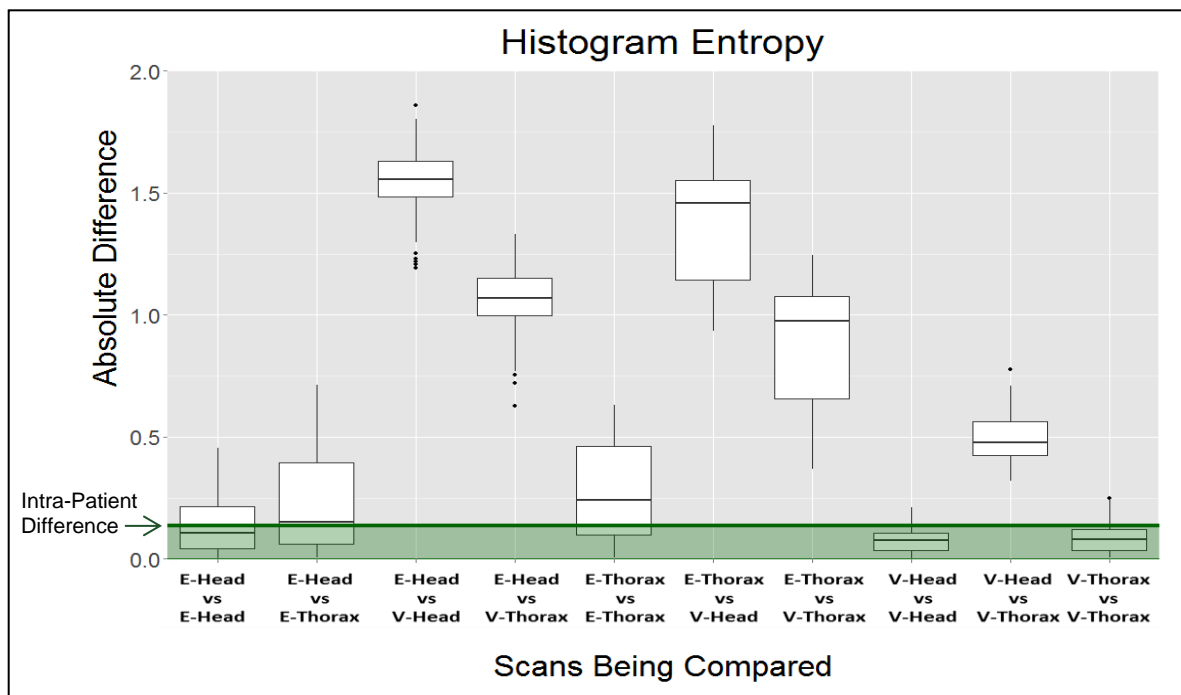


Figure 6.4: The absolute differences between pairs of scans plotted by the types of groups being compared. Scans from machines of different manufacturers had the largest absolute

differences which were above our criteria of the mean intra-patient difference (the horizontal green line).

Table 6.5: Results of the inter-scanner variability test for the shredded rubber ROI. When categories from different manufacturers (E=Elekta and V=Varian) were compared, essentially no comparisons passed the mean intra-patient difference threshold. When both the manufacturer and protocol were the same, many of the features had a passing rate above 50%.

Feature	E-Head	E-Head	E-Head	E-Head	E-Thorax	E-Thorax	E-Thorax	V-Head	V-Head	V-Thorax
	vs	vs	vs	vs	vs	vs	vs	vs	vs	vs
	E-Head	E-Thorax	V-Head	V-Thorax	E-Thorax	V-Head	V-Thorax	V-Head	V-Thorax	V-Thorax
RLMSrhgle	13.9%	0.0%	0.0%	0.0%	2.8%	0.0%	0.0%	11.1%	6.0%	8.9%
RLMSre	16.7%	14.8%	0.0%	0.0%	11.1%	0.0%	0.0%	44.4%	54.0%	73.3%
RLMrlnu	22.2%	16.1%	0.0%	0.0%	13.9%	0.0%	0.0%	35.6%	40.0%	51.1%
RLMrlgle	16.7%	0.0%	0.0%	0.0%	11.1%	23.3%	16.7%	37.8%	25.0%	20.0%
RLMlre	13.9%	12.4%	0.0%	0.0%	11.1%	0.0%	0.0%	48.9%	60.0%	84.4%
RLMlgle	5.6%	0.0%	0.0%	0.0%	16.7%	0.0%	0.0%	8.9%	13.0%	6.7%
RLMhgle	22.2%	0.0%	0.0%	0.0%	11.1%	0.0%	0.0%	11.1%	10.0%	13.3%
RLMglnu	61.1%	49.4%	0.0%	0.0%	44.4%	0.0%	0.0%	100.0%	0.0%	91.1%
NDMcoarse	36.1%	27.2%	0.0%	0.0%	22.2%	0.0%	0.0%	86.7%	56.0%	100.0%
LOGMFmean	47.2%	39.5%	0.0%	0.0%	25.0%	0.0%	0.0%	26.7%	0.0%	37.8%
LOGMFkurt	36.1%	11.1%	15.6%	12.2%	16.7%	1.1%	16.7%	91.1%	0.0%	75.6%
HISTstd	44.4%	33.3%	0.0%	0.0%	22.2%	0.0%	0.0%	20.0%	0.0%	28.9%
HISTskew	83.3%	49.4%	0.0%	21.1%	50.0%	0.0%	6.7%	80.0%	39.0%	51.1%
HISTSCstd	44.4%	33.3%	0.0%	0.0%	22.2%	0.0%	0.0%	17.8%	0.0%	28.9%
HISTSCskew	83.3%	53.1%	0.0%	21.1%	50.0%	0.0%	6.7%	80.0%	43.0%	53.3%
HISTSCmed	27.8%	0.0%	0.0%	0.0%	11.1%	0.0%	0.0%	17.8%	10.0%	20.0%
HISTSCmean	16.7%	0.0%	0.0%	0.0%	13.9%	0.0%	0.0%	8.9%	8.0%	11.1%
HISTSCmax	0.0%	0.0%	0.0%	0.0%	5.6%	1.1%	3.3%	11.1%	0.0%	4.4%
HISTmed	16.7%	0.0%	0.0%	0.0%	11.1%	0.0%	0.0%	8.9%	10.0%	8.9%
HISTmax	2.8%	0.0%	0.0%	0.0%	5.6%	3.3%	3.3%	13.3%	0.0%	11.1%
HISTentropy	55.6%	44.4%	0.0%	0.0%	38.9%	0.0%	0.0%	84.4%	0.0%	80.0%
COMvar	44.4%	37.0%	0.0%	0.0%	19.4%	0.0%	0.0%	15.6%	0.0%	13.3%
COMsumvar	8.3%	4.9%	0.0%	0.0%	2.8%	0.0%	0.0%	4.4%	0.0%	0.0%
COMsument	47.2%	38.3%	0.0%	0.0%	27.8%	0.0%	0.0%	71.1%	0.0%	64.4%
COMsumavg	19.4%	12.4%	0.0%	0.0%	2.8%	0.0%	0.0%	13.3%	0.0%	6.7%
COMinvvar	16.7%	12.4%	0.0%	0.0%	11.1%	0.0%	0.0%	26.7%	24.0%	44.4%
COMinfomc2	77.8%	69.1%	35.6%	0.0%	61.1%	31.1%	0.0%	35.6%	6.0%	57.8%
COMinfomc	58.3%	55.6%	0.0%	0.0%	38.9%	0.0%	0.0%	35.6%	9.0%	84.4%
COMhomog2	13.9%	13.6%	0.0%	0.0%	11.1%	0.0%	0.0%	31.1%	26.0%	51.1%
COMhomog	16.7%	16.1%	0.0%	0.0%	11.1%	0.0%	0.0%	28.9%	25.0%	42.2%
COMdissim	30.6%	23.5%	0.0%	0.0%	13.9%	0.0%	0.0%	4.4%	9.0%	24.4%
COMdifent	16.7%	18.5%	0.0%	0.0%	11.1%	0.0%	0.0%	20.0%	17.0%	42.2%
COMcontrast	47.2%	32.1%	0.0%	0.0%	16.7%	0.0%	0.0%	4.4%	5.0%	11.1%
COMclustend	44.4%	37.0%	0.0%	0.0%	19.4%	0.0%	0.0%	15.6%	0.0%	13.3%
COMclusshade	100.0%	90.1%	0.0%	10.0%	72.2%	0.0%	4.4%	4.4%	0.0%	31.1%
COMclusprom	44.4%	38.3%	0.0%	0.0%	25.0%	0.0%	0.0%	2.2%	0.0%	4.4%
COMautocorrel	8.3%	4.9%	0.0%	0.0%	2.8%	0.0%	0.0%	2.2%	0.0%	0.0%
Features with a Passing Rate >50%	7	4	0	0	4	0	0	7	3	13

Table 6.6: Results of the inter-scanner variability test for the dense cork ROI. Each value represents the percentage of comparisons for that category of comparisons that were less than the mean intra-patient difference value.

Feature	E-Head	E-Head	E-Head	E-Head	E-Thorax	E-Thorax	E-Thorax	V-Head	V-Head	V-Thorax
	vs	vs	vs	vs	vs	vs	vs	vs	vs	vs
	E-Head	E-Thorax	V-Head	V-Thorax	E-Thorax	V-Head	V-Thorax	V-Head	V-Thorax	V-Thorax
RLMSrhgle	25.0%	0.0%	0.0%	4.4%	21.4%	0.0%	0.0%	17.8%	20.0%	22.2%
RLMSre	16.7%	5.6%	0.0%	0.0%	17.9%	0.0%	0.0%	15.6%	25.0%	48.9%
RLMrlnu	30.6%	11.1%	0.0%	0.0%	17.9%	0.0%	0.0%	15.6%	25.0%	44.4%
RLMlrlgle	0.0%	0.0%	0.0%	0.0%	7.1%	0.0%	0.0%	0.0%	1.0%	0.0%
RLMlre	2.8%	1.4%	0.0%	0.0%	7.1%	0.0%	0.0%	11.1%	20.0%	35.6%
RLMlgle	0.0%	0.0%	1.1%	0.0%	7.1%	0.0%	0.0%	2.2%	0.0%	0.0%
RLMhgle	25.0%	0.0%	10.0%	15.6%	28.6%	0.0%	0.0%	17.8%	20.0%	26.7%
RLMglnu	94.4%	69.4%	0.0%	0.0%	67.9%	0.0%	0.0%	77.8%	2.0%	95.6%
NDMFcoarse	25.0%	9.7%	0.0%	0.0%	17.9%	0.0%	0.0%	22.2%	40.0%	77.8%
LOGMFmean	88.9%	54.2%	0.0%	0.0%	75.0%	0.0%	0.0%	40.0%	27.0%	80.0%
LOGMFkurt	44.4%	43.1%	67.8%	51.1%	46.4%	57.5%	70.0%	95.6%	66.0%	88.9%
HISTstd	86.1%	50.0%	0.0%	0.0%	53.6%	0.0%	0.0%	33.3%	2.0%	71.1%
HISTskew	100.0%	100.0%	78.9%	97.8%	100.0%	72.5%	97.5%	100.0%	86.0%	93.3%
HISTSCstd	83.3%	48.6%	0.0%	0.0%	60.7%	0.0%	0.0%	33.3%	3.0%	71.1%
HISTSCskew	100.0%	100.0%	90.0%	100.0%	100.0%	66.3%	90.0%	100.0%	88.0%	95.6%
HISTSCmed	13.9%	0.0%	6.7%	5.6%	25.0%	0.0%	0.0%	11.1%	8.0%	15.6%
HISTSCmean	16.7%	0.0%	3.3%	5.6%	21.4%	0.0%	0.0%	6.7%	6.0%	13.3%
HISTSCmax	11.1%	0.0%	0.0%	0.0%	25.0%	0.0%	0.0%	17.8%	6.0%	11.1%
HISTmed	13.9%	0.0%	3.3%	5.6%	17.9%	0.0%	0.0%	6.7%	6.0%	11.1%
HISTmax	30.6%	0.0%	0.0%	0.0%	42.9%	0.0%	0.0%	42.2%	9.0%	26.7%
HISTentropy	83.3%	50.0%	0.0%	0.0%	60.7%	0.0%	0.0%	53.3%	11.0%	95.6%
COMvar	97.2%	66.7%	0.0%	0.0%	67.9%	0.0%	0.0%	22.2%	0.0%	60.0%
COMsumvar	41.7%	11.1%	0.0%	0.0%	21.4%	0.0%	0.0%	4.4%	3.0%	4.4%
COMsument	66.7%	38.9%	0.0%	0.0%	46.4%	0.0%	0.0%	46.7%	0.0%	75.6%
COMsumavg	47.2%	16.7%	0.0%	0.0%	21.4%	0.0%	0.0%	15.6%	4.0%	13.3%
COMinvvar	72.2%	19.4%	0.0%	0.0%	21.4%	0.0%	0.0%	11.1%	19.0%	24.4%
COMinfomc2	88.9%	75.0%	51.1%	0.0%	100.0%	13.8%	0.0%	31.1%	1.0%	64.4%
COMinfomc	58.3%	12.5%	14.4%	0.0%	60.7%	21.3%	0.0%	33.3%	9.0%	71.1%
COMhomog2	19.4%	8.3%	0.0%	0.0%	17.9%	0.0%	0.0%	11.1%	18.0%	31.1%
COMhomog	22.2%	9.7%	0.0%	0.0%	17.9%	0.0%	0.0%	13.3%	19.0%	33.3%
COMdissim	63.9%	26.4%	0.0%	0.0%	35.7%	0.0%	0.0%	15.6%	23.0%	42.2%
COMdifent	33.3%	18.1%	0.0%	0.0%	17.9%	0.0%	0.0%	13.3%	22.0%	40.0%
COMcontrast	100.0%	68.1%	0.0%	0.0%	92.9%	0.0%	0.0%	15.6%	27.0%	53.3%
COMclustend	97.2%	66.7%	0.0%	0.0%	67.9%	0.0%	0.0%	22.2%	0.0%	60.0%
COMclusshade	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
COMclusprom	100.0%	100.0%	0.0%	0.0%	100.0%	0.0%	0.0%	17.8%	0.0%	62.2%
COMautocorrel	41.7%	11.1%	0.0%	0.0%	21.4%	0.0%	0.0%	4.4%	1.0%	4.4%
Features with a passing rate >50%	17	12	5	4	14	4	4	6	4	17

### Effect of Scatter

In the third part of this study we examined whether adding scatter material created changes in the texture features that were larger than the mean intra-patient difference. Features that changed less than the mean intra-patient difference passed the comparison. Results from the comparisons of features calculated with and without scatter material are in Figure 6.5 and Figure 6.6. For the dense cork cartridge imaged with the thoracic protocol, 25 of the 37 features were reproducible with 1 layer of scatter material. When a second layer of scatter material was added, 16 features were still reproducible. However, for the shredded rubber cartridge imaged with the thoracic protocol only 4 features were reproducible (regardless of the amount of scatter material added).

For the head protocol, only 10 features from dense cork and 11 features from shredded rubber were reproducible with 1 layer of scatter material. These features were not consistent, and only 4 appeared in both groups. With 2 layers of scatter material, the number of reproducible features from dense cork dropped to 4, while the number of reproducible features from shredded rubber remained at 11, 9 of which were the same as before. The most reproducible feature was skewness from the histogram.

The differences between 1 and 2 layers of scatter were smaller than the patient test-retest differences for 23 of the features measured from the thoracic scan of dense cork and 11 for the head scan of dense cork. For the thoracic scan of shredded rubber only 5 features passed and for the head scan of shredded rubber only 8 features passed.

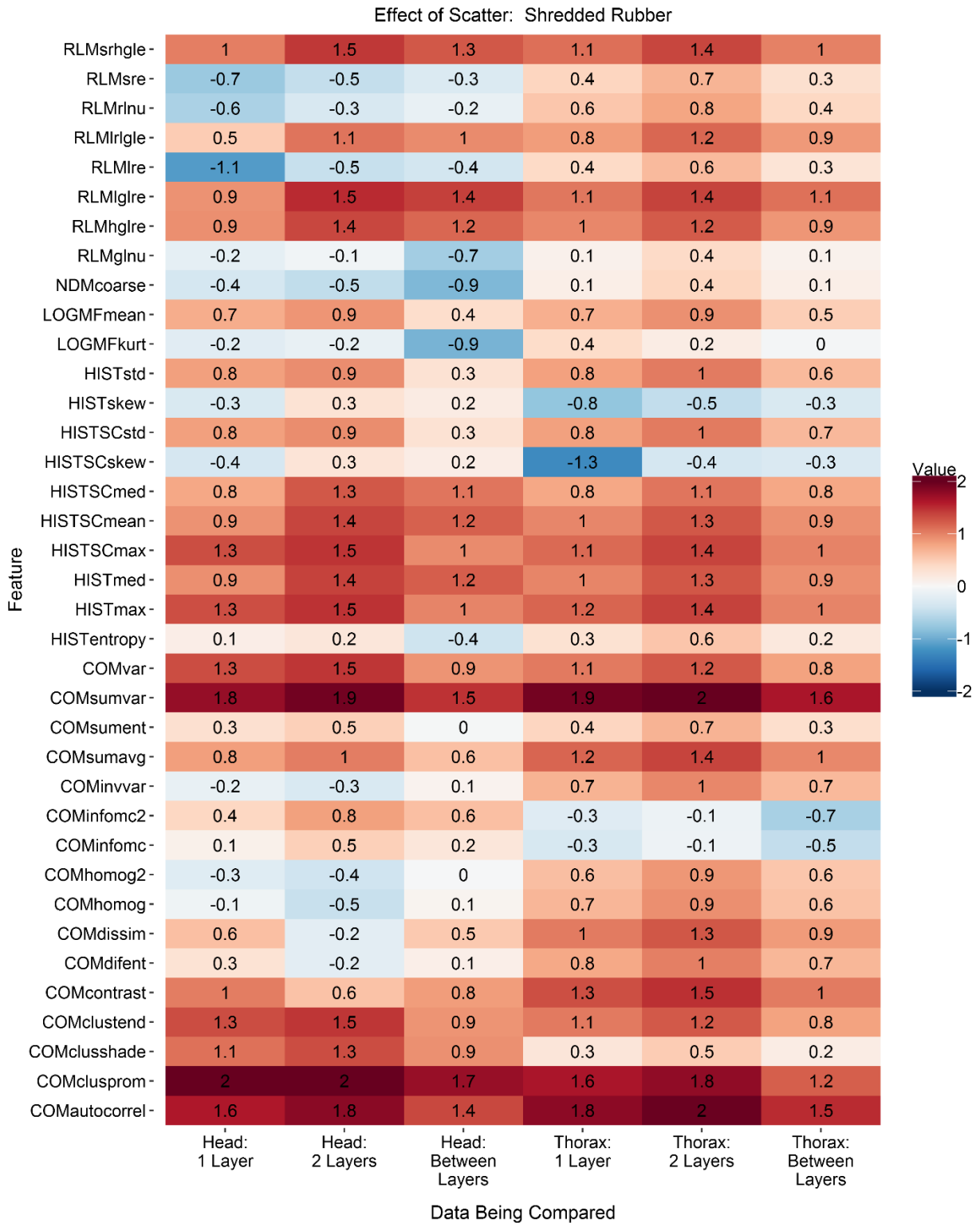


Figure 6.5: The impact of scatter on the reproducibility of radiomics features extracted from the shredded rubber cartridge. For each comparison and feature, the base 10 log ratio of the difference in phantom measurements to the mean intra-patient difference is plotted.

Comparisons are described by the imaging protocol (head or thorax) and the amount of surrounding scatter material (e.g. 1 layer versus no scatter material). Negative values are highlighted in blue and imply smaller differences in the phantom measurements than in the patient test-retest values and thus a “pass.” Positive values are highlighted in red and imply that the phantom difference was larger than the mean intra-patient difference and thus a “fail.”

