5-2013

Radiomics Of Nsclc: Quantitative Ct Image Feature Characterization And Tumor Shrinkage Prediction

Luke Hunter

RADIOMICS OF NSCLC: QUANTITATIVE CT IMAGE FEATURE CHARACTERIZATION AND TUMOR SHRINKAGE PREDICTION

by

Luke A. Hunter, B.S.

APPROVED:

_____
Laurence E. Court, Ph.D.
Supervisory Professor

_____
Stephen Kry, Ph.D.

_____
Francesco Stingo, Ph.D.

_____
Mary Martel, Ph.D.

_____
Haesun Choi, M.D.

APPROVED:

_____
Dean, The University of Texas
Graduate School of Biomedical Sciences at Houston

RADIOMICS OF NSCLC: QUANTITATIVE CT IMAGE FEATURE CHARACTERIZATION AND TUMOR SHRINKAGE PREDICTION

A

THESIS

Presented to the Faculty of
The University of Texas
Health Science Center at Houston
and
The University of Texas
M.D. Anderson Cancer Center
Graduate School of Biomedical Sciences
in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

by

Luke A. Hunter, B.S.
Houston, Texas

May, 2013

Acknowledgements

I would like to thank the Hertz Applied Science Fellowship for both their financial support and inspiration. Our outings in Cambridge, San Jose, Santa Cruz, and Woods Hole put me in contact with talented individuals such as Dr. Sebastian Seung whom started the series of events which led to this project.

I also want to thank Eric Solberg, Carol Helton, Dr. Victoria Knutson, and many others in the GSBS staff. Their support, friendliness, and coordination with the Hertz fellowship greatly enhanced my graduate school experience. In the medical physics program, I would also like to thank Dr. Edward Jackson, Georgeanne Moore, and Gloria Mendoza for their important support and advice.

Many thanks also go to my advisor, Dr. Laurence Court, for his enthusiastic support of this project, valuable contributions, and role as a paragon, both academically and personally, whom I admire greatly. I also want to thank my entire committee for their thoughtful discussions, guidance, and valuable time: Dr. Stephen Kry, Dr. Mary Martel, Dr. Francesco Stingo, and Dr. Haesun Choi. Your diverse viewpoints have illuminated numerous facets of this work that would have otherwise gone unnoticed.

I also think I speak for the entire medical physics student body when I thank Dr. George Starkschall for his exemplary introduction to medical physics course. He genuinely cares about effective teaching and, despite seniority and high stature, is always approachable and open to new ideas, teaching methodologies, and constructive feedback.

I want to thank my parents for their unwavering support and selflessness. By your own example you taught me that you can achieve nearly anything if you work hard and put your mind to it. Through countless book readings, trips to the oceans, lakes, and forests, summer camps, odd pets, and inspirational conversations, you instilled in

me a love and fascination of the natural world, a desire to protect it, and a desire to understand it. Thank you.

Finally, above all, I want to thank my wife, Rachel. Thank you for your patience. Thank you for your encouragement. Thank you for your devotion. Through the years, the seas have swelled up and down, the winds have blown to paths unforeseen—and you've always been there. I'm glad we're together for the adventure of life, and I look forward to seeing where the winds will take us next.

RADIOMICS OF NSCLC: QUANTITATIVE CT IMAGE FEATURE CHARACTERIZATION AND TUMOR SHRINKAGE PREDICTION

Publication No._____

Luke Aaron Hunter, B.S.

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

Abstract

Radiomics is the high-throughput extraction and analysis of quantitative image features. For non-small cell lung cancer (NSCLC) patients, radiomics can be applied to standard of care computed tomography (CT) images to improve tumor diagnosis, staging, and response assessment.

The first objective of this work was to show that CT image features extracted from pre-treatment NSCLC tumors could be used to predict tumor shrinkage in response to therapy. This is important since tumor shrinkage is an important cancer treatment endpoint that is correlated with probability of disease progression and overall survival. Accurate prediction of tumor shrinkage could also lead to individually customized treatment plans.

To accomplish this objective, 64 stage NSCLC patients with similar treatments were all imaged using the same CT scanner and protocol. Quantitative image features were extracted and principal component regression with simulated annealing subset selection was used to predict shrinkage. Cross validation and permutation tests were used to validate the results. The optimal model gave a strong correlation between the observed and predicted shrinkages with $r = 0.81$.