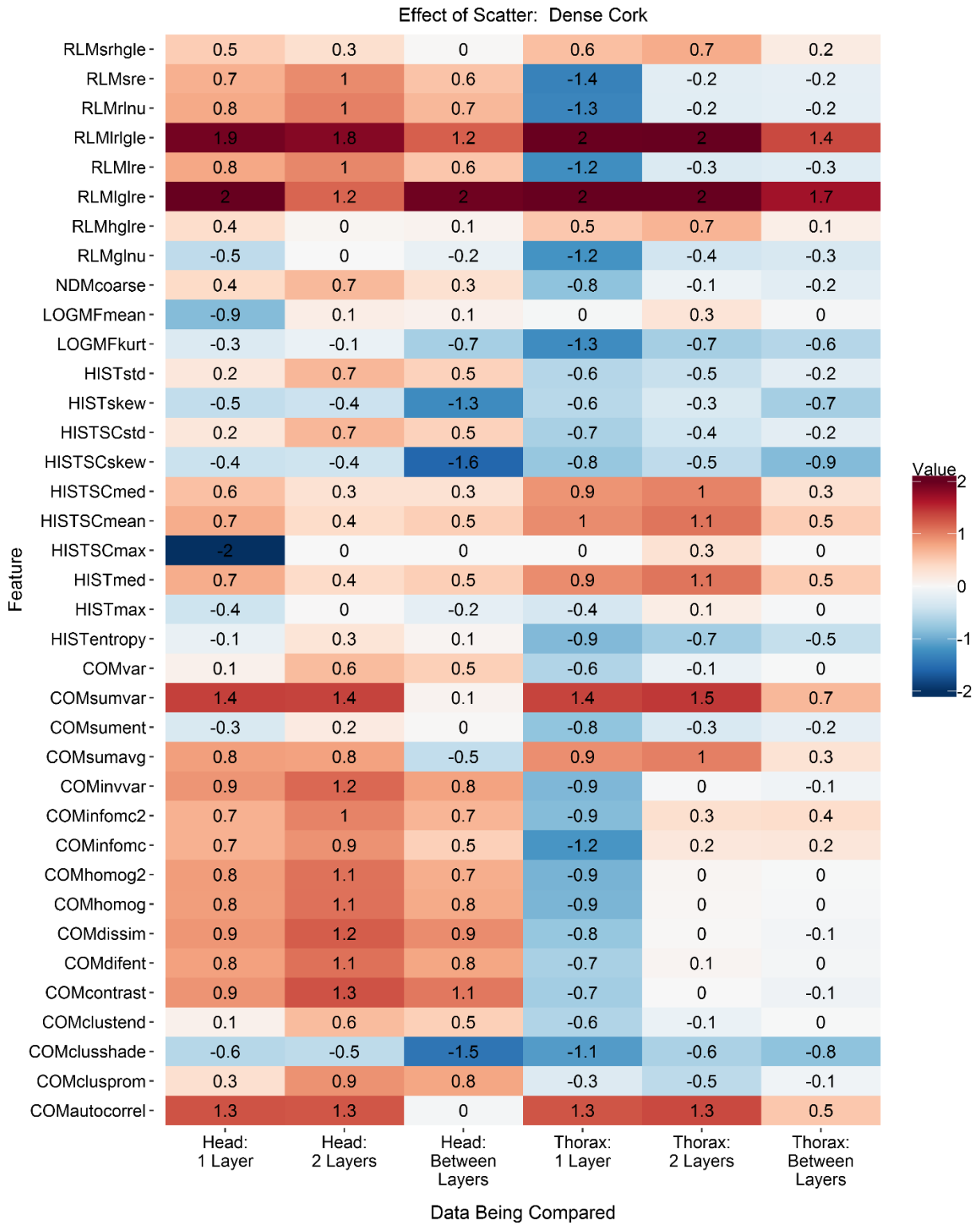


Figure 6.6: The impact of scatter on the reproducibility of radiomics features extracted from the dense cork cartridge. For each comparison and feature, the base 10 log ratio of the difference in phantom measurements to the mean intra-patient difference is plotted. Comparisons are

described by the imaging protocol (head or thorax) and the amount of surrounding scatter material (e.g. 1 layer versus no scatter material). Negative values are highlighted in blue and imply smaller differences in the phantom measurements than in the patient test-retest values and thus a “pass.” Positive values are highlighted in red and imply that the phantom difference was larger than the mean intra-patient difference and thus a “fail.”

### Effect of Motion

In the fourth part of this study we examined whether adding motion produced changes in the texture features that were larger than the mean intra-patient difference. Features that changed less than the mean intra-patient difference passed this comparison. The number of features that were reproducible decreased with increasing motion amplitude, Figure 6.7 and Figure 6.8. Three features: LoG filtered kurtosis, gray-level nonuniformity from the run-length matrix, and entropy from the histogram, were reproducible for motions of 6-8mm when measured from the entire volume. At 4 mm of motion, 12 of the 37 features were reproducible for the entire volume. When only the center image slice was used for feature calculation, seven features were reproducible for up to 6-10 mm of motion. The most consistent features measured from only the center slice were coarseness from the neighborhood difference matrix, high gray-level run emphasis and gray-level nonuniformity from the run-length matrix, sum-average and information measure correlation from the co-occurrence matrix, and scaled mean and entropy from the histogram. At 4 mm of motion, 14 of the 37 features were reproducible for the center slice measurements.

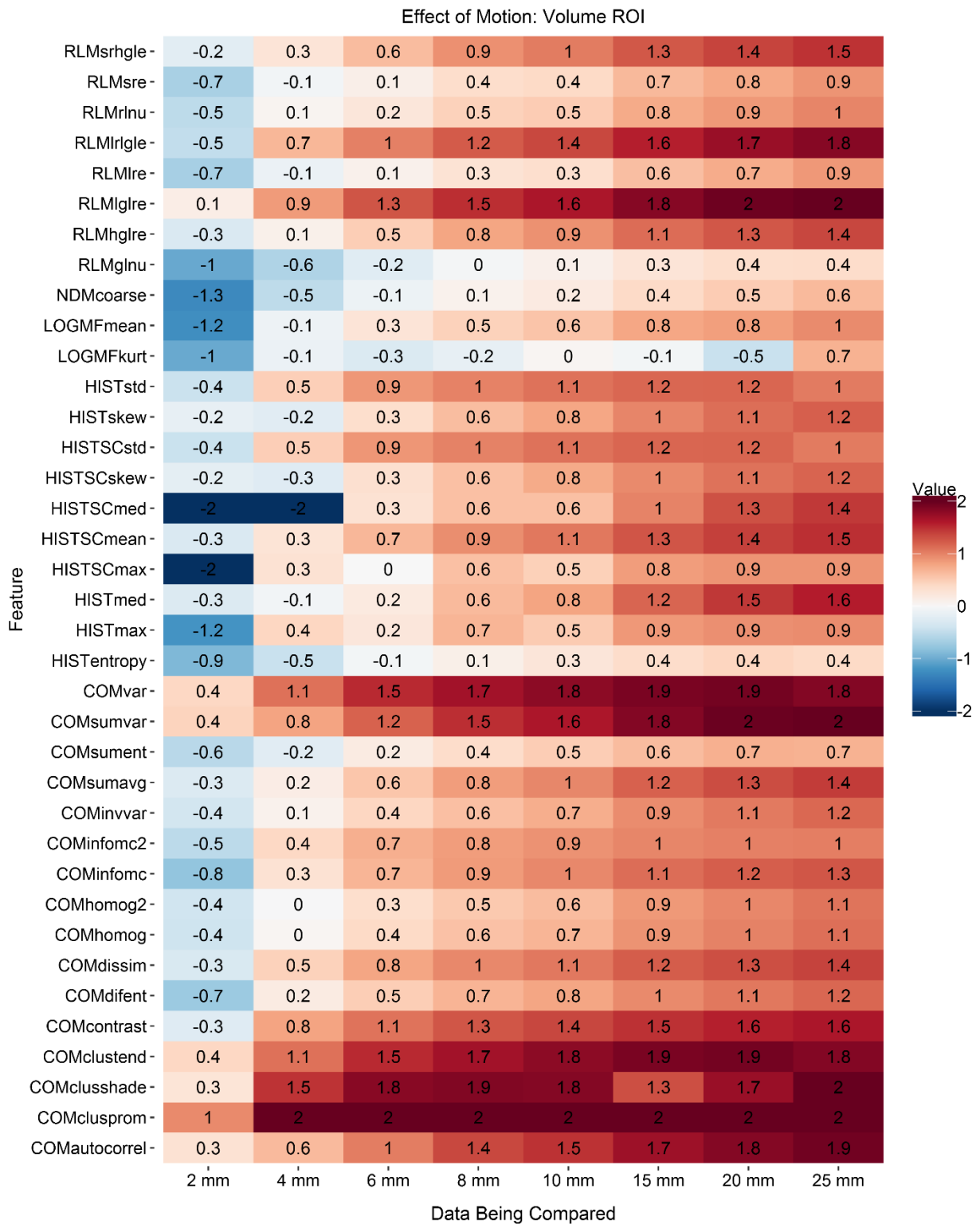


Figure 6.7: The impact of motion on the reproducibility of radiomics features extracted from the dense cork insert in the motion phantom using the full 3D ROI. Comparisons are between different peak-to-peak amplitudes of motion (2mm to 25mm) and no motion. For each

comparison and feature, the base 10 log ratio of the difference in phantom measurements to the mean intra-patient difference is plotted. Negative values are highlighted in blue and imply smaller differences in the phantom measurements than in the patient test-retest values and thus a “pass.” Positive values are highlighted in red and imply that the phantom difference was larger than the mean intra-patient difference and thus a “fail.”

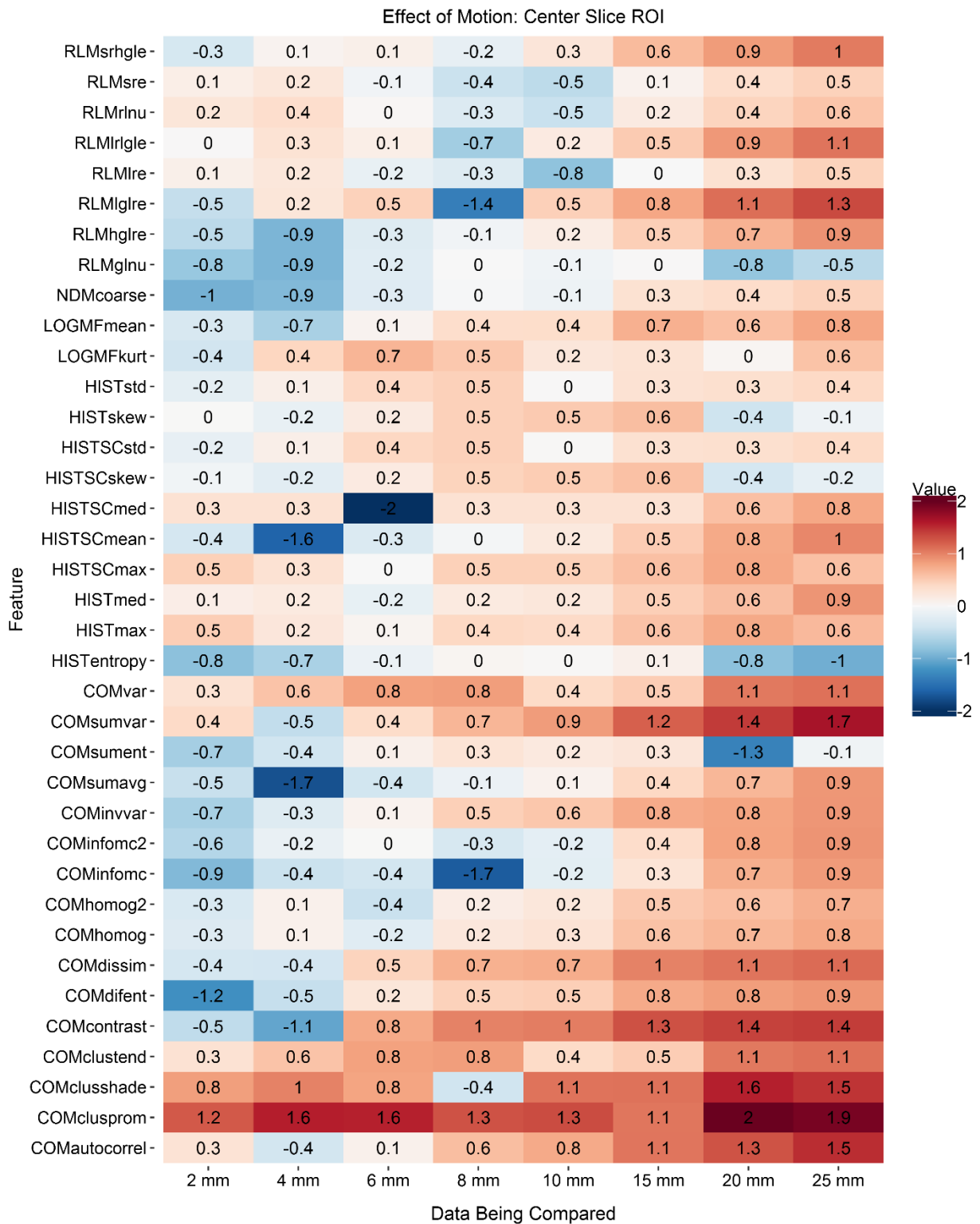


Figure 6.8: The impact of motion on the reproducibility of radiomics features extracted from the dense cork insert in the motion phantom using a 2D ROI from an axial image slice.

Comparisons are between different peak-to-peak amplitudes of motion (2mm to 25mm) and no

motion. For each comparison and feature, the base 10 log ratio of the difference in phantom measurements to the mean intra-patient difference is plotted. Negative values are highlighted in blue and imply smaller differences in the phantom measurements than in the patient test-retest values and thus a “pass.” Positive values are highlighted in red and imply that the phantom difference was larger than the mean intra-patient difference and thus a “fail.”

## Discussion

### Patient Test-Retest

The goal of this study was to determine whether any radiomics features can be reproducibly measured from CBCT images so that features could be tracked periodically through treatment. In order to investigate this question, we initially considered a large number of features. Two tests were used to eliminate features that were not reproducible in a patient test-retest dataset or that were only reproducible due to their volume dependence. Features that were not reproducible even for the same patient on the same machine are extremely unlikely to be useful in future models and could lead to erroneous results. Features that are volume dependent will appear to be reproducible especially in patient datasets where the volume range is large. However because volume is already known to be prognostic and is easy to extract from patient images without the rigor of a texture analysis, volume dependent features would not add meaningful information to future models and could have led to misleading results in this investigative study. Approximately half the features initially considered passed both of these qualifying tests. Interestingly, at least one feature from every feature category was successful in passing these tests. The large number and wide variety of features that passed, offer preliminary support for the possibility of texture analysis in CBCT images. This relatively high pass rate occurred despite the strict criteria for reproducibility ( $CCC \geq 0.9$ ). We deliberately adopted very strict and conservative cutoffs here in order to minimize the possibility of false-positives in this analysis. Many of the excluded features had CCC values in the 0.75-0.89 range representing medium reproducibility and thus we may have excluded features that could potentially be useful in the future.

In order to determine the effects of scanner, scatter, and motion on this reproducibility, phantom measurements were compared to the mean intra-patient difference for each feature. This choice of threshold is one limitation of our study since it is partly arbitrary, and may not accurately reflect the amount of variability in a texture that would significantly influence an

eventual prognostic model. However because the purpose of this study was only to introduce the magnitudes of variability that are created by changes in scanner, motion, and scatter we feel that our choice of threshold is justified. Furthermore by using the mean of the differences measured from patient test-retest data rather than the maximum, it is more likely that our choice of threshold is overly conservative than lenient.

Another limitation of this part of the analysis was the small number of patients with available repeat images. A larger patient dataset may have increased the variability we saw in patient test-retest values which in turn may have increased the number of features that passed each test.

### Inter-Scanner Analysis

The results of our inter-scanner analysis strongly indicated that radiomics values obtained from different imaging protocols or different linac manufacturers should not be compared. This is a useful result for anyone considering extracting radiomics features from CBCT images in order to produce a model. The Elekta values may have differed from the Varian values because of manufacturer differences in Hounsfield unit scaling. CBCT CT numbers can be more variable than the CT numbers in regular CT images because CBCT images include more scatter due to the 2D detector geometry and are acquired with less dose since they are used primarily for image setup versus the primary goals of CT images which include diagnosis, contouring, and dose calculation. Additionally differences between manufacturers in HU mapping from CBCT images could play a role in the observed differences seen between scans of the same phantom.

The inter-scanner analysis also revealed that more features were likely to pass when measured from the dense cork cartridge than the shredded rubber cartridge. This is likely because the dense cork cartridge is physically more uniform than the shredded rubber cartridge. For the same scan, the dense cork standard deviation was typically one-half to one-third the value of the shredded rubber standard deviation. Thus even when magnitude shifts or

varying levels of noise are introduced by using a different scanner or protocol, the dense cork cartridge individual voxels are less varied than those of the shredded rubber cartridge. The patients' standard deviations fell within the range of both cartridges so the values from the shredded rubber cartridge can be assumed to approximate the variability in a patient with heterogeneous texture while the dense cork cartridge may approximate a patient with homogenous texture.

Some features were not reproducible even when both the manufacturer and protocol were kept consistent. These features may be overpowered by the noise in the image making them essentially random. This is probably the reason why features such as the maximum value from the histogram failed each comparison. In other cases, such as low gray-level run emphasis from the run-length matrix, the texture values from the patient and phantom images are essentially always zero because the feature searches for specific patterns that do not exist in images of tumors (such as straight lines).

This analysis also demonstrated that reproducible features could come from any of the feature categories, e.g. skewness from the histogram, cluster shade from the co-occurrence matrix, normalized gray level non uniformity from the run length matrix, and the mean after LoG filtration. This broad spectrum of reproducible features is helpful, because features from different categories may provide independent information about an image and when combined, may be able to provide a more complete picture than one feature alone.

### Effect of Scatter

When the texture phantom was surrounded by scatter material, most of the texture values changed more than the mean of the patient test-retest differences. This result was not surprising, as we know that more surrounding material will result in more scatter and thus a larger amount of noise in the image as well as artifacts from beam hardening and cupping. The differences between 1 layer and 2 layers of scatter material were also in general larger than the mean intra-patient difference. This is a problem because it suggests that two patients with

physiologically alike tumors (i.e. similar levels of heterogeneity) could have very different values for their computed textures if the patients are dissimilar in size. The impact of this problem may be limited if texture features measured from CBCT images are used to observe how the features change for a single patient over time. For that analysis, the relative difference in a texture value could be measured for each patient. The change in the amount of scatter would likely be substantially less than shown here if each patient acted as his or her own control. A recent study investigating c-arm CBCT demonstrated that relative changes in mean Hounsfield units were consistent when measured within patients<sup>93</sup>. Thus, it is possible that relative changes in texture may still be used for future prognostic models despite the effect of changes in scatter levels on the absolute value measured from any one patient.

### Effect of Motion

Most of the features changed substantially with increasing motion of the tumor texture insert. The main reason for this result was hypothesized to be the slices of the ROI at the edge of the texture insert, where the density changes were greatest. This hypothesis was supported by our data for many of the features which did not significantly change with small motion if only the center slice was used for their calculation.

While a majority of features were no longer reproducible beyond 2 mm of motion when the entire tumor volume was used, 12 of the 37 features did still pass at 4 mm of motion and 4 of these even passed at 6mm of motion. The number of reproducible features dropped to zero at 10 mm when the entire tumor volume was used and to 1 when only the center slice was used. Therefore, we recommend a threshold of at most 10 mm and potentially as low as 5mm for future studies. This threshold is not unduly restrictive since a recent study showed a majority of patients with NSCLC had tumor motion less than 5mm and only 10% had motion greater than 10 mm<sup>49</sup>. Thus we think a future study limited to patients with little motion and selecting only these most reproducible features for further investigation would be feasible.

Additionally, either 4D CBCT or breath-hold CBCT could be used to mitigate tumor motion in future studies and may be more successful than shrinking the tumor contour.

Several features, when measured from the center slice, were reproducible at large motions while not being reproducible at smaller motions. For example sum entropy from the co-occurrence matrix was reproducible with 2-4mm and 20-25 mm but not with 6-15 mm of motion. This inconsistency suggests that at large motions the feature may be returning reproducible values by coincidence or because of artifacts. Thus we would not consider this feature reproducible beyond 4 mm.

It should also be noted that while the motion phantom was larger than the texture phantom (32 cm vs 10 cm diameter), it is still smaller than many patients. Thus in a clinically realistic scenario the effects of motion and additional scatter would be combined and may further reduce the number of features or the range of motion that could be considered reproducible.

### Overall Best Performing Features

From our results it appears that select features are reproducible under certain circumstances. Several of these reproducible features have been found useful in studies using CT or contrast-enhanced CT images. One study found skewness, which we showed to be robust to scatter, may aid in identifying tumors with genetic mutations<sup>32</sup> while another study demonstrated it was prognostic for overall survival<sup>25</sup>. LoG filtered kurtosis was useful for identifying tumors with genetic mutations<sup>32</sup> and we showed it was robust to scatter and motion. Gray-level nonuniformity from the run-length matrix was the feature we tested that was most robust to the effects of motion and it has been shown to be useful for predicting survival in NSCLC<sup>23</sup> and differentiating between benign and malignant lymph nodes<sup>21</sup>. Cluster shade from the co-occurrence matrix was able to pass all of the scatter tests for the dense cork material, and recently was shown to be useful for prognosis when used in a radiomics signature of three features for NSCLC patients<sup>28</sup>. These links are encouraging but an independent study will still

be needed to determine if models built on CBCT features alone can be prognostic. However it should be clear that the features which did change more dramatically when measured from a phantom than the calculated mean intra-patient differences in our study are unlikely to be useful in future analyses using CBCT images of NSCLC unless the patient cohort was highly restricted to patients exhibiting low intrascan variability (e.g. negligible tumor motion and minimal weight change).

## **Conclusions**

The goal of this study was to determine if texture features could be reliably extracted from CBCT images under a variety of conditions. A total of 68 features were originally considered. However, 31 of these features were excluded from the analysis because they did not have a high CCC value when measured from a test-retest dataset or had a strong volume-dependence that might be responsible for their high CCC. The remaining 37 features included at least one feature from each feature category that had been studied. These features were then investigated for susceptibilities to differences in scanners, imaging protocols, scatter, and motion. Features changed significantly if they were calculated from images acquired with different protocols or with scanners from different manufacturers. Future studies should attempt to keep their imaging protocols as uniform as possible to avoid this source of error. Almost every feature changed more than the mean intra-patient difference with the addition of scatter. Thus, values of features may not be comparable between patients of different sizes while remaining insensitive to small changes in size of each individual patient. Lastly, no features can be reliably measured if the tumor motion is greater than 1 cm. For motion less than 1 cm, reproducibility is improved if the edges of the tumor are excluded from the ROI for texture calculation. In summary, certain texture features can be reliably measured from CBCT images as long as the imaging protocol is consistent, relative differences are used, and patients are limited to those with less than 1 cm of tumor motion.

# Chapter 7 : Treatment Modality Dependence of Radiomics

## Features

In this chapter we describe the results for Specific Aim 4: Compare the changes in radiomics features from patients treated with intensity modulated radiation therapy (IMRT) to those treated with passive scatter proton therapy (PSPT). Our working hypothesis for this aim was that the changes in radiomics features measured from patients treated with protons would occur earlier in treatment compared to those treated with photons due to the increased relative biological effectiveness (RBE).

### Introduction

If radiomics features change during treatment due to biological changes in the tumor, then the rate or magnitude of their changes may also be affected by the radiation modality used (protons vs. photons). The patients used through the majority of this thesis and described in Chapter 3 were treated with either passive scatter proton therapy (PSPT) or intensity modulated radiation therapy using photons. The clinical relative biological effectiveness (RBE) for protons is assumed to be 1.1 for calculating equivalent dose; however, it is known to vary with depth of penetration and possibly between different tissues<sup>94,95</sup>. The RBE is a complex function of dose per fraction, linear energy transfer (LET), tissue, and cell type<sup>94</sup>. If the radiomics features are measuring biological differences from the images than any deviation from a RBE of 1 could translate to a slower or faster rate of change in those features measured from tumors treated with protons compared to features measured from tumors treated with photons.

This study had several goals. The first was to compare the distribution of values at the beginning and end of treatment between the patients treated with the two modalities. We hypothesized that the distributions would be the same at the beginning of treatment but would have significantly diverged by the end. The second was to determine how early in treatment

these changes begin to occur for each treatment modality. We hypothesized the radiomics features for patients treated with protons would demonstrate significant changes earlier in treatment compared to patients who received IMRT due to the increased RBE of protons.

## **Methods**

### Features

For this study, the same feature set that was designed in Chapter 5 was used. This feature set included 31 features that were calculated with feature-specific image preprocessing, Figure 7.1. The feature-specific image preprocessing was determined by selecting the technique that resulted in a significant ( $p\text{-value} < 0.1$ ) univariate cox proportional hazards fit for the feature using its pre-treatment values and a non-significant ( $p\text{-value} > 0.05$ ) result for the Wilcoxon rank sum test comparing values from two different CT models (GE Discovery ST and the GE Lightspeed RT16). The feature set included features from the histogram, co-occurrence matrix, neighborhood gray-tone difference matrix, run length matrix, and shape categories. Volume was added to this feature set as a comparison metric, bringing the total number of features to 32.



Figure 7.1: List of features included in the analysis of modality dependence and the image preprocessing used to calculate each. None=no extra image preprocessing after image thresholding, Smooth=images are smoothed with a Butterworth filter with an order of 2 and a cutoff frequency of 125 prior to image thresholding, and Both=images are smoothed with a Butterworth filter with an order of 2 and a cutoff frequency of 125, then thresholded, and then resampled to 8 bit depth.

## Patients

For this analysis the patient cohort described in Chapter 3 was used again. These patients had been treated with passive scatter proton therapy (PSPT) or intensity modulated radiation therapy (IMRT). All patients were treated with 2 Gy fractions to either 66 or 74 Gy. Each patient had a 4DCT pre-treatment image acquired for treatment planning and weekly 4DCT images acquired during their treatment. Radiomics features were calculated from the tumor GTV ROI on the end-of-exhale phase of each of these images. This patient set was reduced by different criteria to form 3 different subsets which are described in the next section.

## ANOVA Classification

Of the 110 patients in the main cohort, 68 were treated with photons and 42 were treated with protons. The patients treated with photons were overwhelmingly imaged with a GE Discovery (362 of 448 weekly images) while the patients treated with protons were mainly imaged with a GE LightSpeed RT16 (222 of 282 weekly images). Thus there was a risk that by comparing the two groups, a significant difference in the radiomics feature values could be found that was attributable to a difference in the CT models rather than the tumors. While certain tests had already been performed during the feature-specific image preprocessing selection in Aim 2 to negate this possibility, even stricter criteria were implemented for this analysis to ensure that any discovered changes between treatment modalities could not be attributed to differences in the CT models. To circumvent this possibility, an ANOVA analysis was performed for each feature to evaluate whether it changed as a result of treatment and was independent of CT model. To balance these two factors, a subset of 17 patients was identified that had been imaged on each of the two CT models both early and late in treatment. Images were classified as early in treatment if they were acquired for treatment planning (week 0) or during treatment weeks 1 and 2; images were classified as late in treatment if they were acquired during treatment weeks 4 through 6. Features were considered to be significantly different based on treatment time if the p-value was  $< 0.01$ . Features were considered to be

significantly different between CT models if the p-value was  $<0.2$ . Using these two cutoff values each feature was classified as either (i) treatment dependent, (ii) model dependent, (iii) treatment and model dependent, or (iv) treatment and model independent. A tight cutoff was used to classify a feature as treatment dependent to ensure that only features with particularly large changes were investigated. This also helps reduce the penalty from multiplicity corrections in later tests. Similarly the p-value cutoff for a difference between CT models was large to ensure even features with a weak dependence on CT model were classified as model dependent and thus reduce the number of features classified as independent of CT model.

### Analysis

The treatment dependent and treatment & model dependent groups of features were investigated independently to determine how early during treatment they began to exhibit changes and whether those changes were ever different between patients treated with protons and those treated with photons.

Paired Wilcoxon signed-rank tests were used to evaluate at what time point during treatment the radiomics features first exhibited significant changes from baseline. The test was performed between the values at each week of treatment and the values at baseline. Radiomics features for patients with an eligible CT image at a particular week were compared to the radiomics feature values at pre-treatment using the Wilcoxon test. The final p values for all comparisons were corrected for multiple comparisons using the Bonferroni method because in this case the multiple hypotheses were not independent. Corrected p values  $< 0.05$  were considered significant.

Wilcoxon rank sum tests were used to determine if the radiomics feature values at any week in treatment were different between patients treated with protons versus photons. The final p values were corrected using the Bonferroni method for multiple testing, and corrected p values of  $<0.05$  were considered significant.

Wilcoxon rank sum tests were also used to determine if the net changes in radiomics features from patients treated with proton therapy were significantly different from the net changes for patients treated with IMRT. Net changes were calculated between each week of treatment and baseline. The final p values were corrected using the Bonferroni method for multiple testing, and corrected p values of  $<0.05$  were considered significant.

The subset of images available for each test was different based on the classification of the features being investigated. For the treatment dependent features, the dataset was reduced so that each patient only had values from the CT scanner on which most of their images had been acquired. For the treatment & model dependent features, the dataset was reduced so that all patients only had values from the GE Discovery ST. This was done so that the impact of CT model was removed from the analysis for these features. However using only images acquired on one scanner substantially decreases the power of the test and thus for features that were not model dependent it was preferable to use the larger set of images for each patient. The different cohorts used in this analysis were,

- (i) Primary cohort: 110 patients with between 2-8 images each acquired on any CT model. This was the full original patient cohort and was decreased to form the ANOVA, secondary, and tertiary cohorts as outlined in the guidelines below.
- (ii) ANOVA cohort: 17 patients with 4 images each. These were the patients who had received a 4DCT on each of the primary CT models both early in treatment (weeks 0-2) and late in treatment (weeks 4-6). This dataset was only used to classify features as treatment dependent, model dependent, treatment & model dependent, or treatment & model independent.
- (iii) Secondary cohort: 81 patients with between 2-8 images each, where for a given patient all of their images were acquired on either the GE Discovery ST or GE LightSpeed RT16, they had an image acquired on their main scanner by week 1,

and had at least 2 images remaining after these criteria were implemented. In this case any net change that was measured would be between the same scanner. This cohort was used for comparing features that were not significantly dependent on CT model. The number of patients with an image available at pre-treatment was heavily weighted towards photon patients (43 photon patients vs 6 proton patients) thus for this dataset, baseline was defined as Week 1 of treatment where the distribution was more even (31 photon patients vs 28 proton patients). This meant only 59 of the possible 81 patients were used in calculations.

- (iv) Tertiary cohort: 73 patients with 2 images each, where all of the images were acquired on a GE Discovery ST. Each patient had a pre-treatment (Week 0) 4DCT for treatment planning and a mid-treatment 4DCT acquired on the GE Discovery ST. The mid-treatment 4DCT was acquired on weeks 4-6. If the patient had more than one image available during this period, one was chosen based in this order of preference: week 5, week 6, week 4. This cohort was used for analyzing features that were significantly dependent on CT model (the treatment & model dependent feature set). The mid-treatment 4DCTs were grouped together into week 5 for calculations since each patient only had 1 mid-treatment 4DCT.

The feature set of primary interest were those that were treatment dependent with no significant dependence on CT model. However using the tertiary cohort we were also able to analyze features that were treatment & model dependent in a way that controlled for the model dependence. Features that did not exhibit a dependence on treatment in the ANOVA analysis were not examined.

The number of patients treated with each modality with an image at each week of treatment for each of the cohorts is presented in Table 7.1. Then in Table 7.2 the number of

images at each week used in the calculations after taking into account the baseline shift for the secondary cohort and the mid-treatment grouping of images for the tertiary cohort are shown.

Table 7.1: Number of patients with an image available at each week for each cohort.

Cohort	Modality	Week 0	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7
Primary Cohort (All images)	Photon	68	38	59	59	62	57	58	47
	Proton	42	29	37	40	36	34	35	29
	Total	110	67	96	99	98	91	93	76
Secondary Cohort (1 CT model per patient)	Photon	43	31	41	43	43	37	44	32
	Proton	6	28	27	30	23	17	23	19
	Total	49	59	68	73	66	54	67	51
Tertiary Cohort <sup>1</sup> (1 CT model in the cohort)	Photon	54	0	0	0	1	46	7	0
	Proton	19	0	0	0	4	11	4	0
	Total	73	0	0	0	5	57	11	0

Table 7.2: Number of patients at each week in each cohort used in calculations.

Cohort	Modality	Week 0	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7
Primary Cohort (All images)	Photon	68	38	59	59	62	57	58	47
	Proton	42	29	37	40	36	34	35	29
	Total	110	67	96	99	98	91	93	76
Secondary Cohort (1 CT model per patient)	Photon	--	31	25	27	28	23	27	18
	Proton	--	28	25	28	21	15	21	18
	Total	--	59	50	55	49	38	48	36
Tertiary Cohort <sup>1</sup> (1 CT model in the cohort)	Photon	54	0	0	0	0	54	0	0
	Proton	19	0	0	0	0	19	0	0
	Total	73	0	0	0	0	73	0	0

<sup>1</sup>The baseline for the secondary cohort was set to Week 1 instead of Week 0 because the number of patients with an image available at Week 0 were heavily biased to those treated with photons. <sup>2</sup>Values for the tertiary cohort from Week 4-6 were grouped together into Week 5 to increase power in calculations leading to a total of 54 photon patients and 19 proton patients for that analysis.

### Clinical Factors

Clinical factors were compared between patients receiving IMRT versus PSPT using the full patient set of 110 patients. A chi-square test was used for categorical factors and a Wilcoxon rank-sum test was used for continuous factors. The characteristics for each group are tabulated in Table 7.3 and the test results are summarized in Table 7.4 and show that smoking status was the only covariate significantly differed between the patients treated with photons versus protons.

Table 7.3: Clinical characteristics of the patients treated with protons versus photons.

Clinical Factor	Values	Proton Patients (n=42)	Photon Patients (n=68)
Age	Mean (range)	66.8 (51-76)	64.4 (47-80)
Sex	Female	15	32
	Male	27	36
KPS	70	6	2
	80	14	35
	90	22	30
	100	0	1
ECOG	0	3	3
	>0	39	65
Smoking	Never	2	7
	Former	19	48
	Current	21	13
Pack years	Mean (range)	58.6 (4-200)	54.1 (5-180)
Histology	Adenocarcinoma	23	40
	Squamous cell carcinoma	19	28
T Stage	0-1	19	30
	2-3	23	38
N stage	0-1	12	12
	2-3	30	56
Stage	II	8	4
	III	34	62
	IV	0	2
Prescribed dose	66	12	26
	74	30	42

Table 7.4: Differences in clinical factors for patients treated with photons versus protons (n=110).

Clinical Factor	Test Used	P value
Age	Mann-Whitney	0.164
Sex	Chi-square	0.332
KPS	Chi-square	0.057
ECOG performance status	Chi-square	0.857
Smoking status	Chi-square	0.003
Pack years	Mann-Whitney	0.873
Histology	Chi-square	0.826
T category	Chi-square	1.000
N category	Chi-square	0.267
Overall stage	Chi-square	0.059
Prescribed dose	Chi-square	0.407

ECOG: Eastern Cooperative Oncology Group; KPS: Karnofsky Performance Status

## Results

### ANOVA Results

Of the 32 features tested, 10 demonstrated both a strong dependence on dose in the ANOVA ( $p < 0.01$ ) and no significant dependence on the CT scanner model used to acquire the images ( $p > 0.2$ ), Figure 7.2. These were classified as treatment dependent. Of the other 21 features, 18 were dependent on treatment and CT model, 3 were only dependent on the CT model, and 1 was independent of both treatment and CT model. Scaled values for one feature from each group are displayed as boxplots in Figure 7.3. The p-values for both covariates after correction are in Table 7.5

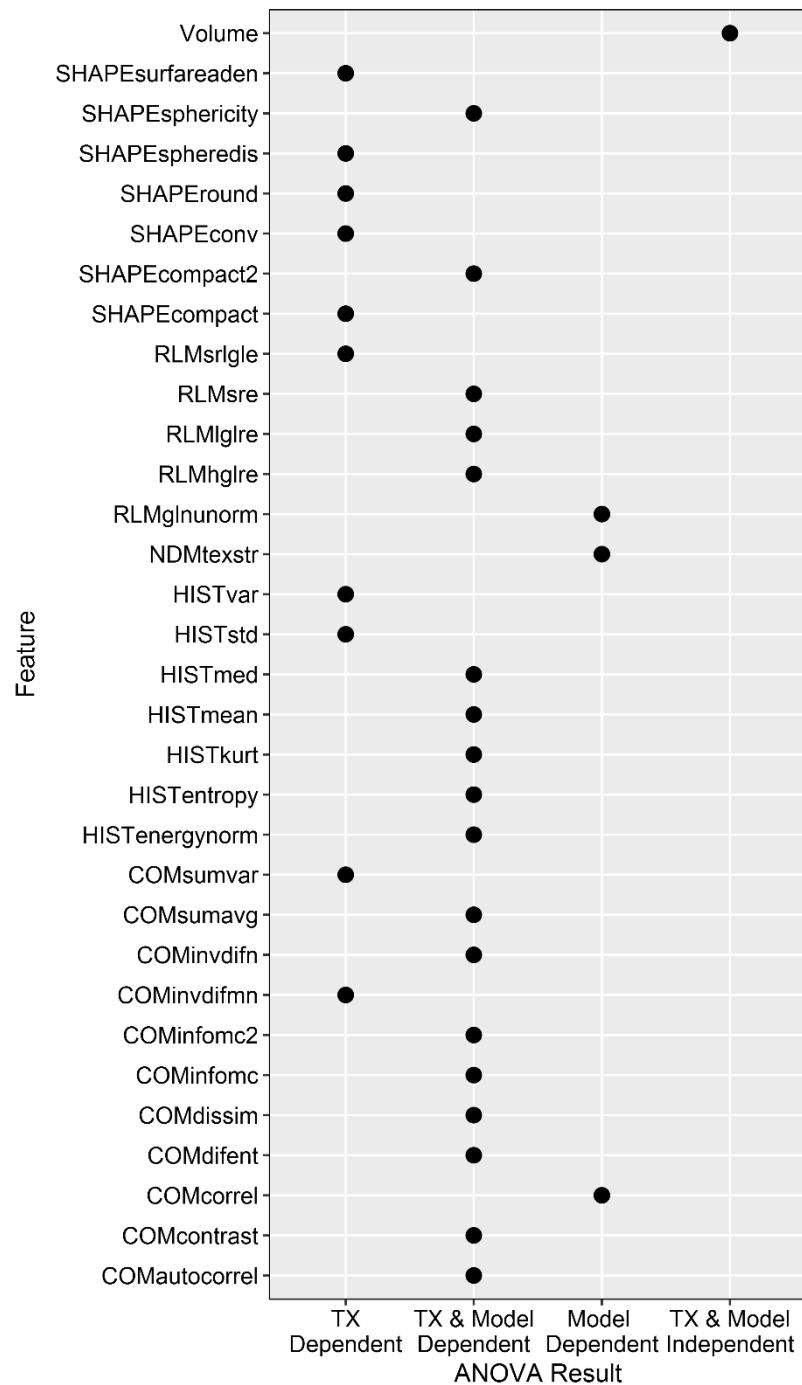


Figure 7.2: Results of the ANOVA analysis for each feature to determine dependence on treatment and CT scanner model. Features were categorized as treatment (TX) dependent, model dependent, model and treatment dependent, or model and treatment independent. The p values for both treatment and model were corrected for multiplicity using the Benjamini-

Hochberg correction. A feature was considered dependent on treatment if its p value was  $<0.01$  and dependent on model if its p value was  $<0.2$ .

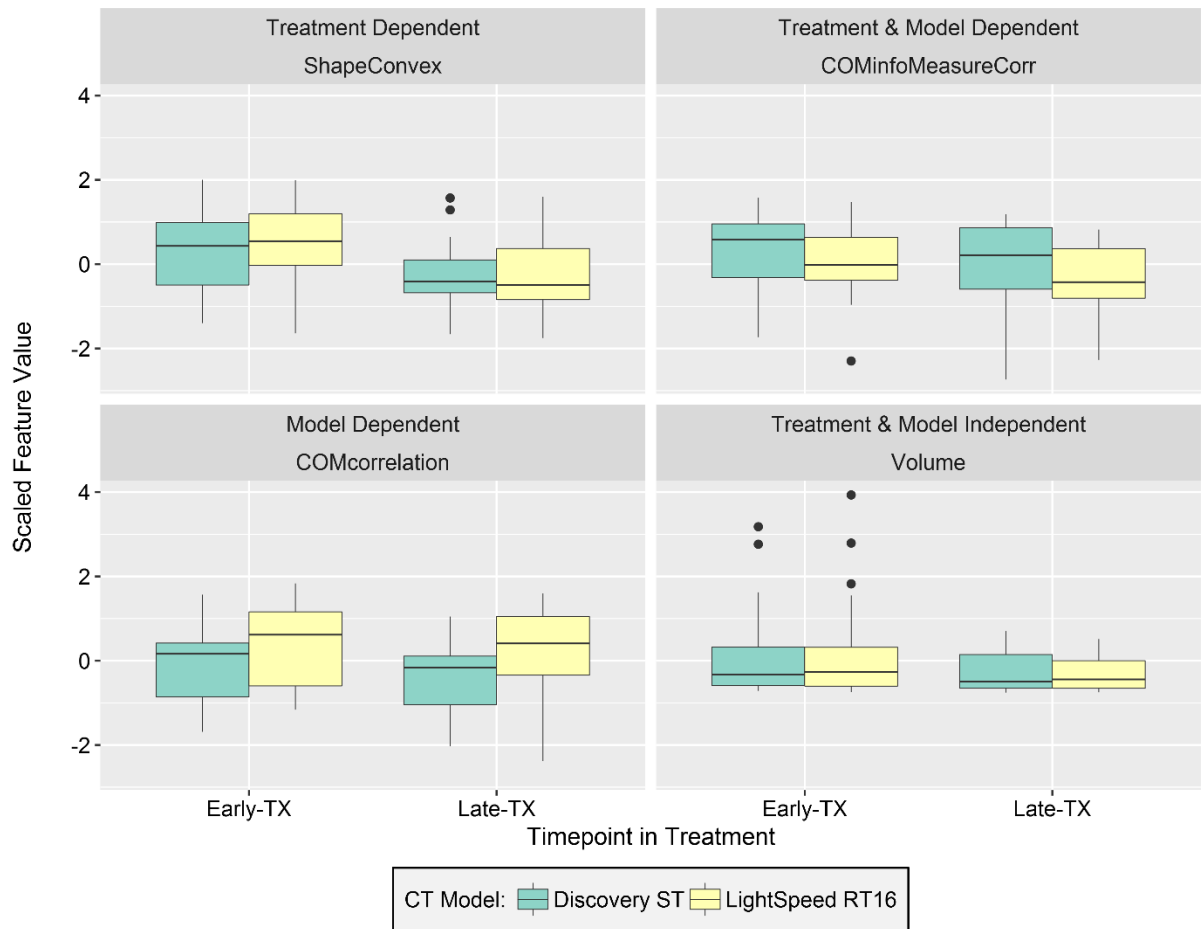


Figure 7.3: Example features for each of the ANOVA results. Boxplots of one feature from each category of the ANOVA results were plotted to demonstrate the relative difference in feature values with treatment (early-TX versus late-TX) and CT scanner model (Discovery ST versus LightSpeed RT16). Feature values were scaled for visualization in this figure by subtracting the mean and dividing by the standard deviation.