The second objective of this work was to identify sets of NSCLC CT image features that are reproducible, non-redundant, and informative across multiple

machines. Feature sets with these qualities are needed for NSCLC radiomics models to be robust to machine variation and spurious correlation.

To accomplish this objective, test-retest CT image pairs were obtained from 56 NSCLC patients imaged on three CT machines from two institutions. For each machine, quantitative image features with concordance correlation coefficient values greater than 0.90 were considered reproducible. Multi-machine reproducible feature sets were created by taking the intersection of individual machine reproducible feature sets. Redundant features were removed through hierarchical clustering.

The findings showed that image feature reproducibility and redundancy depended on both the CT machine and the CT image type (average cine 4D-CT imaging vs. end-exhale cine 4D-CT imaging vs. helical inspiratory breath-hold 3D CT). For each image type, a set of cross-machine reproducible, non-redundant, and informative image features was identified. Compared to end-exhale 4D-CT and breath-hold 3D-CT, average 4D-CT derived image features showed superior multi-machine reproducibility and are the best candidates for clinical correlation.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Background and Significance

Lung cancer is the leading cause of cancer-related mortality both worldwide and in the United States (1). Over 200,000 new lung cancer cases are diagnosed and over 160,000 people die due to the disease every year in the United States (2). Non-small cell lung cancer (NSCLC) accounts for approximately 85% of these lung cancer cases, and overall 5-year lung cancer survival is 15% (3). The variation of current incidence rates is primarily explained by variations in historical tobacco use (4). Therefore, as a result of decreased tobacco consumption, incidence rates in the United States are declining. However, worldwide incidence rates are rising rapidly due to increased tobacco use in China and other developing nations (5).

The tumor node metastasis (TNM) staging system is a systematic way of representing the spread of lung cancer by assessing tumor size, nodal involvement, and metastatic extent (6). Staging information is obtained using thoracic computed tomography (CT) and is often augmented by positron emission tomography (PET) which can detect metabolically active microscopic disease (7). The three TNM categories can be combined (i.e. stage grouped) to give an overall tumor stage which has important implications for treatment and prognosis. Stage I lung cancer is present in the lungs only, stage II has nearby lymph node involvement, stage III (locally advanced disease) has more distant lymph node involvement, and stage IV (advanced disease) has additional organ-system involvement. Surgical resection is recommended for stages I and II, and has relatively good clinical outcomes. However, approximately 70% of patients present with stage III or stage IV disease (2). Chemotherapy is generally recommended for stage IV patients, and radiation with concurrent chemotherapy is recommended for stage III treatment.

At its core, the TNM staging system is an exclusively anatomically-oriented system that does not take advantage of additional data which may be relevant to prognosis and treatment, and patients that are assigned the same stage often have large variations in

clinical outcome. The addition of information from various non-anatomical sources could help reduce this problem. For example, the simple physiological features of patient age and body mass index have been shown to improve NSCLC staging (8). Integration of genetic information (e.g. single nucleotide polymorphism genotyping) improves the ability to predict patient survival (9-11), and serum biomarkers have also shown promise for NSCLC recurrence prediction (12). The literature shows that these and other molecular biomarkers can be used for indicating etiology, prognosis, and therapy response (13-15). However, external validation studies for molecular biomarker studies have largely been unsuccessful, and their clinical application seems stunted (16). This could be due to variations in institution demographics, sample collection, and/or analysis techniques. Additionally, many solid tumors are known to be temporally heterogeneous (i.e. genotype changes through time) and/or spatially heterogeneous (i.e. genotype varies spatially within the tumor). Therefore, a biopsy sampling a random spatial sub-region of a tumor at a single time point may not be able to accurately reflect true complexity of the tumor (17).

In light of these complications, CT imaging seems like a viable alternative for probing the information content of tumors. CT imaging is readily available, and any additional information gleaned from imaging is essentially “free” since it is performed as part of the standard of care for lung cancer. Additionally, because of its non-invasive nature, CT imaging can be used to achieve high temporal resolution via frequent imaging (17). CT imaging also has high spatial resolution which can allow for quantification and assessment of intra-tumor spatial heterogeneity (17). This is important since identification of an image heterogeneity – treatment response link could be used to further optimize the therapeutic ratio of radiation treatments (18).