Table 7.5: P-values from the ANOVA analysis.

Feature	Treatment P-value	CT Model P-value	Classification
COMautocorrel	0.001	0.040	Treatment & Model Dependent
COMinfomc	0.001	0.011	Treatment & Model Dependent
COMinfomc2	0.005	0.018	Treatment & Model Dependent
COMsumavg	0.001	0.040	Treatment & Model Dependent
COMsumvar	0.001	0.246	Treatment Dependent
RLMhglre	0.001	0.045	Treatment & Model Dependent
RLMlglre	0.001	0.045	Treatment & Model Dependent
RLMsrlgle	0.001	0.333	Treatment Dependent
HISTenergynorm	0.001	0.045	Treatment & Model Dependent
HISTmean	0.001	0.045	Treatment & Model Dependent
HISTmed	0.001	0.108	Treatment & Model Dependent
SHAPEcompact	0.010	0.492	Treatment Dependent
SHAPEcompact2	0.005	0.125	Treatment & Model Dependent
SHAPEconv	0.006	0.675	Treatment Dependent
SHAPEround	0.005	0.308	Treatment Dependent
SHAPEspheredis	0.006	0.481	Treatment Dependent
SHAPEsphericity	0.004	0.142	Treatment & Model Dependent
SHAPEsurfareaden	0.001	0.948	Treatment Dependent
NDMtexstr	0.016	0.108	Model Dependent
COMcontrast	0.003	0.000	Treatment & Model Dependent
COMcorrel	0.063	0.001	Model Dependent
COMdissim	0.002	0.000	Treatment & Model Dependent
COMinvdifmn	0.006	0.471	Treatment Dependent
COMinvdifn	0.004	0.040	Treatment & Model Dependent
RLMglnunorm	0.024	0.127	Model Dependent
HISTstd	0.007	0.308	Treatment Dependent
HISTkurt	0.006	0.125	Treatment & Model Dependent
HISTvar	0.006	0.481	Treatment Dependent
COMdifent	0.002	0.000	Treatment & Model Dependent
RLMsre	0.002	0.011	Treatment & Model Dependent
HISTentropy	0.006	0.156	Treatment & Model Dependent
Volume	0.024	0.848	Treatment & Model Independent

### Treatment Dependent Features

The results of the Wilcoxon sign-rank test comparing patients' radiomics feature values at each week of treatment to their values at week 1 for the 10 treatment dependent features are shown in Figure 7.4. Of these features, 6 began to show significant changes in their values by week 2 and all of the subsequent weeks were also significantly different from the baseline. 3 features began to show significant changes by week 3 that were also consistent through the remainder of treatment. While 1 shape feature, roundness, was different by week 3 and throughout treatment with the exception of one mid-treatment week, week 5. Boxplots of the feature values at each week of treatment are shown in Figure 7.5.

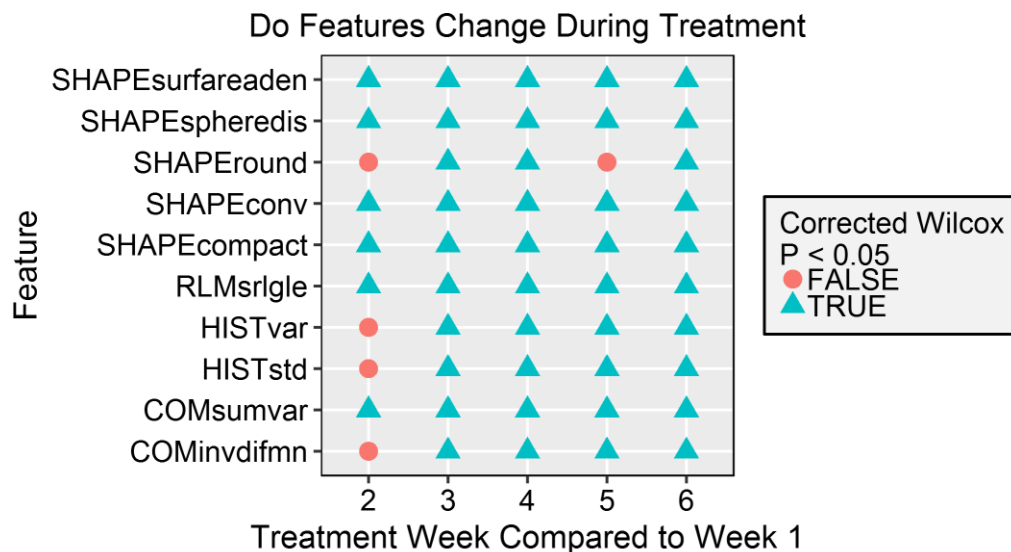


Figure 7.4: Results of Wilcoxon sign-rank test comparing patients' radiomics feature values at each week of treatment to their values at week 1 for the treatment dependent features. The majority of these features began to significantly change by treatment weeks 2 and 3.

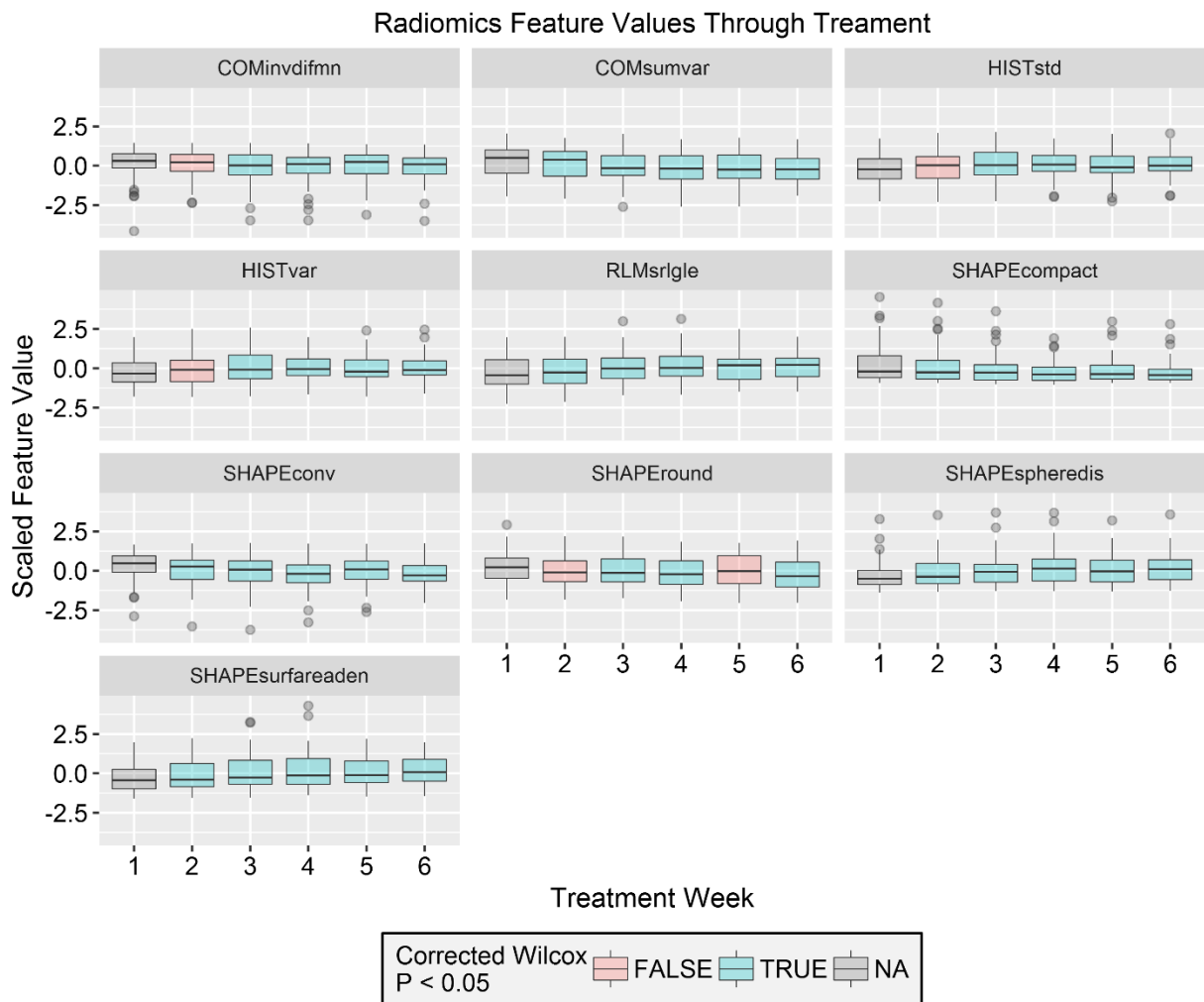


Figure 7.5: Boxplots of the radiomics features values through treatment for the treatment dependent features. The radiomics feature values were scaled by subtracting the mean and dividing by the standard deviation. Each boxplot is colored based on whether the change from treatment week 1 to that week was significant with red indicating that the change was not significant and blue indicating the change was significant.

When the values in the radiomics feature values at each week or when the net changes from week 1 were compared for the two treatment modalities, no feature demonstrated a significant difference between the two treatment modalities. Results for each test can be seen in Figure 7.6 and Figure 7.7. Boxplots for the feature values at each time point in treatment are in Figure 7.8.

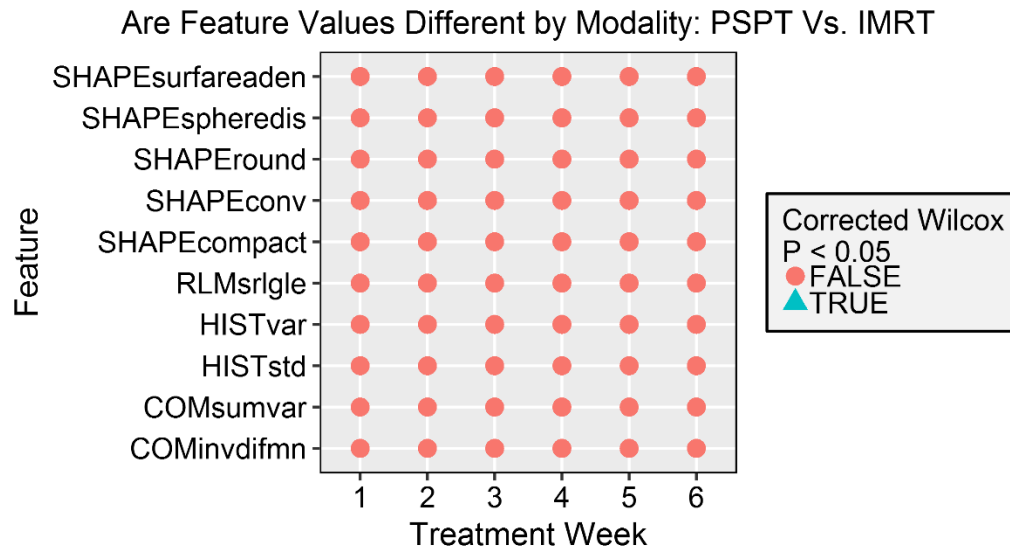


Figure 7.6: Results of Wilcoxon rank-sum test comparing the patients' radiomics feature values by treatment modality at each week for the treatment dependent features. At no point in treatment were the values in the two groups significantly different.

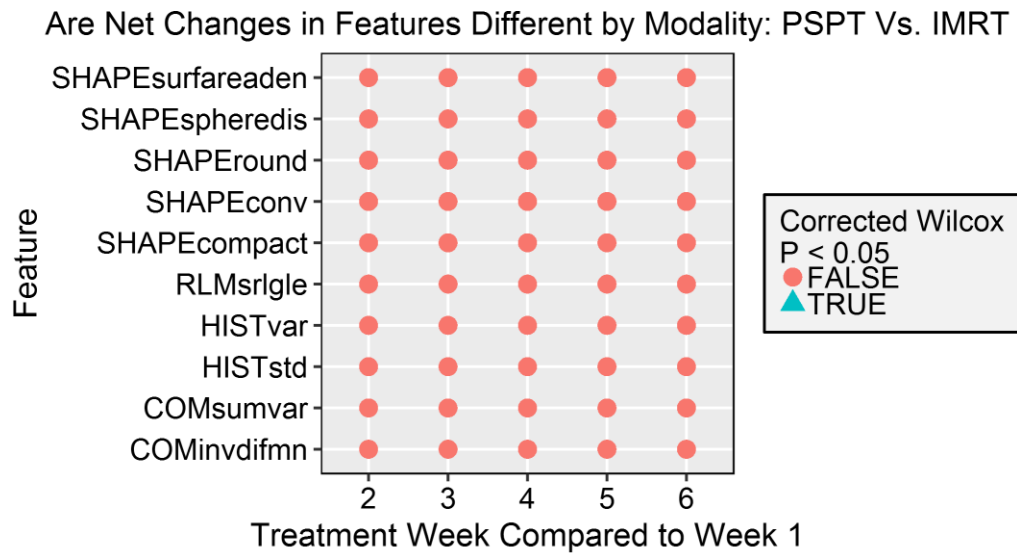


Figure 7.7: Results of Wilcox rank-sum test comparing the net changes between the patients' radiomics feature values at each week of treatment to their values at week 1 by modality for the treatment dependent features. At no point in treatment were the net changes significantly different for the two modalities.

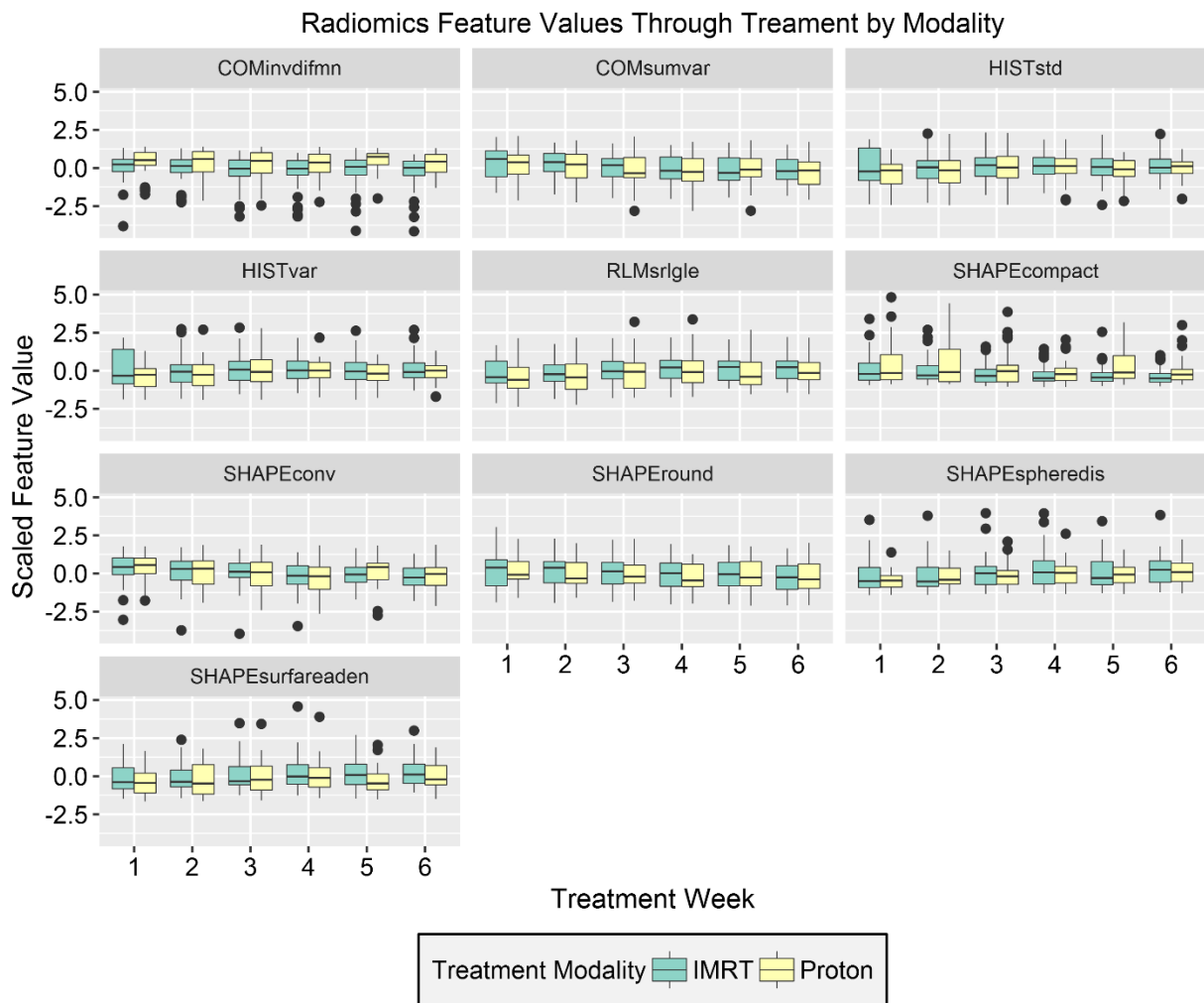


Figure 7.8: Boxplots of the radiomics feature values through treatment and by treatment modality for the treatment dependent features. The radiomics feature values were scaled by subtracting the mean and dividing by the standard deviation. At no point in treatment were the differences between the two modalities significant.

Treatment & Model Dependent Features

For the 18 treatment and model dependent features, the tertiary cohort was used for the analysis. Thus changes were only evaluated between weeks 0 and week 5. The results of the Wilcox sign-rank test comparing patients' radiomics feature values at week 5 to their values at week 0 are shown in Figure 7.9. All of these features, showed significant changes at week 5. Boxplots of the feature values at each week of treatment are shown in Figure 7.10.

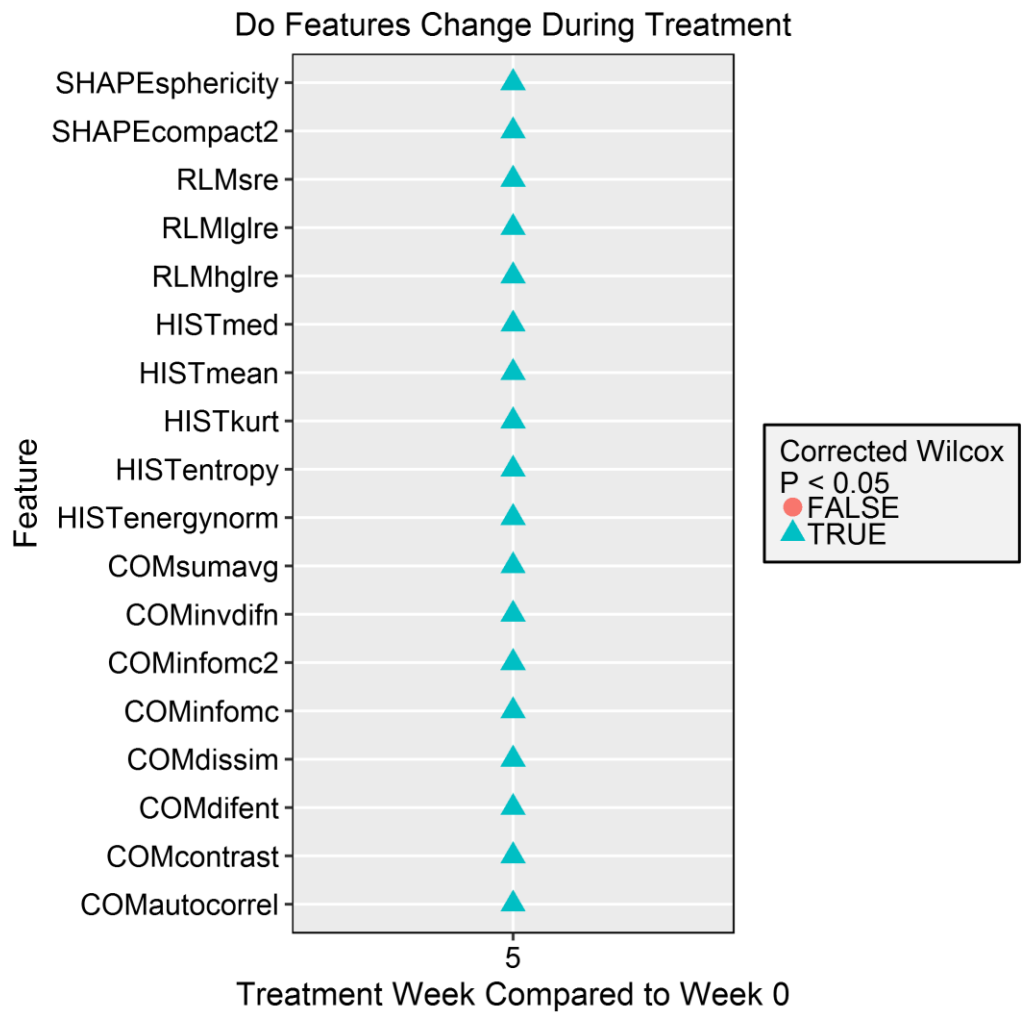


Figure 7.9: Results of Wilcox sign-rank test comparing patients' radiomics feature values at week 5 to their value at week 1 for the treatment & model dependent features. All of the features demonstrated a significant change by mid-treatment.

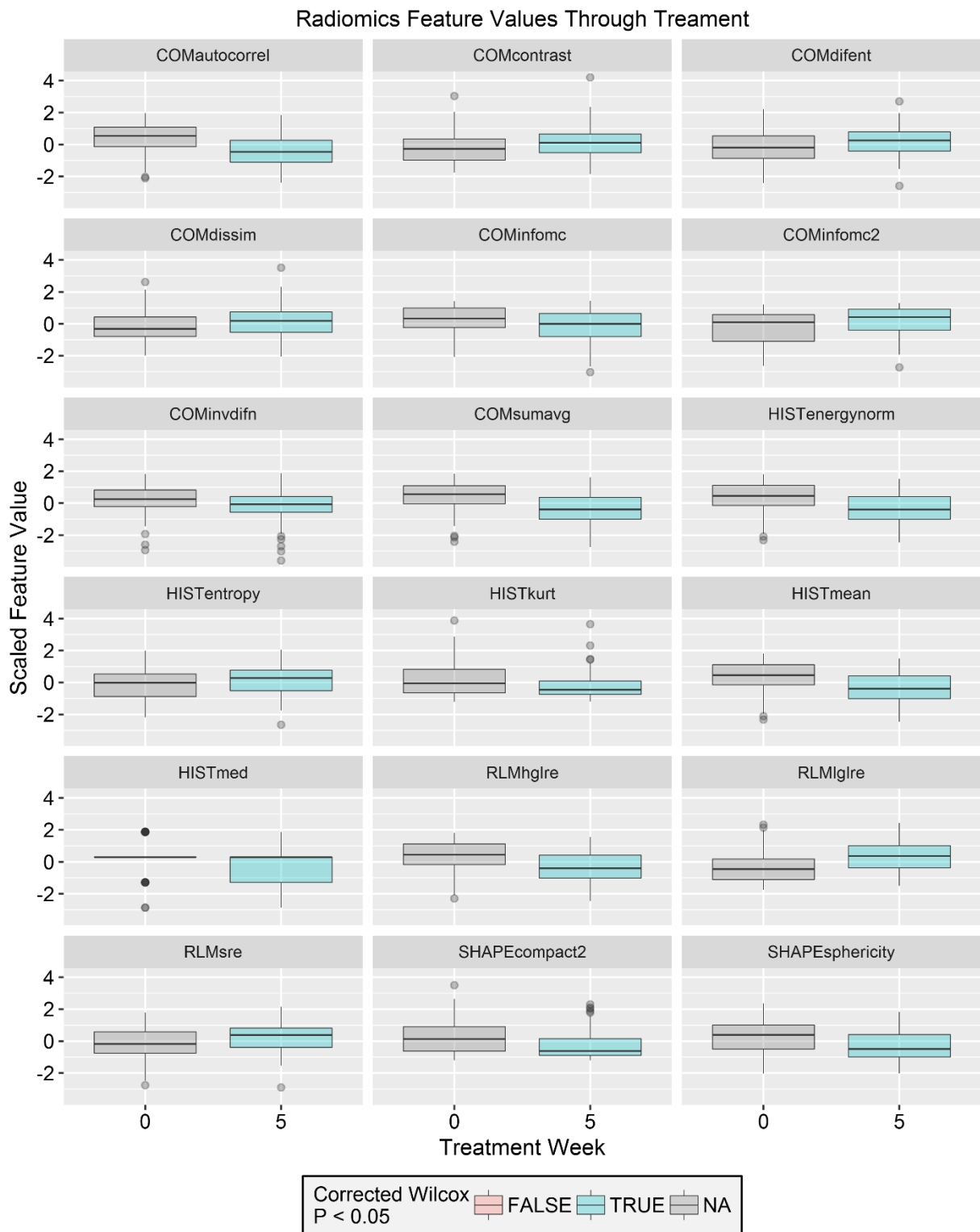


Figure 7.10: Boxplots of the radiomics features values at weeks 0 and 5 for the treatment & model dependent features. The radiomics feature values were scaled by subtracting the mean and dividing by the standard deviation. Each boxplot is colored based on whether the change

from treatment week 0 was significant with red indicating that the change was not significant and blue indicating the change was significant.

When the values in the radiomics feature values at week 0 and week 5 or when the net changes from week 0 were compared for the two treatment modalities, no feature demonstrated a significant difference between the two treatment modalities. Results for each test can be seen in Figure 7.11 and Figure 7.12. Boxplots for the feature values at each time point in treatment are in Figure 7.13.

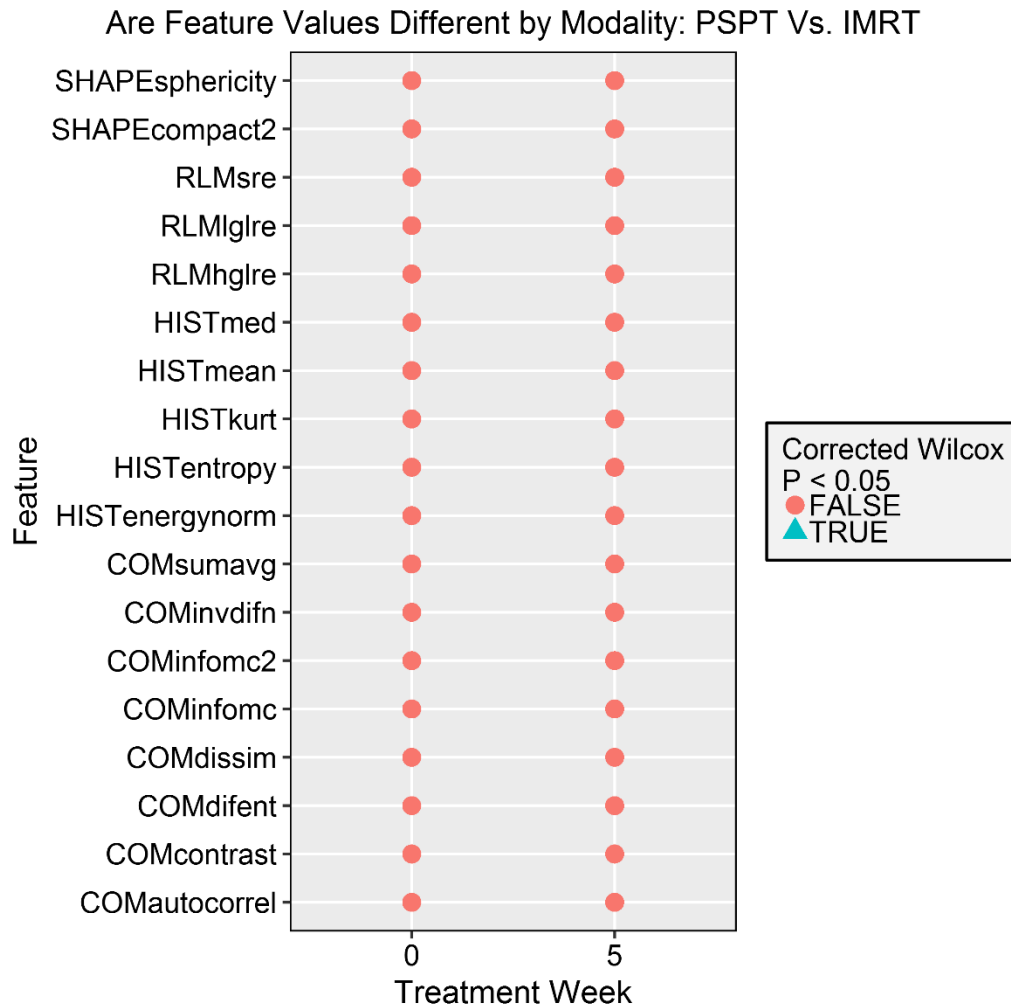


Figure 7.11: Results of Wilcox rank-sum test comparing the patients' radiomics feature values by treatment modality at weeks 0 and 5 for the treatment & model dependent features. At no point in treatment were the values in the two groups significantly different.

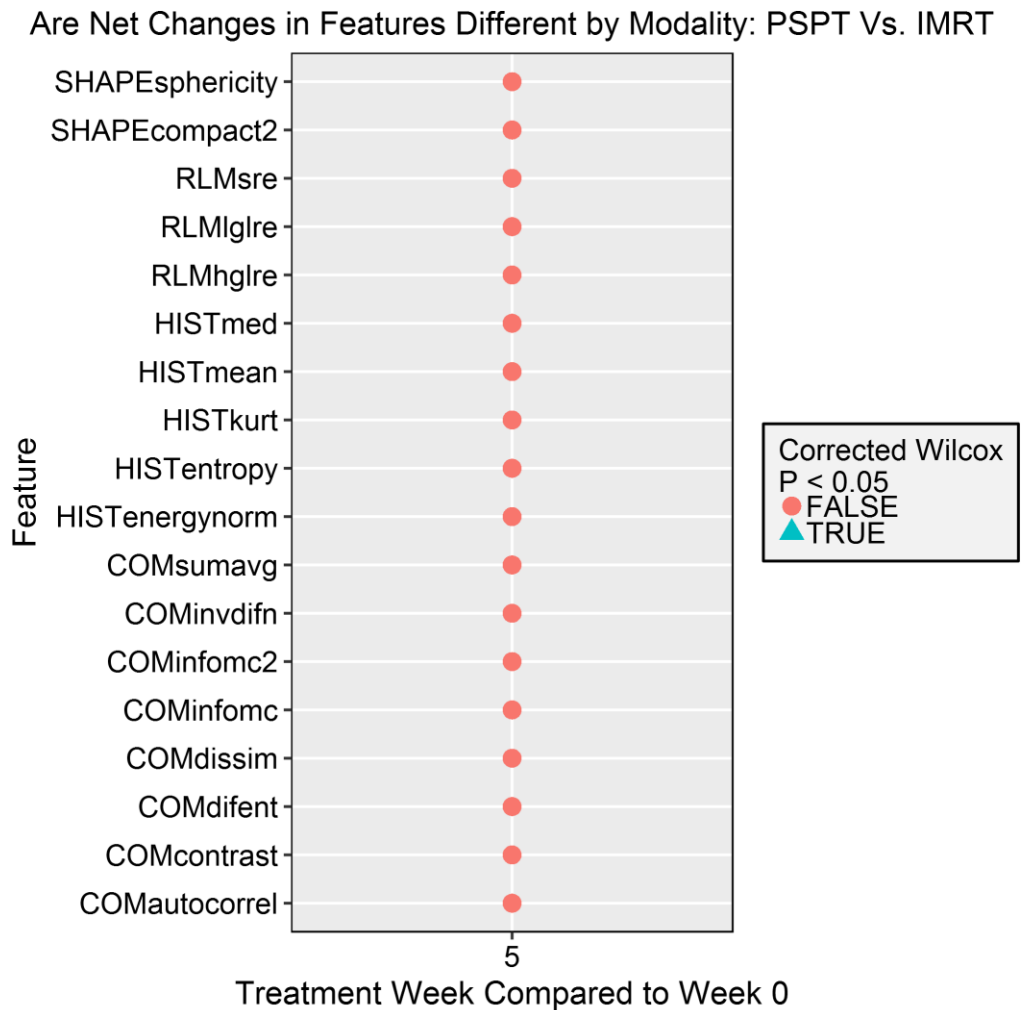


Figure 7.12: Results of Wilcoxon rank sum test comparing the net changes between the patients' radiomics feature values at week 0 to week 5 by modality for the treatment & model dependent features. At no point in treatment were the net changes significantly different for the two modalities.

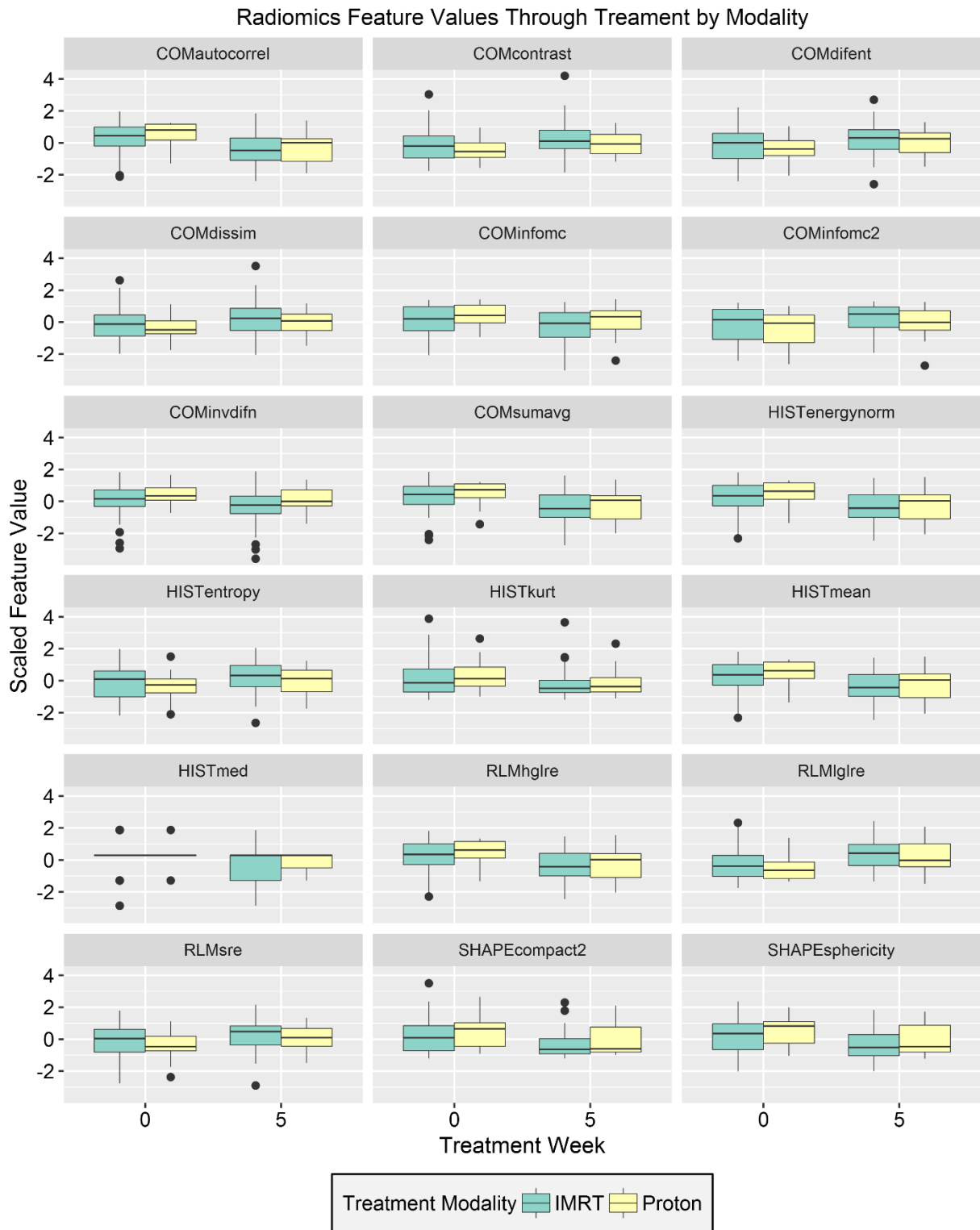


Figure 7.13: Boxplots of the radiomics feature values through treatment and by treatment modality for the treatment & model dependent features. The radiomics feature values were scaled by subtracting the mean and dividing by the standard deviation.

## Discussion

The goal of this analysis was to determine which radiomics features exhibited significant therapy-induced changes during treatment, how early in treatment these changes occurred, and whether they differed between patients who received PSPT and those who received IMRT. A subset of features that exhibited significant, consistent changes as early as 2-3 weeks into treatment was found. This is the same treatment window reported by radiomics and imaging biomarker studies in other imaging modalities including FDG-PET<sup>96,97</sup>, diffusion weighted MRI<sup>98</sup>, kV CT-on-rails<sup>99</sup>, and perfusion-CT<sup>100</sup>. Thus this may be the earliest possible detection of tumor response for conventionally fractionated radiotherapy. The fact that radiomics features demonstrated changes early in treatment for all patients may indicate that radiomics features are related to early biologic responses to radiation therapy. It is also possible that patients whose radiomics features change substantially from baseline early in treatment may be classified as having a better response to radiation therapy. However, we explored this hypothesis in Chapter 5 with this same patient set and found that changes in radiomics features were not more prognostic than the feature values at the beginning of treatment or typical clinical factors.

Interestingly, there were no significant differences between feature changes for patients treated with protons and those treated with photons. A hypothesis of this study was that the tumors treated with protons would demonstrate changes earlier in treatment as a result of the higher RBE. This negative result may indicate that the use of the value 1.1 for proton RBE accurately accounts for differences in dose deposition characteristics of proton therapy compared to photon therapy in the tumor. Additionally for passive scattering proton therapy, an increase in RBE above 1.1 may only occur at distal edges of the beams which are almost always beyond the distal edge of the target volume and into the surrounding normal tissues. Further, the net changes in radiomics features throughout the course of treatment, while statistically significantly different from baseline, were small and may not be sensitive enough to

detect even smaller differences owing to radiation therapy modality. Other sources of uncertainty, such as respiratory motion and inter-fractional anatomy changes, may also have further obscured the differences if they exist.

In this study, we focused on radiomics features that demonstrated changes throughout therapy. However, during the ANOVA analysis, a set of features that were consistent throughout treatment was also identified. The fact that this set of features did not change during therapy most likely means that they are not measuring anything meaningful from the tumor and are likely dominated by image noise.

Strict criteria were used in the ANOVA analysis for the classification of features to reduce the feature dimensionality in the subsequent analyses and thus the multiplicity correction penalization. Thus it is possible that some of the features classified by the ANOVA as treatment independent, do in fact change during treatment but that their changes are smaller than those features that were classified as treatment-dependent. For example, Volume which is well known to change during treatment<sup>101–103</sup> was classified as both treatment & model independent because its corrected p-value for treatment was 0.024 and for model was 0.85. If our criteria for treatment dependence had been loosened to 0.05 instead of 0.01, Volume would have been included in the treatment-dependent group and 2 of the model dependent features would have been included in the treatment and model dependent group. Because the main goal of this analysis was to determine whether features demonstrate modality-specific differences, it was more important to reduce the feature dimensionality in order to allow for the possibility of finding those small changes than to explore when in treatment each feature changes.

This study had a few limitations. First, the dataset was relatively small. Also, the number of patients with CT images at each dose point differed as a consequence of the data condition used to eliminate variation introduced by CT scanner model. This difference in the number of patients being compared at different points in treatment may have been a cause of the fact that

some features demonstrated significant changes at certain points during treatment but not throughout the treatment. Additionally while we controlled for differences between scanners, there are also known variabilities in radiomics features measured from test-retest patients on the same scanner<sup>104</sup>. Thus some of the measured differences in treatment could be due to variability in the features and not changes in the tumor. However this possibility seems small as for all the patients, the features changed in the same direction (increased or decreased) over the course of treatment. If the values were different only because of test-retest variability then some of the patients' values should have decreased and some of the patients' values should have increased leading to a negligible net effect for the patient population.

## **Conclusions**

A set of radiomics features that exhibited significant radiation-induced changes over the course of treatment was identified. At no point in treatment did these changes significantly differ between patients treated with IMRT and those treated with proton therapy. Thus if any changes in tumor biology are different based on treatment modality, they may not be large enough to be measured using radiomics features.