Various CT image features can be useful descriptors. In NSCLC, tumors are often qualitatively described as spiculated or cavitated, and several groups are working to develop a constrained vocabulary for lung tumor CT image annotation (19, 20). Tumors are also commonly described by quantitative one-dimensional (Response Evaluation Criteria in Solid

Tumors) or two-dimensional (World Health Organization) size measurements. However, the nuances of tumor morphology are not well-captured by these two measures, and in some cases their changes do no strongly correlate with therapeutic benefit (21, 22). Other commonly obtained quantitative image features such as the mean voxel intensity are typically simple calculations over the region of interest (ROI) (23). More advanced quantitative measures are usually statistical, model, or transform based and reflect variations in tumor morphology, heterogeneity, and/or texture (24). Statistics-based image features such as the run-length matrix (RLM) and co-occurrence matrix (COM) are based on probabilities of voxels occurring in certain combinations (25, 26). Model-based image features such as the fractal dimension quantify texture irregularity and roughness (24). Transform-based methods such as the wavelet transform are used to quantify textures in the frequency space (24). Compared to simple ROI-averaged image features, these more advanced features tend to have significantly improved prognostic power (27-30).

Inspired by the high-throughput success of the “omics” (genomics, proteomics, metabolomics, etc.), the newly created field of radiomics is centered around the high-throughput extraction of advanced quantitative image features (17). The radiomics hypothesis is that these image features are related to a tumor’s underlying genotype and phenotype. This was first shown to be true in liver cancer where 28 image features were able to reconstruct 78% of the gene expression profile (31). Other studies have shown that genomic heterogeneity, treatment resistance, and metastatic probability are each associated with tumor image heterogeneity (27, 32). In addition, CT texture analysis can indicate tumor invasion and estrogen receptor status in patients with breast cancer, can partition high and low grade cerebral gliomas, and can indicate overall survival rates in colorectal cancer (33-35). In lung cancer, image features have also been used to distinguish between the adenocarcinoma and squamous cell carcinoma subtypes of NSCLC (36). The fractal dimension of lung tumors has been used to classify pulmonary nodules and also correlates with tumor stage (37, 38). NSCLC CT texture analysis by Ganeshan *et al.* has also

discovered image feature associations for tumor stage, metabolism, hypoxia, and angiogenesis (39, 40).

Particularly in NSCLC CT radiomics applications, authors often lament the lack of data standardization and uniformity (41). This is an important point to consider, because NSCLC radiomics models could be easily confounded by patient variability (different treatment types, different disease stages, different demographics, etc.), image variability (motion management, voxel size, contrast vs. no contrast, helical vs. axial, etc.), or by interpretation variability (inter-operator segmentation differences, subtle differences in image feature extraction algorithms, etc.). Each of these sources of variability needs to be carefully minimized. In theory, interpretation variability can be controlled for through automated segmentation and feature extraction standardization. Patient variability, on the other hand, cannot be controlled for on data sets studied retrospectively. This is a challenge because training a model requires as many patient-outcome pairs as possible and this is usually at odds with patient uniformity. Image variability also introduces the same problem: obtaining a larger sample to study introduces more variability.

To address some of these concerns, the Quantitative Imaging Biomarkers Alliance, organized by the Radiological Society of North America in 2007, is working to harmonize imaging standards across institutions and CT machine venders (42). Additionally, in a simple water phantom study, several CT texture features have been shown to be relatively robust to kVp and mAs variation (43). For more complex human data, there is a publically available "test-retest" dataset of 32 NSCLC patients in which each patient was scanned twice on the same CT machine with a short break in between (44). This dataset was utilized by Kumar *et al.* (23) to identify 39 out of 327 image features that were reproducible, non-redundant, and informative. Reproducibility was quantified by the concordance correlation coefficient (45) and is desirable to ensure imaging biomarker model fidelity and generality. Informativeness was quantified using the dynamic range and is important for discerning patients. Non-

redundance was quantified by feature correlation coefficients and is desirable to increase model interpretability and reduce overfitting.