## Chapter 8 : Discussion

The main goal of this study was to determine whether radiomics features calculated from computed tomography (CT) images of non-small cell lung cancer (NSCLC) patients change over the course of treatment, and if those changes were prognostic for outcome. We also independently investigated if those features were reproducible in CBCT images, and if the changes in features were different based on treatment modality. To accomplish this, we began in Aim 1 by evaluating the impact of different image preprocessing techniques on the radiomics features. Then in Aim 2 we used the results of Aim 1 to curate a set of radiomics features with feature-specific preprocessing and evaluated their significance in univariate and multivariate models. In Aim 3, radiomics features were measured from CBCT images of a texture phantom to determine the influence of different imagers, scatter, and motion on their reproducibility. Finally in Aim 4, we pinpointed when the features first began to change and if that rate or magnitude of change was ever different between patients treated with protons versus photons.

In Aim 1 we calculated radiomics features from the pre-treatment images using a variety of different image preprocessing techniques. For each feature and image preprocessing combination we calculated the feature's dependence on volume, dependence on CT model, and prognostic significance. While other groups have investigated the impact of gray-level discretization on the reproducibility of features<sup>44</sup>, we believe we are the first to propose these metrics for deciding which version of a feature is the most useful. We found that most features were more likely to be prognostic and independent of CT model if images were smoothed before feature calculation either alone or in conjunction with a bit depth resampling step. However the individual response was feature specific even within a category of features such as the co-occurrence matrix. In addition to this main result we also identified five features that were inherently volume dependent and proposed corrective factors for each. These corrected features can be of use for future studies or can serve as a precautionary tale of some of the

possible weaknesses in using features not designed for analyzing medical images. Then we designed a set of digital phantoms that can be easily used to test a feature's independence from volume or evaluate a user's own understanding of a particular feature.

In Aim 2, the results of Aim 1 were used to design a methodology for selecting feature-specific image preprocessing for each radiomics feature. Then this set of features was calculated from weekly 4DCT images to evaluate the ability of delta-radiomics features to predict overall survival, time until distant metastases, and time until local-regional recurrence in univariate and multivariate studies. In contrast to previous studies using the relative net changes in features, this study included the slope of a linear regression of the feature, and the value at the end of treatment as independent covariates for analysis. While there was strong evidence that the features changed during treatment, these changes did not translate into strong prognostic factors in either univariate or multivariate models. The delta-radiomics features were able to significantly improve the model fit for overall survival but did not substantially improve its performance. For time until distant metastases, only clinical factors and radiomics features measured at pre-treatment were significant for both model fit and performance. For time until local-regional recurrence, the final model included only one covariate and it was texture-strength measured from the neighborhood difference matrix at the end of treatment. Using this univariate model resulted in significant stratification of high and low-risk patients. While these results did not provide strong evidence of the prognostic potential of delta-radiomics features they were the first to investigate these features in NSCLC. Additionally a framework for model building was generated which could be used with different features designed to measure tumor heterogeneity or to investigate features in different image modalities such as contrast-CT. Finally, one pre-treatment feature, compactness 2, was successful at improving the prognostic performance of models predicting overall survival or time until distant metastases. This supports several other analyses which have also found this feature to be useful<sup>23,105</sup>.

In Aim 3, we investigated whether features could be measured from CBCT images because this would facilitate the aggregation of larger datasets to measure delta-radiomics features in future studies. Almost half of the features we studied were not reproducible even in a patient test-retest cohort where the same patient was imaged on the same CBCT scanner on the same day. We found that the remaining features were highly impacted by differences in motion, scatter, and imaging equipment when measured from a radiomics phantom under varying conditions. Thus it will be important for any future studies seeking to use radiomics features calculated from CBCT images to keep their imaging protocols as uniform as possible and minimize motion. This was the first, and to our knowledge only, study using a radiomics phantom to evaluate variability in radiomics features for CBCT images. This is an important landmark because when the different sources of uncertainty are understood they can be accounted for within the setup or analysis.

In Aim 4, we pinpointed when in treatment the features first began to significantly change from baseline and determined if there was any difference in the rate or magnitude of the changes for patients treated with PSPT or IMRT. We established that a subset of features showed significant changes after 2-3 weeks of treatment but that there was no difference in either the rate or magnitudes of these changes between patients treated with protons versus photons. While this was a negative result, it did support the choice of 1.1 for the RBE when calculating equivalent dose for proton treatments.

One limitation that was prevalent through all of these studies was the patient cohort size. This became an even larger limitation when we tried to remove some of the variability in the dataset prior to performing measurements. For example, the initial cohort of 157 patients was decreased to 107 patients after several exclusion criteria (images acquired with 4DCT not breathhold, primary GTV available in the treatment plan, and tumor volume>5cc) were implemented prior to each of the aims. But then in Aim 4, this set was even further reduced to 73 patients in order to control for the fact that different patients had been imaged on different

CT scanners throughout their treatment. This removed variability from the scanners but also reduced the power of each statistical test and thus made it more difficult to find a significant difference if one existed. A similar limitation was the lack of an independent validation set which would have been particularly useful for Aims 1 and 2 where univariate and multivariate cox proportional hazard models were built. We balanced this limitation by performing leave-one-out cross validation in these instances.

The segmentation of the lesions may also have been a weakness in this study. We defined the tumor ROI on each of the weekly images by deforming the primary GTV contour from the treatment plan to each of the weekly images using an in-house, clinically validated deformation software called CT-Assisted Targeting. Then I went through each of the images and modified the contour to ensure consistency. While the vast majority of contours required no edits by hand, many of the deformed contours for tumors connected to the mediastinum would stray into it and thus slight manual edits would be necessary. In these cases, I erred on the side of caution and would decrease the volume encapsulated by the contour to what I was positive was the same area that had been identified as tumor in the treatment planning images. This step would have benefited from input from a trained radiation oncologist familiar with thoracic tumors. The segmentation can impact all of the radiomics features but is especially a concern for the shape-based features. As mentioned in Chapter 5, those features that bordered the mediastinum often were edited to have one smooth side as it was not clear where the tumor ended and the mediastinum began and this smooth side could affect the shape based radiomics features to have values closer to smooth floating style tumors that are typically less aggressive than spiculated tumors.

Another limitation of this study, is the uncertainty about which radiomics features should be used and how they should be calculated. There are hundreds of radiomics features that are currently being explored by different groups and a practically infinite number of ways in which each can be calculated. So far none of these features have been directly linked to a ground

truth such as a genetic mutation. Thus it is impossible to know whether the subset of features used in a study are the best features that could be used or if they are being calculated in a way that would make them useful. Many radiomics studies handle this uncertainty by beginning with a large set of features and whittling them down to those that are reproducible and cover a large dynamic range in a test-retest patient set<sup>10,42,55,106</sup>. These are important steps however reproducibility does not guarantee that the feature is linked to biology or even that it is independent from volume. In this study we attempted to shed some light on these uncertainties by calculating each feature after multiple different image preprocessing techniques and evaluating the impact on the feature's volume dependence, CT model dependence, and prognostic significance. We then developed a method for selecting the optimal feature preprocessing for each feature that was based largely on its prognostic ability when calculated from pre-treatment images. While this allowed us to establish a pseudo-ground truth it also may have biased the remainder of our analyses which sought to find the prognostic significance of the delta radiomics version of these features. The original requirement that used the pre-treatment version of the features was chosen for two reasons: first, because several publications have shown that pretreatment radiomics features have informative value and thus changes in the features that are already prognostic may reflect actual biological changes in the tumor, and second, if model building was limited to delta-radiomics features that were significant in univariate analyses the results could also be biased and overly optimistic.

## **Future Directions**

One avenue for future research would be the use of auto-segmentation to contour the tumors in this analysis. A study by Parmar *et al* demonstrated that feature reproducibility increased when contours were segmented using a semiautomatic region growing algorithm<sup>107</sup>. This result is not unexpected as auto segmentation typically leads to more reproducible contours and the radiomics features are all influenced by the specific voxels contained within the contour. This reproducibility is of especial importance in longitudinal studies such as this

one, to make sure that the weekly contours are all defined in a systematic and reproducible way so that the measured changes in the features are true changes in the tumor and are not due to changes in the contours.

In this analysis all of the studied features were calculated globally. What this means is that the final value is representative of the entire 3D tumor ROI. Thus if the tumor is large and mainly homogenous with one small area that is heterogeneous, the signal from the heterogeneous area can be lost amongst all the signal from the substantially larger homogenous area. In future studies, features could be calculated locally instead to prevent this effect. When features are calculated locally a pre-defined window is established (such as a 5mm x 5mm window) and then this window is slid over the tumor and at each interval the features are recalculated for only the voxels within the window. The maximum or minimum value for the feature from any of the windows can then be calculated and compared between patients. So if for example, the feature was entropy from the co-occurrence matrix, it could be more meaningful to evaluate whether a patient has a single region with high entropy at the beginning of treatment. Then if the entropy value of this region decreases during treatment, that decrease could signify that the most heterogeneous region of the tumor was destroyed during treatment.

In order for radiomics to move from retrospective analyses to prospective clinical trials and eventually alter patient treatment, the features must be standardized. This is an important area of research that has been largely neglected in the past ten years of radiomics publications as mentioned in many recent review articles<sup>34,68,108–112</sup>. The image biomarker standardization initiative was recently formed to help tackle this important challenge and has already begun by defining a list of features and their algorithms in specific detail<sup>113</sup>. A future step could be the development of a set of test images to be available for download online along with a table of calculated values for each of these standard features from the test images. This would allow different researchers to calculate these same features with their own software and ensure that

they are obtaining the same values. Then multi-institutional studies could be performed that aggregate feature values from those institutions that have shown they can correctly measure these features.

Another interesting area of future research would be the designing of new features specifically for the analysis of medical images in radiomics studies. The majority of the features used in the work presented here were designed in the 1970's or 1980's for comparing satellite images<sup>11-14</sup>. As a result they may not be sensitive enough to pick up the relatively smaller differences in heterogeneity or homogeneity in different tumors. They also, as we demonstrated in Aim 1, are more likely to be sensitive to differences in volumes between the two tumors being compared. Recent work by Prasanna and Tiwari is perhaps the start of a new wave of features designed and validated for use specifically in radiomics<sup>114</sup>. They designed a feature they named co-occurrence of local anisotropic gradient orientations (CoLIAGE) and evaluated its ability to distinguish benign and malignant phenotypes in T1 weighted MRIs of the brain (classified radiation necrosis versus recurrent tumor), DCE-MRI of the breast (classified two molecular sub-types), and non-contrast CT images of the lung (classified granulomas versus adenocarcinomas)<sup>114</sup>. In every case, this feature outperformed classical radiomics features including the Haralick co-occurrence matrix features analyzed in this study.

One effort that could aid in designing useful features would be to begin by imaging and contouring healthy tissues and comparing the values. Ideally, a set of features that can differentiate between tissues such as the liver, lungs, kidney, muscle, blood in the heart, and brain would be identified. If a feature cannot differentiate between these tissues which we know are very biologically different then it is very unlikely that it would be able to pick out the more subtle differences in heterogeneity of two lung tumors. Additionally, the ROIs for this evaluation could all be kept the same size so that there is no impact from volume and multiple ROIs of that size could be drawn in each tissue in the patient to assess both intra- and inter- patient variability. Once a set of features that can differentiate these or other tissues was found, the

experiment could be repeated using different size ROIs up to encapsulating the entire healthy tissue to evaluate the effect of volume on these now known values. Once the normal range of feature values in healthy tissue is well characterized, similar ROIs could be drawn on tumor tissue and compared. For example, values from healthy liver could be compared to liver lesions. Taking this one step further, these features could possibly even be used to train a machine learning algorithm to classify voxels as healthy or tumorous tissue and thus allow for the eventual auto-segmentation of tumors using radiomics features as was demonstrated by Markel *et al*/ using the classical features<sup>115</sup>.

In addition to designing features specifically for radiomics analyses, designing imaging protocols with the sensitivity necessary to pick up the tumor heterogeneity differences we are trying to measure could be useful. However one of the strengths of radiomics is that these studies can make use of medical images that are already routinely acquired. This makes it easier to collect the large patient datasets which are necessary for high-throughput feature analysis. Additionally imaging protocols are already carefully designed to balance visualization needs with minimum radiation dose to the patient and short acquisition times. Thus instead of changing the actual imaging protocols, it would be more feasible to harmonize the resulting images prior to radiomics analysis. In this study, we achieved this harmonization by filtering the images with a smoothing algorithm or bit depth resample in order to remove differences between the two main scanners in this study, a GE LightSpeed RT16 and GE Discovery ST. However these images had still been acquired using the same protocol and from the same manufacturer. When images in the future are collated between institutions using different imaging protocols or a greater variety of scanners, the differences in the images will be larger<sup>40,41</sup> and more effort will be needed to make the images equivalent via post-processing. Our group is currently evaluating this need by imaging a new radiomics phantom on over 100 CT scanners in Texas using the default imaging protocol at each clinic. The results could be

used to calibrate the images so that the resulting radiomics features are standardized and open the door for larger multi-institutional radiomics studies.

Finally, while this study demonstrated that delta-radiomics features are not significantly prognostic for NSCLC when measured from CT images, a similar study using contrast-CT images could have better results. Contrast-CT images are acquired after iodine is injected as a contrast agent into the patient's blood stream intravenously<sup>116</sup>. The resulting images highlight the vascularity of the tumor and allow for regions of necrosis to be identified. Thus radiomics features measured from these images may be more meaningful than features measured from non-contrast CT where the inter-tumor differences are more subtle. Radiomics features from contrast-CT images have been successfully used to predict tumor response for soft tissue sarcomas<sup>39</sup> and predict patient outcomes for NSCLC<sup>26</sup> when measured at baseline. Additionally many features have been shown to be reproducible to differences in time between injection and scan acquisition in contrast-CT<sup>71,117</sup>. While no radiomics studies have used delta-radiomics features from contrast-CT images yet, one study by Lind *et al* showed that significant decreases in the blood flow within NSCLC tumors treated with anti-angiogenic chemotherapy<sup>100</sup> could be measured from serial contrast-CT scans. Thus using more sophisticated metrics such as radiomics features could further improve our understanding of the changes taking place in NSCLC tumors over the course of chemoradiotherapy and aid in predicting prognosis. The same analysis methods presented in our study could be applied to this scientific question for both the development of features and evaluation of their utility.

## Conclusions

In this study we carefully constructed a set of features that were not dependent on CT model or volume through the use of different image preprocessing techniques to evaluate whether changes in features, delta-radiomics, were prognostic for NSCLC patients. We found strong evidence that radiomics features change over the course of treatment. These changes began to be significant as early as the second or third week of treatment for many features and

were independent of treatment modality. However, both in univariate and multivariate models, changes in features were less likely to be significant prognostic factors than their counterparts measured from pre-treatment images or than classical clinical factors. Two features that were prognostic were compactness 2 measured at pre-treatment for overall survival and time until distant metastases and texture-strength measured at the end of treatment for local-regional recurrence. When features were measured from CBCT images they varied substantially with the choice of imaging unit, amount of scatter, and range of motion. Our results demonstrate the importance of establishing uniform imaging protocols for use in radiomics studies as well as developing rigorous standards for model building in radiomics analyses.

## Appendix A: IBEX Parameter Sheets

When radiomics features are calculated using the IBEX software, the output includes three excel sheets. The first sheet has the quantitative feature value for each image and feature, the second sheet has the image characteristics such as the voxel size, and the third sheet includes all of the parameters that were used to calculate the features. The following two tables in this appendix are the parameter sheets for the feature sets used in Aims 2-4. The parameter sheet is made up of 6 columns in the following order: Category, Parameters, Feature, Parameters, Preprocess, and Parameters. The three 'Parameters' columns refer respectively to the parameters used for the Category, Feature, and Preprocess columns. These columns are grouped under 'FeatureItem' because in IBEX, the user begins by selecting a feature category such as Shape. This category becomes the 'FeatureItem' in the parameter sheet. Then under that particular category a number of different features can be selected and calculated using the same category parameters and imaging preprocessing parameters. This saves the user having to reselect these for each individual feature. Thus in Table A1, the 7 features listed under 'FeatureItem-1' are all shape features that were calculated with the 'Threshold\_Image\_Mask' preprocessing technique. For more details please refer to the paper by Zhang *et al* on IBEX<sup>50</sup>.

Table A1: The feature and parameters sheet produced by the radiomics analysis software IBEX for the calculation of the radiomics features used in Aim 2 and 4.

FeatureItem-1					
Category	Parameters	Feature	Parameters	Preprocess	Parameters
Shape		Compactness1		Threshold_Image_Mask	ThresholdLow=900; ThresholdHigh=1200; ErosionDist=0;
		Compactness2			
		Convex			
		Roundness			
		SphericalDisproportion			
		Sphericity			
		SurfaceAreaDensity			
FeatureItem-2					
Category	Parameters	Feature	Parameters	Preprocess	Parameters
GrayLevelCooccurrenceMatrix25	Direction=0 45 90 135; AdaptLimitLevel=1; GrayLimits=0 4096; NumLevels=4096; Offset=1; Symmetric=0;	InformationMeasureCorr1		Threshold_Image_Mask	ThresholdLow=900; ThresholdHigh=1200; ErosionDist=0;
		InformationMeasureCorr2			

FeatureItem-3					
Category	Parameters	Feature	Parameters	Preprocess	Parameters
GrayLevelRunLengthMatrix25	Direction=0 90; GrayLimits=1 4096; NumLevels=4096;	GrayLevelNonuniformity		Butterworth_Smooth	cutoff=125; order=2; x_padded_size=512; y_padded_size=512; draw_before_after=0; images_folder=;
		HighGrayLevelRunEmpha		Threshold_Image_Mask	ThresholdLow=900; ThresholdHigh=1200; ErosionDist=0;
		LowGrayLevelRunEmpha			
		ShortRunLowGrayLevelEmpha			
FeatureItem-4					
Category	Parameters	Feature	Parameters	Preprocess	Parameters
NeighborIntensityDifference25	NHood=5; NHoodSym=1; IncludeEdge=0; AdaptLimitLevel=1; RangeMin=0; RangeMax=4096; NBins=4096;	TextureStrength		Butterworth_Smooth	cutoff=125; order=2; x_padded_size=512; y_padded_size=512; draw_before_after=0; images_folder=;

			Threshold_Image_Mask	ThresholdLow=900; ; ThresholdHigh=1200; ErosionDist=0;	
FeatureItem-5					
Category	Parameters	Feature	Parameters	Preprocess	Parameters
IntensityDirect	ThresholdLow=1; ThresholdHigh=8000; ErosionDist=0; OnlyUseMaxSlice=0;	EnergyNorm		Butterworth_Smooth	cutoff=125; order=2; x_padded_size=512; y_padded_size=512; draw_before_after=0; images_folder=;
		GlobalMean		Threshold_Image_Mask	ThresholdLow=900; ; ThresholdHigh=1200; ErosionDist=0;
		GlobalStd			
		Kurtosis			
		Variance			
FeatureItem-6					
Category	Parameters	Feature	Parameters	Preprocess	Parameters

GrayLevelCooccurrenceMatrix25	Direction=0 45 90 135; AdaptLimitLevel=1; GrayLimits=0 4096; NumLevels=4096; Offset=1; Symmetric=0;	AutoCorrelation		Butterworth_Smooth	cutoff=125; order=2; x_padded_size=512; y_padded_size=512; draw_before_after=0; images_folder=;
		Contrast		Threshold_Image_Mask	ThresholdLow=900; ThresholdHigh=1200; ErosionDist=0;
		Dissimilarity			
		InverseDiffMomentNorm			
		InverseDiffNorm			
<b>FeatureItem-7</b>					
<b>Category</b>	<b>Parameters</b>	<b>Feature</b>	<b>Parameters</b>	<b>Preprocess</b>	<b>Parameters</b>
GrayLevelRunLengthMatrix25	Direction=0 90; GrayLimits=1 256; NumLevels=256;	ShortRunEmphasis		Butterworth_Smooth	cutoff=125; order=2; x_padded_size=512; y_padded_size=512; draw_before_after=0; images_folder=;
				Threshold_Image_Mask	ThresholdLow=900;

					ThresholdHigh=1200; ErosionDist=0;
					BitDepthRescale_Range RangeMin=0; RangeMax=4096; RangeFix=1; BitDepth=8;
FeatureItem-8					
Category	Parameters	Feature	Parameters	Preprocess	Parameters
IntensityDirect	ThresholdLow=1; ThresholdHigh=8000; ErosionDist=0; OnlyUseMaxSlice=0;	GlobalEntropy	NBins=256; RangeMin=0; RangeMax=4096; RangeFix=0;	Butterworth_Smooth	cutoff=125; order=2; x_padded_size=512; y_padded_size=512; draw_before_after=0; images_folder=;
		GlobalMedian		Threshold_Image_Mask	ThresholdLow=900; ThresholdHigh=1200; ErosionDist=0;
				BitDepthRescale_Range	RangeMin=0; RangeMax=4096; RangeFix=1; BitDepth=8;
FeatureItem-9					
Category	Parameters	Feature	Parameters	Preprocess	Parameters

GrayLevelCooccurrenceMatrix25	Direction=0 45 90 135; AdaptLimitLevel=1; GrayLimits=0 256; NumLevels=256; Offset=1; Symmetric=0;	Correlation	Butterworth_Smooth	cutoff=125; order=2; x_padded_size=512; y_padded_size=512; draw_before_after=0; images_folder=;
	DifferenceEntropy		Threshold_Image_Mask	ThresholdLow=900; ThresholdHigh=1200; ErosionDist=0;
	SumAverage		BitDepthRescale_Range	RangeMin=0; RangeMax=4096; RangeFix=1; BitDepth=8;
	SumVariance			

Table A2: The feature and parameters sheet produced by the radiomics analysis software IBEX for the calculation of the radiomics features used in Aim 3.