Basu *et al.* (2012) extracted these 39 features from a separate set of 95 NSCLC patients, binarized patient outcomes around 2 year survival, and tested various classifiers to obtain an area under the curve (AUC) of 0.68 (41). However, this study is limited by treatment variability and imaging variability within its patient dataset. Patients received different types of radiation treatments and chemotherapy, and imaging was done on several different machines with considerable parameter variation. For example, slice thickness was highly variable for the images and likely limited the usefulness of RLM and COM features since they assume a static voxel size. Aerts *et al.* (2012) performed a similar survival prediction study on a larger data set (412 patients) and obtained an AUC of 0.70 (46, 47). Besides being limited by treatment variability and imaging variability, these two studies assumed that the reproducible feature set found by Kumar *et al.* (23) would also be reproducible on the machines used to collect their data, but the validity of this assumption is unclear. Therefore, to expand upon previous work in NSCLC CT radiomics, the significance of this project is to 1) predict NSCLC treatment outcome from patients with uniform treatment and imaging and 2) use public and internal NSCLC CT test-retest datasets to identify image features that are robust across multiple CT machines.

1.2 Hypotheses

1. *CT image features extracted from pre-treatment NSCLC tumors can be used to predict tumor shrinkage in response to therapy.*
2. *Using several selection metrics, a small subset of multi-machine reproducible, non-redundant, and informative image features can be identified for each of several different CT image types.*

1.3 Specific Aims

The hypotheses will be tested through the following specific aims:

1. *Develop MATLAB code which can access ROIs from the Pinnacle^3 TPS and extract 3D quantitative image features from an ROI's geometry, intensity histogram, absolute gradient image, run-length matrices, and co-occurrence matrices.*
2. *Predict NSCLC tumor shrinkage for patients treated with IMRT and PSPT using models based on pre-treatment CT image features.*
3. *Develop a user-friendly graphical user interface program to leverage previously developed code by streamlining Pinnacle ROI location, verification, and image feature extraction.*
4. *Characterize the reproducibility, non-redundance, and informativeness of CT image features and identify subsets of image features that are optimal under image acquisition protocols.*

1.4 Thesis Organization

This thesis is intended to serve as a permanent record of the work that was completed to evaluate the hypotheses of the project. Chapters 2 and 3 are self-contained studies, each including an introduction, methods, results, discussion, and conclusions. These chapters each describe separate portions of the work completed for this project.

Chapter 2 addresses specific aims 1 and 2, providing reasonably accurate predictions of NSCLC tumor shrinkage from pre-treatment CT images. Chapter 3 addresses specific aims 3 and 4, and lists relatively small image feature sets that are reproducible, non-redundant, and informative for each of the following image types: cine 4D-CT average images, cine 4D-CT end-exhale images, and inspiratory breath-hold helical 3D-CT images. In closing, Chapter 4 is a summary of the overall research project. This section assesses the hypotheses, draws overall conclusions, and proposes future related research.

Chapter 2: NSCLC tumor shrinkage prediction using quantitative image features

2.1 Introduction

A goal of oncology therapies is to utilize information gained from treating previous patients to deliver treatment specific to the patient and disease. The current tumor node metastasis (TNM) staging system for non-small cell lung cancer (NSCLC) is used for this purpose and utilizes anatomical information such as the tumor size, location, spread, and lymph node involvement (6). Although this system is based on the study of over 67,000 NSCLC cases, NSCLC patients with the same TNM staging often have very different clinical outcomes. To improve this, many models have been proposed which add additional non-anatomical features to the TNM staging system. For example, the readily available features of age and body mass index have been shown to improve NSCLC staging (8). Others have used single nucleotide polymorphism genotyping to predict survival (9-11). NSCLC recurrence prediction from serum biomarkers has also been explored (12). However, these molecular assays are invasive, much less commonly obtained than age and body mass index, and could be sensitive to variations in institution demographics, sample collection, and/or analysis techniques. Moreover, molecular assays are also limited by the fact that many tumors are spatially and/or temporally heterogeneous (48, 49).

CT images, on the other hand, are ubiquitous, non-invasive, and increasingly quantitative and standardized. They can also assess the tumor both spatially and through time (48). In basic quantitative CT imaging, such as the Response Evaluation Criteria in Solid Tumors (RECIST) guidelines, tumor response to therapy is gauged by one-dimensional measurements of tumor size (50, 51). However, using more complex quantifiable image features such as the tumor heterogeneity and radiodensity can result in a significant prognostic improvement over RECIST (22). Other quantitative image features are usually statistical, model, or transform based and show similar prognostic promise (24, 27-30). In light of these findings, the nascent field of radiomics aims to achieve automated high-throughput extraction of various quantitative image features under the hypothesis that they

are related to gene expression and phenotype. Supporting this claim, Segal *et al.* (2007) showed that 28 image features could reconstruct 78% of a liver cancer gene expression profile (31). Additionally, it has been demonstrated that features such as the tumor image heterogeneity are associated with genomic heterogeneity and are correlated with increased treatment resistance and metastatic probability (27, 32).

Recently, radiomics has been applied to NSCLC survival prediction. Kumar *et al.* (2012) extracted 327 3D features for 32 patients who underwent two CT imaging sessions spaced 15 minutes apart on the same machine (52). To assess intra-patient (test-retest) reproducibility, they calculated feature concordance correlation coefficients and dynamic ranges and determined that 39 of these 327 features were highly reproducible. Using this information, Basu *et al.* (2012) extracted these 39 features from a separate set of 95 NSCLC patients, binarized patient outcomes around 2 year survival, and tested various classifiers to obtain an area under the curve (AUC) of 0.68 (41). Aerts *et al.* (2012) used similar techniques on a larger data set (412 patients) to obtain an AUC of 0.70 (46, 47).

There are several limitations to these studies, however. First, both used multiple CT machines to establish their CT data sets. At our institution we have found that quantitative image features vary significantly across different machines, even of the same model. Given the large size of the databases used in these two studies, it is likely that several machines were used and it is possible that their prognostic power was limited by inter-machine variability. Moreover, inter-machine variability could also affect the intra-patient variability. It is not clear that the same set of features identified by Kumar *et al.* (52) would be reproducible on other machines as was assumed by Basu *et al.* (41). Finally, both the work by Basu *et al.* (41) and Aerts *et al.* (46) have inter-patient treatment variability which could similarly limit the prognostic power of their models.

In summary, although increasing patient database size is desirable to provide enough training inputs to create models with sufficient expressiveness, it may be counterproductive if it results in increased image and/or treatment heterogeneity that

confounds modelling. Therefore, the purpose of our study was to develop a quantitative image feature-derived prediction model for NSCLC volume shrinkage (a proxy for survival) on a set of patients with similar treatments all imaged with the same CT scanner and protocol.

2.1 Materials and methods

Data acquisition

We obtained simulation and weekly free-breathing, non-contrast 4DCT images from 66 patients with locally advanced, pathologically proven stage II-IIIB NSCLC. As part of a separate prospective trial comparing treatment modalities, these patients were randomly assigned treatment either by IMRT (36 patients) or protons (30 patients). The protons were delivered with passively scattered proton therapy with an assumed relative biological effectiveness (RBE) of 1.1. Both groups were treated to a 74 Gy (RBE) dose level and had concurrent chemotherapy. All simulation images were acquired with the same GE Medical Systems LightSpeed 16 machine (GE Healthcare, Milwaukee, WI) with helical scans using kVp = 120, mAs = 450, and a standard reconstruction convolution kernel. Axial images were 512 x 512 pixels with voxel dimensions of 0.98 x 0.98 x 2.5 mm^3.

To minimize the effects of respiratory motion, physicians contoured the gross tumor volumes (GTVs) on the end-exhale (T50) phase of planning 4DCT images. These contours were propagated onto weekly images using an in-house demons-based deformation algorithm (53, 54). For each patient, the weekly GTV volume divided by the planning GTV volume was defined as the *tumor response* for a particular week.

Although both treatment regimens lasted longer than six weeks, for logistical reasons some patients did not have 4DCT scans after week six. Therefore, week six was chosen as the “final” time point for prediction of the tumor response from the initial planning images. For modelling, both the IMRT and proton data sets were pooled in order to obtain a larger set of observations, and one patient from each group was thrown out for being a tumor response outlier (deviating by more than two standard deviations). Results from (55) support pooling

the two data sets by showing that each group had remarkably similar tumor responses (see Table 2.1). In fact, the p-value that the two groups' tumor responses have equal means is 0.733.

Table 2.1: Week six tumor responses by treatment modality. The weekly tumor response is defined as the GTV volume for the week divided by the GTV volume on the planning image. Intensity modulated radiation therapy (IMRT) and passively scatted proton therapy (PSPT) patients showed remarkably similar tumor responses and were pooled for the purpose of modelling.