FeatureItem-1					
Category	Parameters	Feature	Parameters	Preprocess	Parameters
IntensityDirect	ThresholdLow=850; ThresholdHigh=1200; ErosionDist=0; OnlyUseMaxSlice=0;	Energy		Threshold_Image_Mask	ThresholdLow=850; ThresholdHigh=1200; ErosionDist=0;
		GlobalEntropy	NBins=256; RangeMin=0; RangeMax=4096; RangeFix=0;		
		GlobalMax			
		GlobalMean			
		GlobalMedian			
		GlobalStd			
		GlobalUniformity	NBins=256; RangeMin=0; RangeMax=4096; RangeFix=0;		
		Kurtosis			
		Skewness			
FeatureItem-2					
Category	Parameters	Feature	Parameters	Preprocess	Parameters
IntensityDirect	ThresholdLow=1; ThresholdHigh=256; ErosionDist=0;	Energy		Threshold_Image_Mask	ThresholdLow=850;

OnlyUseMaxSlice=0;			ThresholdHigh=1200; ErosionDist=0;		
	GlobalEntropy	NBins=256; RangeMin=0; RangeMax=4096; RangeFix=0;	BitDepthRescale_ Range	RangeMin=0; RangeMax=4096; RangeFix=1; BitDepth=8;	
	GlobalMax				
	GlobalMean				
	GlobalMedian				
	GlobalStd				
	GlobalUniformity	NBins=256; RangeMin=0; RangeMax=4096; RangeFix=0;			
	Kurtosis				
	Skewness				
FeatureItem-3					
Category	Parameters	Feature	Parameters	Preprocess	Parameters
GrayLevelCooccurrenceMatrix25	Direction=0 45 90 135; AdaptLimitLevel=1; GrayLimits=0 2100; NumLevels=100; Offset=1; Symmetric=1;	AutoCorrelation		Threshold_Image_Mask	ThresholdLow=850 ; ThresholdHigh=1200; ErosionDist=0;
		ClusterProminence		BitDepthRescale_Range	RangeMin=0; RangeMax=4096; RangeFix=1; BitDepth=8;
		ClusterShade			
		ClusterTendency			
		Contrast			
		Correlation			

		DifferenceEntropy			
		Dissimilarity			
		Energy			
		Entropy			
		Homogeneity			
		Homogeneity2			
		InformationMeasureCorr1			
		InformationMeasureCorr2			
		InverseDiffMomentNorm			
		InverseDiffNorm			
		InverseVariance			
		MaxProbability			
		SumAverage			
		SumEntropy			
		SumVariance			
		Variance			
FeatureItem-4					
Category	Parameters	Feature	Parameters	Preprocess	Parameters
GrayLevelRunLengthMatrix25	Direction=0 90; GrayLimits=1 256; NumLevels=256;	GrayLevelNonuniformity		Threshold_Image_Mask	ThresholdLow=850; ; ThresholdHigh=1200; ErosionDist=0;
		HighGrayLevelRunEmpha		BitDepthRescale_Range	RangeMin=0; RangeMax=4096; RangeFix=1; BitDepth=8;
		LongRunEmphasis			
		LongRunHighGrayLevelEmpha			

		LongRunLowGrayLevelEmpha			
		LowGrayLevelRunEmpha			
		RunLengthNonuniformity			
		RunPercentage			
		ShortRunEmphasis			
		ShortRunHighGrayLevelEmpha			
		ShortRunLowGrayLevelEmpha			
<b>FeatureItem-5</b>					
<b>Category</b>	<b>Parameters</b>	<b>Feature</b>	<b>Parameters</b>	<b>Preprocess</b>	<b>Parameters</b>
NeighborIntensityDifference25	NHood=5; NHoodSym=1; IncludeEdge=0; AdaptLimitLevel=1; RangeMin=0; RangeMax=4096; NBins=256;	Busyness		Threshold_Image_Mask	ThresholdLow=850; ThresholdHigh=1200; ErosionDist=0;
		Coarseness		BitDepthRescale_Range	RangeMin=0; RangeMax=4096; RangeFix=1; BitDepth=8;
		Complexity			
		Contrast			
		TextureStrength			
<b>FeatureItem-6</b>					
<b>Category</b>	<b>Parameters</b>	<b>Feature</b>	<b>Parameters</b>	<b>Preprocess</b>	<b>Parameters</b>
IntensityDirect	ThresholdLow=1; ThresholdHigh=100; ErosionDist=0;	GlobalEntropy	NBins=256; RangeMin=0;	Threshold_Image_Mask	ThresholdLow=850;

	OnlyUseMaxSlice=0;		RangeMax=4096; RangeFix=0;		ThresholdHigh=1200; ErosionDist=0;
		GlobalMean		Log_Filter	Size=5; Sigma=1; FillROIOutOn=1; FillROIOutValue=5000;
		GlobalStd			
		GlobalUniformity	NBins=256; RangeMin=0; RangeMax=4096; RangeFix=0;		
		Kurtosis			
		Skewness			
<b>FeatureItem-7</b>					
<b>Category</b>	<b>Parameters</b>	<b>Feature</b>	<b>Parameters</b>	<b>Preprocess</b>	<b>Parameters</b>
IntensityDirect	ThresholdLow=1; ThresholdHigh=100; ErosionDist=0; OnlyUseMaxSlice=0;	GlobalEntropy	NBins=256; RangeMin=0; RangeMax=4096; RangeFix=0;	Threshold_Image_Mask	ThresholdLow=850; ThresholdHigh=1200; ErosionDist=0;
		GlobalMean		Log_Filter	Size=7; Sigma=1.5; FillROIOutOn=1; FillROIOutValue=5000;
		GlobalStd			
		GlobalUniformity	NBins=256; RangeMin=0; RangeMax=4096; RangeFix=0;		
		Kurtosis			
		Skewness			
<b>FeatureItem-8</b>					

Category	Parameters	Feature	Parameters	Preprocess	Parameters
IntensityDirect	ThresholdLow=1; ThresholdHigh=10 0; ErosionDist=0; OnlyUseMaxSlice=0;	GlobalEntropy	NBins=256; RangeMin=0; RangeMax=4096; RangeFix=0;	Threshold_Image_ Mask	ThresholdLow=850 ; ThresholdHigh=12 00; ErosionDist=0;
		GlobalMean		Log_Filter	Size=11; Sigma=2.5; FillROIOutOn=1; FillROIOutValue=5 000;
		GlobalStd			
		GlobalUniformity	NBins=256; RangeMin=0; RangeMax=4096; RangeFix=0;		
		Kurtosis			
		Skewness			
<b>FeatureItem-9</b>					
Category	Parameters	Feature	Parameters	Preprocess	Parameters
Shape		NumberOfVoxel	EdgeVoxelFraction =0.5;	Threshold_Image_ Mask	ThresholdLow=850 ; ThresholdHigh=12 00; ErosionDist=0;
		Volume	EdgeVoxelFraction =0.5;		



## Bibliography

1. SEER stat fact sheets: Lung and bronchus cancer. (2014). at <http://seer.cancer.gov/statfacts/html/lungb.html>
2. Molina, J. R., Yang, P., Cassivi, S. D., Schild, S. E. & Adjei, A. A. Non-small cell lung cancer: Epidemiology, risk factors, treatment, and survivorship. *Mayo Clin. Proc.* **83**, 584–94, (2008).
3. Spiro, S. G., Tanner, N. T., Silvestri, G. a, Janes, S. M., Lim, E., Vansteenkiste, J. F. & Pirker, R. Lung cancer: progress in diagnosis, staging and therapy. *Respirology*. **15**, 44–50, (2010).
4. Chen, H.-Y., Yu, S.-L., Chen, C.-H., Chang, G.-C., Chen, C.-Y., Yuan, A., Cheng, C.-L., Wang, C.-H., Terng, H.-J., Kao, S.-F., Chan, W.-K., Li, H.-N., Liu, C.-C., Singh, S., Chen, W. J., Chen, J. J. W. & Yang, P.-C. A Five-Gene Signature and Clinical Outcome in Non-Small-Cell Lung Cancer. *N. Engl. J. Med.* **356**, 11–20, (2007).
5. Dehing-Oberije, C., Yu, S., De Ruyscher, D., Meersschout, S., Van Beek, K., Lievens, Y., Van Meerbeeck, J., De Neve, W., Rao, B., van der Weide, H. & Lambin, P. Development and external validation of prognostic model for 2-year survival of non-small-cell lung cancer patients treated with chemoradiotherapy. *Int. J. Radiat. Oncol. Biol. Phys.* **74**, 355–62, (2009).
6. Dehing-Oberije, C., Aerts, H., Yu, S., De Ruyscher, D., Menheere, P., Hilvo, M., Van Der Weide, H., Rao, B. & Lambin, P. Development and validation of a prognostic model using blood biomarker information for prediction of survival of non-small-cell lung cancer patients treated with combined chemotherapy and radiation or radiotherapy alone (NCT00181519, NCT00573040, and NCT0. *Int. J. Radiat. Oncol. Biol. Phys.* **81**, 360–

- 368, (2011).
7. Xu, Z., Gao, Y., Hao, Y., Li, E., Wang, Y., Zhang, J., Wang, W., Gao, Z. & Wang, Q. Application of a microfluidic chip-based 3D co-culture to test drug sensitivity for individualized treatment of lung cancer. *Biomaterials*. **34**, 4109–4117, (2013).
  8. Beer, D. G., Kardia, S. L. R., Huang, C.-C., Giordano, T. J., Levin, A. M., Misek, D. E., Lin, L., Chen, G., Gharib, T. G., Thomas, D. G., Lizyness, M. L., Kuick, R., Hayasaka, S., Taylor, J. M. G., Iannettoni, M. D., Orringer, M. B. & Hanash, S. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.* **8**, 816, (2002).
  9. Yu, S. L., Chen, H. Y., Chang, G. C., Chen, C. Y., Chen, H. W., Singh, S., Cheng, C. L., Yu, C. J., Lee, Y. C., Chen, H. S., Su, T. J., Chiang, C. C., Li, H. N., Hong, Q. S., Su, H. Y., Chen, C. C., Chen, W. J., Liu, C. C., Chan, W. K., Chen, W. J., Li, K. C., Chen, J. J. W. & Yang, P. C. MicroRNA Signature Predicts Survival and Relapse in Lung Cancer. *Cancer Cell*. **13**, 48–57, (2008).
  10. Kumar, V., Gu, Y., Basu, S., Berglund, A., Eschrich, S. a, Schabath, M. B., Forster, K., Aerts, H. J. W. L., Dekker, A., Fenstermacher, D., Goldgof, D. B., Hall, L. O., Lambin, P., Balagurunathan, Y., Gatenby, R. a & Gillies, R. J. Radiomics: the process and the challenges. *Magn. Reson. Imaging*. **30**, 1234–48, (2012).
  11. Haralick, R. M. Statistical and structural approaches to texture. *Proc. IEEE*. **67**, 786–804, (1979).
  12. Haralick, R. M., Shanmugam, K. & Dinstein, I. Textural features for image classification. *IEEE Trans. Syst. Man. Cybern.* **3**, 610–621, (1973).
  13. Galloway, M. M. Texture analysis using gray level run lengths. *Comput. Graph. Image Process.* **4**, 172–179, (1975).

14. Amadasun, M. & King, R. Textural features corresponding to textural properties. *IEEE Trans. Syst. Man. Cybern.* **19**, 1264–1274, (1989).
15. Yates, L. R., Gerstung, M., Knappskog, S., Desmedt, C., Gundem, G., Van Loo, P., Aas, T., Alexandrov, L. B., Larsimont, D., Davies, H., Li, Y., Ju, Y. S., Ramakrishna, M., Haugland, H. K., Lilleng, P. K., Nik-Zainal, S., McLaren, S., Butler, A., Martin, S., Glodzik, D., Menzies, A., Raine, K., Hinton, J., Jones, D., Mudie, L. J., Jiang, B., Vincent, D., Greene-Colozzi, A., Adnet, P. Y., Fatima, A., Maetens, M., Ignatiadis, M., Stratton, M. R., Sotiriou, C., Richardson, A. L., Lønning, P. E., Wedge, D. C. & Campbell, P. J. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. **21**, 751–759, (2015).
16. De Bruin, E. C., McGranahan, N., Mitter, R., Salm, M., Wedge, D. C., Yates, L., Jamal-Hanjani, M., Shafi, S., Murugaesu, N., Rowan, A. J., Grönroos, E., Muhammad, M. A., Horswell, S., Gerlinger, M., Varela, I., Jones, D., Marshall, J., Voet, T., Van Loo, P., Rasmussen, D. M., Rintoul, R. C., Janes, S. M., Lee, S. M., Forster, M., Ahmad, T., Lawrence, D., Falzon, M., Capitanio, A., Harkins, T. T., Lee, C. C., Tom, W., Teefe, E., Chen, S. C., Begum, S., Rabinowitz, A., Phillimore, B., Spencer-Dene, B., Stamp, G., Szallasi, Z., Matthews, N., Stewart, A., Campbell, P. & Swanton, C. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. **346**, 251–256, (2014).
17. Schwarz, R. F., Ng, C. K. Y., Cooke, S. L., Newman, S., Temple, J., Piskorz, A. M., Gale, D., Sayal, K., Murtaza, M., Baldwin, P. J., Rosenfeld, N., Earl, H. M., Sala, E., Jimenez-Linan, M., Parkinson, C. A., Markowitz, F. & Brenton, J. D. Spatial and Temporal Heterogeneity in High-Grade Serous Ovarian Cancer: A Phylogenetic Analysis. *PLOS Med.* **12**, e1001789, (2015).
18. Sheth, S., Jilani, D., Bos, A., Ahmed, O., Patel, M. & Zangan, S. Core Lung Biopsy for

- Biomarker Analysis. *J. Thorac. Imaging.* **30**, 314–318, (2015).
19. Basu, S., Hall, L. O., Goldgof, D. B., Gu, Y., Kumar, V., Choi, J., Gillies, R. J. & Gatenby, R. A. Developing a classifier model for lung tumors in CT-scan images. in *2011 IEEE Int. Conf. Syst. Man, Cybern.* 1306–1312, (2011). doi:10.1109/ICSMC.2011.6083840
  20. Song, J., Liu, Z., Zhong, W., Huang, Y., Ma, Z., Dong, D., Liang, C. & Tian, J. Non-small cell lung cancer: quantitative phenotypic analysis of CT images as a potential marker of prognosis. *Sci. Rep.* **6**, 38282, (2016).
  21. Bayanati, H., Thornhill, R. E., Souza, C. A., Sethi-Virmani, V., Gupta, A., Maziak, D., Amjadi, K. & Dennie, C. Quantitative CT texture and shape analysis: Can it differentiate benign and malignant mediastinal lymph nodes in patients with primary lung cancer? *Eur. Radiol.* **25**, 480–7, (2015).
  22. Wang, H., Guo, X. H., Jia, Z. W., Li, H. K., Liang, Z. G., Li, K. C. & He, Q. Multilevel binomial logistic prediction model for malignant pulmonary nodules based on texture features of CT image. *Eur. J. Radiol.* **74**, 124–9, (2010).
  23. Aerts, H. J. W. L., Velazquez, E. R., Leijenaar, R. T. H., Parmar, C., Grossmann, P., Cavalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., Rietveld, D., Hoebers, F., Rietbergen, M. M., Leemans, C. R., Dekker, A., Quackenbush, J., Gillies, R. J. & Lambin, P. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **5**, 1–8, (2014).
  24. Win, T., Miles, K. A., Janes, S. M., Ganeshan, B., Shastry, M., Endozo, R., Meagher, M., Shortman, R. I., Wan, S., Kayani, I., Ell, P. J. & Groves, A. M. Tumor heterogeneity and permeability as measured on the CT component of PET/CT predict survival in patients with non-small cell lung cancer. *Clin. Cancer Res.* **19**, 3591–3599, (2013).

25. Ahn, S. Y., Park, C. M., Park, S. J., Kim, H. J., Song, C., Lee, S. M., Mcadams, H. P. & Goo, J. M. Prognostic Value of Computed Tomography Texture Features in Non – Small Cell Lung Cancers Treated With Definitive Concomitant Chemoradiotherapy. *Invest. Radiol.* **50**, 719–725, (2015).
26. Fried, D. V, Tucker, S. L., Zhou, S., Liao, Z., Mawlawi, O., Ibbott, G. & Court, L. E. Prognostic value and reproducibility of pretreatment CT texture features in stage III non-small cell lung cancer. *Int. J. Radiat. Oncol. Biol. Phys.* **90**, 834–842, (2014).
27. Ganeshan, B., Panayiotou, E., Burnand, K., Dizdarevic, S. & Miles, K. Tumour heterogeneity in non-small cell lung carcinoma assessed by CT texture analysis: A potential marker of survival. *Eur. Radiol.* **22**, 796–802, (2012).
28. Coroller, T. P., Grossmann, P., Hou, Y., Rios Velazquez, E., Leijenaar, R. T. H., Hermann, G., Lambin, P., Haibe-Kains, B., Mak, R. H. & Aerts, H. J. W. L. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiother. Oncol.* **114**, 345–350, (2015).
29. Depeursinge, A., Yanagawa, M., Leung, A. N. & Rubin, D. L. Predicting adenocarcinoma recurrence using computational texture models of nodule components in lung CT. *Med. Phys.* **42**, 2054–2063, (2015).
30. Koo, T., Moon, S., Lim, Y., Kim, J., Kim, Y., Kim, T., Cho, K., Han, J.-Y., Lee, Y., Yun, T., Kim, H. & Lee, J. The effect of tumor volume and its change on survival in stage III non-small cell lung cancer treated with definitive concurrent chemoradiotherapy. *Radiat. Oncol.* **9**, 283, (2014).
31. Mattonen, S. a, Palma, D. a, Haasbeek, C. J. a, Senan, S. & Ward, A. D. Early prediction of tumor recurrence based on CT texture changes after stereotactic ablative radiotherapy (SABR) for lung cancer. *Med. Phys.* **41**, 33502, (2014).

32. Weiss, G. J., Ganeshan, B., Miles, K. A., Campbell, D. H., Cheung, P. Y., Frank, S. & Korn, R. L. Noninvasive image texture analysis differentiates K-ras mutation from pan-wildtype NSCLC and is prognostic. *PLoS One*. **9**, e100244, (2014).
33. Ozkan, E., West, A., Dedelow, J. A., Chu, B. F., Zhao, W., Yildiz, V. O., Otterson, G. A., Shilo, K., Ghosh, S., King, M., White, R. D. & Erdal, B. S. CT gray-level texture analysis as a quantitative imaging biomarker of epidermal growth factor receptor mutation status in adenocarcinoma of the lung. *Am. J. Roentgenol*. **205**, 1016–1025, (2015).
34. Miles, K. A. How to use CT texture analysis for prognostication of non-small cell lung cancer. *Cancer Imaging*. **16**, 10, (2016).
35. Jaffe, C. C. Measures of response: RECIST, WHO, and new alternatives. *J. Clin. Oncol*. **24**, 3245–3251, (2006).
36. Nishino, M., Jackman, D. M., Hatabu, H., Yeap, B. Y., Cioffredi, L.-A., Yap, J. T., Jänne, P. A., Johnson, B. E. & Van den Abbeele, A. D. New response evaluation criteria in solid tumors (RECIST) guidelines for advanced non–small cell lung cancer: Comparison with original RECIST and impact on assessment of tumor response to targeted therapy. *Am. J. Roentgenol*. **195**, W221–W228, (2010).
37. Cunliffe, A., Armato, S. G., Castillo, R., Pham, N., Guerrero, T. & Al-Hallaq, H. A. Lung texture in serial thoracic computed tomography scans: Correlation of radiomics-based features with radiation therapy dose and radiation pneumonitis development. *Int. J. Radiat. Oncol. Biol. Phys*. **91**, 1048–1056, (2015).
38. Rao, S. X., Lambregts, D. M., Schnerr, R. S., Beckers, R. C., Maas, M., Albarello, F., Riedl, R. G., Dejong, C. H., Martens, M. H., Heijnen, L. A., Backes, W. H., Beets, G. L., Zeng, M.-S. & Beets-Tan, R. G. CT texture analysis in colorectal liver metastases: A better way than size and volume measurements to assess response to chemotherapy?