Treatment Group	N	Mean	Median	Std. Dev.	Max	Min
IMRT	36	0.83	0.85	0.19	1.39	0.47
PSPT	30	0.81	0.81	0.19	1.45	0.46
IMRT+PSPT	66	0.82	0.82	0.19	1.45	0.46

Feature extraction

We extracted quantitative image features from the pre-treatment planning GTVs using in-house software that used the following 3D feature sources: geometry, intensity histogram, absolute gradient image, co-occurrence matrix (COM), and run-length matrix (RLM). See Table 2.2 for a complete listing of features extracted from each feature source. For the absolute gradient image, each voxel is defined as the difference between paired adjacent voxels in all three directions added in quadrature. For a given image and direction, the run-length matrix $p(i,j)$ is defined as the number of voxel runs in the image with intensity i and run length j (26). RLM features were extracted for 13 different 3D directions. For a given image and displacement vector $\vec{\Delta}$, a co-occurrence matrix $c(i,j)$ is defined as the probability that two voxels separated by $\vec{\Delta}$ will have the intensity values of i and j, respectively (25). COM features were extracted for six different one-voxel displacements: left, right, superior, inferior, anterior, and posterior.

Table 2.2: List of quantitative image features extracted from each planning CT image GTV. The RLM features were generated for 13 different 3D directions, and the COM features were generated for 6 different 3D displacements.

Geometry	Intensity Histogram	Absolute Gradient	Run-Length Matrix (RLM)	Co-Occurrence Matrix (COM)
Volume (V)	Mean	Mean	Run Length Nonuniformity	Angular 2nd Moment
Area (A)	Median	Variance	Grey Level Nonuniformity	Contrast
V/A Ratio	Variance	Skewness	Long Run Emphasis	Correlation
Pruned Volume	Skewness	Kurtosis	Short Run Emphasis	Sum of Squares
Pruned Area	Kurtosis	% Non-Zero	Fx. of Image in Runs	Inv. Diff. Moment
Pruned V/A Ratio	Minimum			Sum Average
Fx. Volume Pruned	1th percentile			Sum Variance
	10th percentile			Sum Entropy
	90th percentile			Entropy
	99th percentile			Diff. Variance
	Maximum			Diff. Entropy
	Entropy			

Prior to feature extraction, we pre-processed the GTV regions of interest (ROIs) by pruning (removing voxels below a certain Hounsfield unit [HU] cut-off). This was done to reduce contouring variability, remove air cavities, and better define the solid tumor. However, the optimal HU cut-off was not known *a priori*, so each unprocessed ROI gave rise to several variants, each of which had different HU cut-offs. Because the optimal voxel intensity bit depths were also unknown, each of these variants then gave rise to additional variants for which the gradient, RLM, and COM image bit depths were varied. Thus, for each patient's ROI, many additional ROIs with different pruning and bit depths were created. Figure 2.1 illustrates this process. For each unique HU cut-off and bit depth combination, the resulting ROIs were pooled from the 64 patients and features were extracted to create a *feature matrix* with each row corresponding to a patient and each column corresponding to a feature.