*United Eur. Gastroenterol. J.* **4**, 257–63, (2016).

39. Tian, F., Hayano, K., Kambadakone, A. R. & Sahani, D. V. Response assessment to neoadjuvant therapy in soft tissue sarcomas: using CT texture analysis in comparison to tumor size, density, and perfusion. *Abdom. Imaging.* (2014). doi:10.1007/s00261-014-0318-3
40. Mackin, D., Fave, X., Zhang, L., Fried, D., Yang, J., Taylor, B., Rodriguez-rivera, E., Dodge, C., Jones, A. K. & Court, L. Measuring computed tomography scanner variability of radiomics features. *Invest. Radiol.* **50**, 757–765, (2015).
41. 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. *Comput. Med. Imaging Graph.* **44**, 4–11, (2015).
42. Balagurunathan, Y., Gu, Y., Wang, H., Kumar, V., Grove, O., Hawkins, S., Kim, J., Goldgof, D. B., Hall, L. O., Gatenby, R. a. & Gillies, R. J. Reproducibility and Prognosis of Quantitative Features Extracted from CT Images. *Transl. Oncol.* **7**, 72–87, (2014).
43. Hunter, L. A., Krafft, S., Stingo, F., Choi, H., Martel, M. K., Kry, S. F. & Court, L. E. High quality machine-robust image features: identification in nonsmall cell lung cancer computed tomography images. *Med. Phys.* **40**, 121916.1-12, (2013).
44. Shafiq-ul-Hassan, M., Zhang, G. G., Latifi, K., Ullah, G., Hunt, D. C., Balagurunathan, Y., Abdalah, M. A., Schabath, M. B., Goldgof, D. G., Mackin, D., Court, L. E., Gillies, R. J. & Moros, E. G. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. *Med. Phys.* **44**, 1050–1062, (2017).
45. The University of Texas MD Anderson Cancer Center. Image-guided adaptive conformal photon versus proton therapy. at

<<https://clinicaltrials.gov/ct2/show/record/NCT00915005>>

46. Seppenwoolde, Y., Shirato, H., Kitamura, K., Shimizu, S., van Herk, M., Lebesque, J. V & Miyasaka, K. Precise and real-time measurement of 3D tumor motion in lung due to breathing and heartbeat, measured during radiotherapy. *Int. J. Radiat. Oncol.* **53**, 822–834, (2002).
47. Wang, H., Dong, L., Lii, M. F., Lee, A. L., de Crevoisier, R., Mohan, R., Cox, J. D., Kuban, D. A. & Cheung, R. Implementation and validation of a three-dimensional deformable registration algorithm for targeted prostate cancer radiotherapy. *Int. J. Radiat. Oncol. Biol. Phys.* **61**, 725–35, (2005).
48. Chao, K. S. C., Bhide, S., Chen, H., Asper, J., Bush, S., Franklin, G., Kavadi, V., Liengswangwong, V., Gordon, W., Raben, A., Strasser, J., Koprowski, C., Frank, S., Chronowski, G., Ahamad, A., Malyapa, R., Zhang, L. & Dong, L. Reduce in variation and improve efficiency of target volume delineation by a computer-assisted system using a deformable image registration approach. *Int. J. Radiat. Oncol. Biol. Phys.* **68**, 1512–21, (2007).
49. Liu, H. H., Balter, P., Tutt, T., Choi, B., Zhang, J., Wang, C., Chi, M., Luo, D., Pan, T., Hunjan, S., Starkschall, G., Rosen, I., Prado, K., Liao, Z., Chang, J., Komaki, R., Cox, J. D., Mohan, R. & Dong, L. Assessing respiration-induced tumor motion and internal target volume using four-dimensional computed tomography for radiotherapy of lung cancer. *Int. J. Radiat. Oncol. Biol. Phys.* **68**, 531–40, (2007).
50. Zhang, L., Fried, D., Fave, X., Hunter, L., Yang, J. & Court, L. E. IBEX: An open infrastructure software platform to facilitate collaborative work in radiomics. *Med. Phys.* **42**, 1341–1353, (2015).
51. Ganeshan, B., Abaleke, S., Young, R. C. D., Chatwin, C. R. & Miles, K. A. Texture

- analysis of non-small cell lung cancer on unenhanced computed tomography: Initial evidence for a relationship with tumour glucose metabolism and stage. *Cancer Imaging*. **10**, 137–143, (2010).
52. Ganeshan, B., Miles, K. A., Young, R. C. D. & Chatwin, C. R. Texture analysis in non-contrast enhanced CT: Impact of malignancy on texture in apparently disease-free areas of the liver. *Eur. J. Radiol.* **70**, 101–110, (2009).
  53. Fried, D. Investigation of Quatitative Image Features From Pretreatment Ct and Fdg-Pet Scans in Stage Iii Nsclc Patients Undergoing Defintive Radiation Therapy. *UT GSBS Diss. Theses (Open Access)*. (2015).
  54. Krafft, S. P. Utilizing Computed Tomography Image Features to Advance Prediction of Radiation Pneumonitis. (2016).
  55. Basu, S. Developing Predictive Models for Lung Tumor Analysis. (2012).
  56. RCoreTeam. R: A language and enviroment for statistical computing. (2015). at <<https://www.r-project.org/>>
  57. Therneau, T. A package for survival analysis in S. (2015). at <<http://cran.r-project.org/package=survival>>
  58. Schroeder, M. S., Culhane, A. C., Quackenbush, J. & Haibe-Kains, B. survcomp: an R/Bioconductor package for performance assessment and comparison of survival models. *Bioinformatics*. **27**, 3206–3208, (2011).
  59. Bates, D., Machler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48, (2015).
  60. Wickham, H. ggplot2: Elegant graphics for data analysis. (2009). at <<http://ggplot2.org>>

61. Neuwirth, E. RColorBrewer: ColorBrewer Palettes. (2014).
62. Sachs, M. C. ggkm: Survival plots for ggplot2. (2016).
63. Auguie, B. gridExtra: Miscellaneous Functions for 'Grid' Graphics. (2016).
64. Galavis, P. E., 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 Oncol.* **49**, 1012–6, (2010).
65. Hunter, L. A., Krafft, S., Stingo, F., Choi, H., Martel, M. K., Kry, S. F. & Court, L. E. High quality machine-robust image features: identification in nonsmall cell lung cancer computed tomography images. *Med. Phys.* **40**, 121916, (2013).
66. Leijenaar, R. T. H., Nalbantov, G., Carvalho, S., van Elmpt, W. J. C., Troost, E. G. C., Boellaard, R., Aerts, H. J. W. L., Gillies, R. J. & Lambin, P. The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis. *Sci. Rep.* **5**, 11075, (2015).
67. Leijenaar, R. T. H., Carvalho, S., Velazquez, E. R., van Elmpt, W. J. C., Parmar, C., Hoekstra, O. S., Hoekstra, C. J., Boellaard, R., Dekker, A. L. A. J., Gillies, R. J., Aerts, H. J. W. L. & Lambin, P. Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter-observer variability. *Acta Oncol.* **52**, 1391–7, (2013).
68. Chalkidou, A., O'Doherty, M. J. & Marsden, P. K. False Discovery Rates in PET and CT Studies with Texture Features: A Systematic Review. *PLoS One.* **10**, e0124165, (2015).
69. Hatt, M., Majdoub, M., Vallieres, M., Tixier, F., Cheze Le Rest, C., Groheux, D., Hinde, E., Martineau, A., Pradier, O., Hustinx, R., Perdrisot, R., Guillemin, R., El Naqa, I. & Visvikis, D. FDG PET uptake characterization through texture analysis: investigating the complementary nature of heterogeneity and functional tumor volume in a multi-cancer

- site patient cohort. *J. Nucl. Med.* **56**, 38–44, (2014).
70. Brooks, F. J. & Grigsby, P. W. The effect of small tumor volumes on studies of intratumoral heterogeneity of tracer uptake. *J. Nucl. Med.* **55**, 37–42, (2013).
  71. Yang, J., Zhang, L., Fave, X. J., Fried, D. V., Stingo, F. C., Ng, C. S. & Court, L. E. Uncertainty analysis of quantitative imaging features extracted from contrast-enhanced CT in lung tumors. *Comput. Med. Imaging Graph.* **48**, 1–8, (2016).
  72. Fave, X., Mackin, D., Yang, J., Zhang, J., Fried, D., Balter, P., Followill, D., Gomez, D., Kyle Jones, A., Stingo, F., Fontenot, J. & Court, L. Can radiomics features be reproducibly measured from CBCT images for patients with non-small cell lung cancer? *Med. Phys.* **42**, 6784–6797, (2015).
  73. Segal, E., Sirlin, C. B., Ooi, C., Adler, A. S., Gollub, J., Chen, X., Chan, B. K., Matcuk, G. R., Barry, C. T., Chang, H. Y. & Kuo, M. D. Decoding global gene expression programs in liver cancer by noninvasive imaging. *Nat. Biotechnol.* **25**, 675–80, (2007).
  74. Pal, C., Chakrabarti, A., Ghosh, R. & Choudhury, A. K. Procedia Computer Science A Brief Survey of Recent Edge-Preserving Smoothing Algorithms on Digital Images. *Procedia Comput. Sci.* **0**, 1–40, (2015).
  75. Zhang, L., Fried, D., Fave, X., Mackin, D., Yang, J. & Court, L. SU-E-J-261: The Importance of Appropriate Image Preprocessing to Augment the Information of Radiomics Image Features. *Med. Phys.* **42**, 3326–3327, (2015).
  76. Mackin, D., Court, L., Ng, C., Yang, J., Zhang, L. & Fave, X. SU-F-R-09: Homogenization of CT Images for Radiomics Studies: It's Like Butter(worth). *Med. Phys.* **43**, 3374–3375, (2016).
  77. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple

Testing . at <[http://www.jstor.org/stable/2346101?seq=1#page\\_scan\\_tab\\_contents](http://www.jstor.org/stable/2346101?seq=1#page_scan_tab_contents)>

78. Dafni, U. Landmark analysis at the 25-year landmark point. *Circ. Cardiovasc. Qual. Outcomes*. **4**, 363–371, (2011).
79. Anderson, J., Cain, K. & Gelber, R. Analysis of survival by tumor response. *J. Clin. Oncol.* **1**, 710–719, (1983).
80. Fave, X., Zhang, L., Yang, J., Mackin, D., Balter, P., Gomez, D., Followill, D., Jones, A. K., Stingo, F. & Court, L. E. Impact of image preprocessing on the volume dependence and prognostic potential of radiomics features in non-small cell lung cancer. *Transl. Cancer Res.* **5**, 349–363, (2016).
81. Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L. & Rosati, R. A. Evaluating the yield of medical tests. *JAMA J. Am. Med. Assoc.* **247**, 2543–2546, (1982).
82. Venables, W. N. & Ripley, B. D. *Modern applied statistics with S*. (Springer, 2002).
83. Fushiki, T. Estimation of prediction error by using K-fold cross-validation. *Stat. Comput.* **21**, 137–146, (2011).
84. Simon, R. M., Subramanian, J., Li, M.-C. & Menezes, S. Using cross-validation to evaluate predictive accuracy of survival risk classifiers based on high-dimensional data. *Brief. Bioinform.* **12**, 203–214, (2011).
85. Hilsenbeck, S. G., Clark, G. M. & McGuire, W. L. Why do so many prognostic factors fail to pan out? *Breast Cancer Res. Treat.* **22**, 197–206, (1992).
86. Hilsenbeck, S. G. & Clark, G. M. Practical p-value adjustment for optimally selected cutpoints. *Stat. Med.* **15**, 103–112, (1996).
87. Carvalho, S., Leijenaar, R. T. H., Troost, E. G. C., van Elmpt, W., Muratet, J.-P., Denis,

- F., De Ruyscher, D., Aerts, H. J. W. L. & Lambin, P. Early variation of FDG-PET radiomics features in NSCLC is related to overall survival - the 'delta radiomics' concept. in *Radiother. Oncol.* **118**, S20–S21, (2016).
88. McBride, G. B. A proposal for strength-of-agreement criteria for Lin's Concordance Correlation Coefficient. *NIWA Client Rep. HAM2005-062*. 1–14, (2005).
  89. Zou, K. H., Tuncali, K. & Silverman, S. G. Correlation and simple linear regression. *Radiology*. **227**, 617–22, (2003).
  90. Mukaka, M. M. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Med. J.* **24**, 69–71, (2012).
  91. Miles, K. a, Ganeshan, B. & Hayball, M. P. CT texture analysis using the filtration-histogram method: What do the measurements mean? *Cancer Imaging*. **13**, 400–6, (2013).
  92. Lujan, A. E., Larsen, E. W., Balter, J. M. & Ten Haken, R. K. A method for incorporating organ motion due to breathing into 3D dose calculations. *Med. Phys.* **26**, 715–720, (1999).
  93. Jones, A. K. & Mahvash, A. Evaluation of the potential utility of flat panel CT for quantifying relative contrast enhancement. *Med. Phys.* **39**, 4149–54, (2012).
  94. Paganetti, H., Niemierko, A., Ancukiewicz, M., Gerweck, L. E., Goitein, M., Loeffler, J. S. & Suit, H. D. Relative biological effectiveness (RBE) values for proton beam therapy. *Int. J. Radiat. Oncol.* **53**, 407–421, (2002).
  95. Gerweck, L. E. & Kozin, S. V. Relative biological effectiveness of proton beams in clinical therapy. *Radiother. Oncol.* **50**, 135–142, (1999).
  96. Usmanij, E. A., de Geus-Oei, L.-F., Troost, E. G. C., Peters-Bax, L., van der Heijden, E.

- H. F. M., Kaanders, J. H. A. M., Oyen, W. J. G., Schuurbiers, O. C. J. & Bussink, J. 18F-FDG PET early response evaluation of locally advanced non-small cell lung cancer treated with concomitant chemoradiotherapy. *J. Nucl. Med.* **54**, 1528–34, (2013).
97. Nahmias, C., Hanna, W. T., Wahl, L. M., Long, M. J., Hubner, K. F. & Townsend, D. W. Time course of early response to chemotherapy in non-small cell lung cancer patients with 18F-FDG PET/CT. *J. Nucl. Med.* **48**, 744–51, (2007).
  98. Yabuuchi, H., Hatakenaka, M., Takayama, K., Matsuo, Y., Sunami, S., Kamitani, T., Jinnouchi, M., Sakai, S., Nakanishi, Y. & Honda, H. Non–Small Cell Lung Cancer: Detection of Early Response to Chemotherapy by Using Contrast-enhanced Dynamic and Diffusion-weighted MR Imaging. *Radiology*. **261**, 598–604, (2011).
  99. Lee, Y. H., Kim, Y. S., Lee, H. C., Lee, S. W., Kang, Y. N., Kang, J. H., Hong, S. H., Kim, Y. K., Kim, S. J., Ahn, M. I., Han, D. H., Yoo, I. R., Park, J. G., Sung, S. W. & Lee, K. Y. Tumour volume changes assessed with high-quality KVCT in lung cancer patients undergoing concurrent chemoradiotherapy. *Br. J. Radiol.* **88**, 20150156, (2015).
  100. Lind, J. S. W., Meijerink, M. R., Dingemans, A.-M. C., van Kuijk, C., Öllers, M. C., de Ruysscher, D., Postmus, P. E. & Smit, E. F. Dynamic contrast-enhanced CT in patients treated with sorafenib and erlotinib for non-small cell lung cancer: a new method of monitoring treatment? *Eur. Radiol.* **20**, 2890–2898, (2010).
  101. Eisenhauer, E. A., Therasse, P., Bogaerts, J., Schwartz, L. H., Sargent, D., Ford, R., Dancey, J., Arbuck, S., Gwyther, S., Mooney, M., Rubinstein, L., Shankar, L., Dodd, L., Kaplan, R., Lacombe, D. & Verweij, J. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *Eur. J. Cancer*. **45**, 228–47, (2009).
  102. Sonke, J.-J. & Belderbos, J. Adaptive Radiotherapy for Lung Cancer. *Semin. Radiat. Oncol.* **20**, 94–106, (2010).

103. Kupelian, P. A., Ramsey, C., Meeks, S. L., Willoughby, T. R., Forbes, A., Wagner, T. H. & Langen, K. M. Serial megavoltage CT imaging during external beam radiotherapy for non–small-cell lung cancer: Observations on tumor regression during treatment. *Int. J. Radiat. Oncol.* **63**, 1024–1028, (2005).
104. Balagurunathan, Y., Kumar, V., Gu, Y., Kim, J., Wang, H., Liu, Y., Goldgof, D. B., Hall, L. O., Korn, R., Zhao, B., Schwartz, L. H., Basu, S., Eschrich, S., Gatenby, R. A. & Gillies, R. J. Test–Retest Reproducibility Analysis of Lung CT Image Features. *J. Digit. Imaging.* **27**, 805–823, (2014).
105. Leijenaar, R. T. H., Carvalho, S., Hoebbers, F. J. P., Aerts, H. J. W. L., van Elmpt, W. J. C., Huang, S. H., Chan, B., Waldron, J. N., O’Sullivan, B., Lambin, P., O’sullivan, B., Lambin, P., O’Sullivan, B. & Lambin, P. External validation of a prognostic CT-based radiomic signature in oropharyngeal squamous cell carcinoma. *Acta Oncol. (Madr)*. **54**, 1423–1429, (2015).
106. Grove, O., Berglund, A. E., Schabath, M. B., Aerts, H. J. W. L., Dekker, A., Wang, H., Velazquez, E. R., Lambin, P., Gu, Y., Balagurunathan, Y., Eikman, E., Gatenby, R. A., Eschrich, S. & Gillies, R. J. Quantitative computed tomographic descriptors associate tumor shape complexity and intratumor heterogeneity with prognosis in lung adenocarcinoma. *PLoS One*. **10**, e0118261, (2015).
107. Parmar, C., Rios Velazquez, E., Leijenaar, R., Jermoumi, M., Carvalho, S., Mak, R. H., Mitra, S., Shankar, B. U., Kikinis, R., Haibe-Kains, B., Lambin, P. & Aerts, H. J. W. L. Robust radiomics feature quantification using semiautomatic volumetric segmentation. *PLoS One*. **9**, e102107, (2014).
108. Mclean, E., Goh, V. & Cook, G. J. Imaging Heterogeneity in Lung Cancer: Techniques, Applications, and Challenges. 1–11, (2016). doi:10.2214/AJR.15.15864

109. Depeursinge, A., Foncubierta-Rodriguez, A., Van De Ville, D. & Müller, H. Three-dimensional solid texture analysis in biomedical imaging: Review and opportunities. *Med. Image Anal.* **18**, 176–96, (2014).
110. Paul, T., Banerjee, P., Mukherjee, A. & Bandhyopadhyay, S. Technologies in Texture Analysis – A Review. *Br. J. Appl. Sci. Technol.* **13**, 1–21, (2016).
111. Castellano, G., Bonilha, L., Li, L. M. & Cendes, F. Texture analysis of medical images. *Clin. Radiol.* **59**, 1061–9, (2004).
112. Court, L. E., Fave, X., Mackin, D., Lee, J., Yang, J. & Zhang, L. Computational resources for radiomics. **5**, 340–348, (2016).
113. Zwanenburg, A., Leger, S., Vallières, M. & Löck, S. Image biomarker standardisation initiative - feature definitions. *eprint arXiv:1612.07003 [cs.CV]*. (2016).
114. Prasanna, P., Tiwari, P. & Madabhushi, A. Co-occurrence of Local Anisotropic Gradient Orientations (CoLIAGe): A new radiomics descriptor. *Sci. Rep.* **6**, 37241, (2016).
115. Markel, D., Caldwell, C., Alasti, H., Soliman, H., Ung, Y., Lee, J. & Sun, A. Automatic segmentation of lung carcinoma using 3D texture features in 18-FDG PET/CT. *Int. J. Mol. Imaging.* **2013**, 1–13, (2013).
116. Bae, K. T. Intravenous Contrast Medium Administration and Scan Timing at CT: Considerations and Approaches. **256**, (2010).
117. Kim, H., Park, C. M., Park, S. J., Song, Y. S., Lee, J. H., Hwang, E. J. & Goo, J. M. Temporal Changes of Texture Features Extracted From Pulmonary Nodules on Dynamic Contrast-Enhanced Chest Computed Tomography: How Influential Is the Scan Delay? *Invest. Radiol.* **51**, 569–574, (2016).

## Vita

Xenia Janice Favè was born in Basle, Switzerland on April 19, 1990, the daughter of Dominique Janine Favè and Carl Phillipp Favè. After completing her work at Frank W. Cox High School, Virginia Beach, Virginia in 2008, she entered Florida Institute of Technology in Melbourne, Florida. She received the degree of Bachelor of Science with a major in physics from Florida Institute of Technology in December, 2011. In September of 2012 she entered The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences.

Permanent address:

113B 83<sup>rd</sup> Street

Virginia Beach, VA 23451