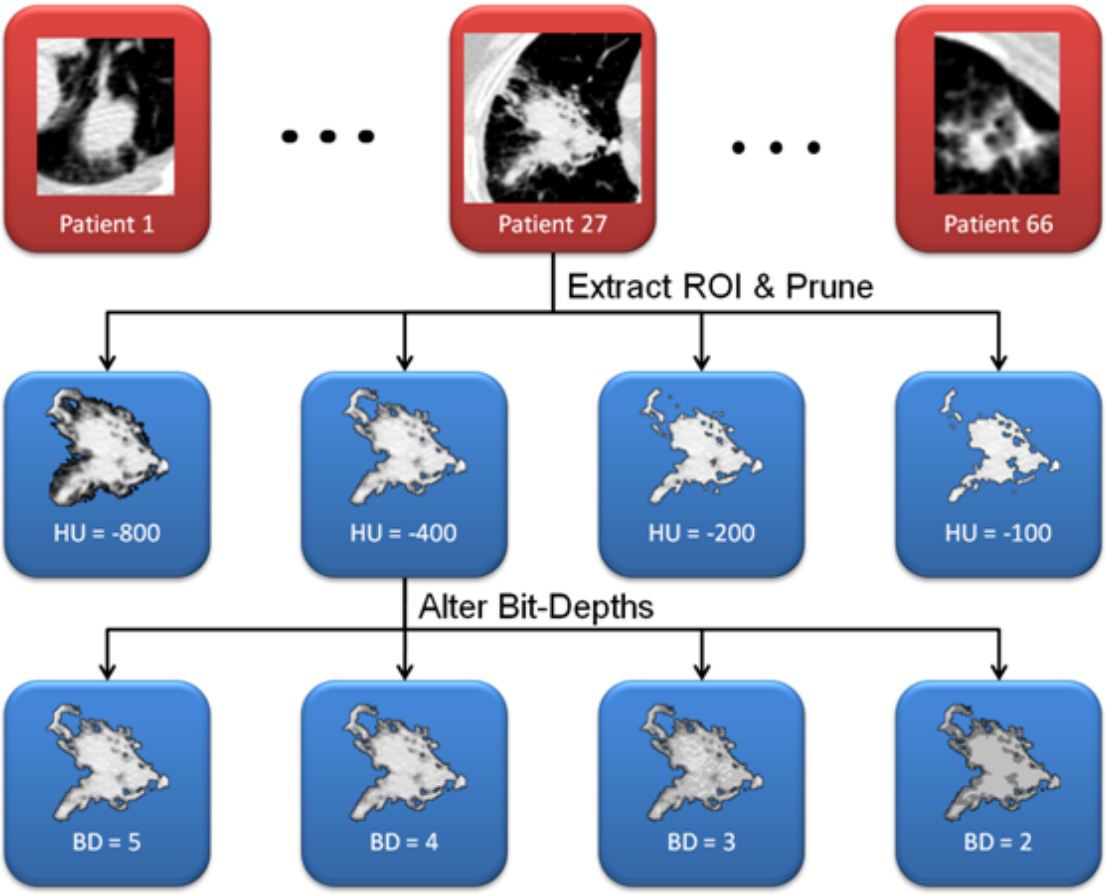

Figure 2.1: Region of interest (ROI) pre-processing. For each patient, the ROI was extracted according to physician-generated contours. Multiple pruned variants of these ROIs were created by removing voxels below various HU cut-offs. From each of these variants, additional variants were created and saved using several different bit depths (BD).

We define a *feature set* as a collection of feature matrices where each feature matrix within the feature set was generated using a unique set of extraction parameters (e.g. HU cutoff and bit depths). To investigate the predictive power of various feature sources, we created four feature sets. Each of of the four feature sets used different feature sources and feature extraction parameters as indicated by Table 2.3.

Table 2.3: Feature set feature sources and extraction parameters. The feature set column denotes the name of a feature set as well as the source of features used to create the feature matrices that it contains (GEO = geometry, INT = intensity histogram, GRAD = absolute gradient image). The feature extraction parameter column shows the different feature extraction parameters that were used for each feature set (HU = Hounsfield unit, BD = bit depth). To populate the associated feature set, a feature matrix was generated for each unique combination.

Feature Set	Feature Extraction Parameters
GEO + COM	HU_{cutoff} = -250, -200, -150, -100, -50 BD_{COM} = 5, 6, 7, 8, 9
GEO + RLM	HU_{cutoff} = -250, -200, -150, -100, -50 BD_{RLM} = 5, 6, 7, 8, 9
GEO + INT + GRAD	HU_{cutoff} = -250, -225, -200,-175, -150, -125, -100, -75, -50 BD_{GRAD} = 4, 5, 6, 7, 8
ALL	HU_{cutoff} = -250, -200, -150, -100, -50 BD_{COM} = 5, 6, 7 BD_{RLM} = 5, 6, 7 BD_{GRAD} = 5, 6, 7

Modelling

The fitness of prediction models was quantified using the *mean squared error* (MSE) between predictions and observations. The method for obtaining the MSE of a feature matrix F with column selection vector $\vec{s}$ using leave-one-out cross-validation is described here and in Figure 2.2. First, a row is omitted from a feature matrix and the remaining training observations are projected into z-score space. These are then projected into principal component (PC) space. Dimensionality reduction is achieved by removing PC dimensions that have trivial components for all training observations. Next, multiple linear regression is performed using the PC space columns indicated by the Boolean vector $\vec{s}$. Tumor responses are used as the regression target. The resulting model is known as the principal component regression (PCR) model (56).

Finally, the omitted row is converted into the same z-score PC space as the training observations and fed into the PCR model in order to obtain a tumor response prediction for the hidden observation. The squared difference between the predicted and observed tumor response is recorded. This process is repeated for each row to obtain the leave-one-out cross-validation MSE.

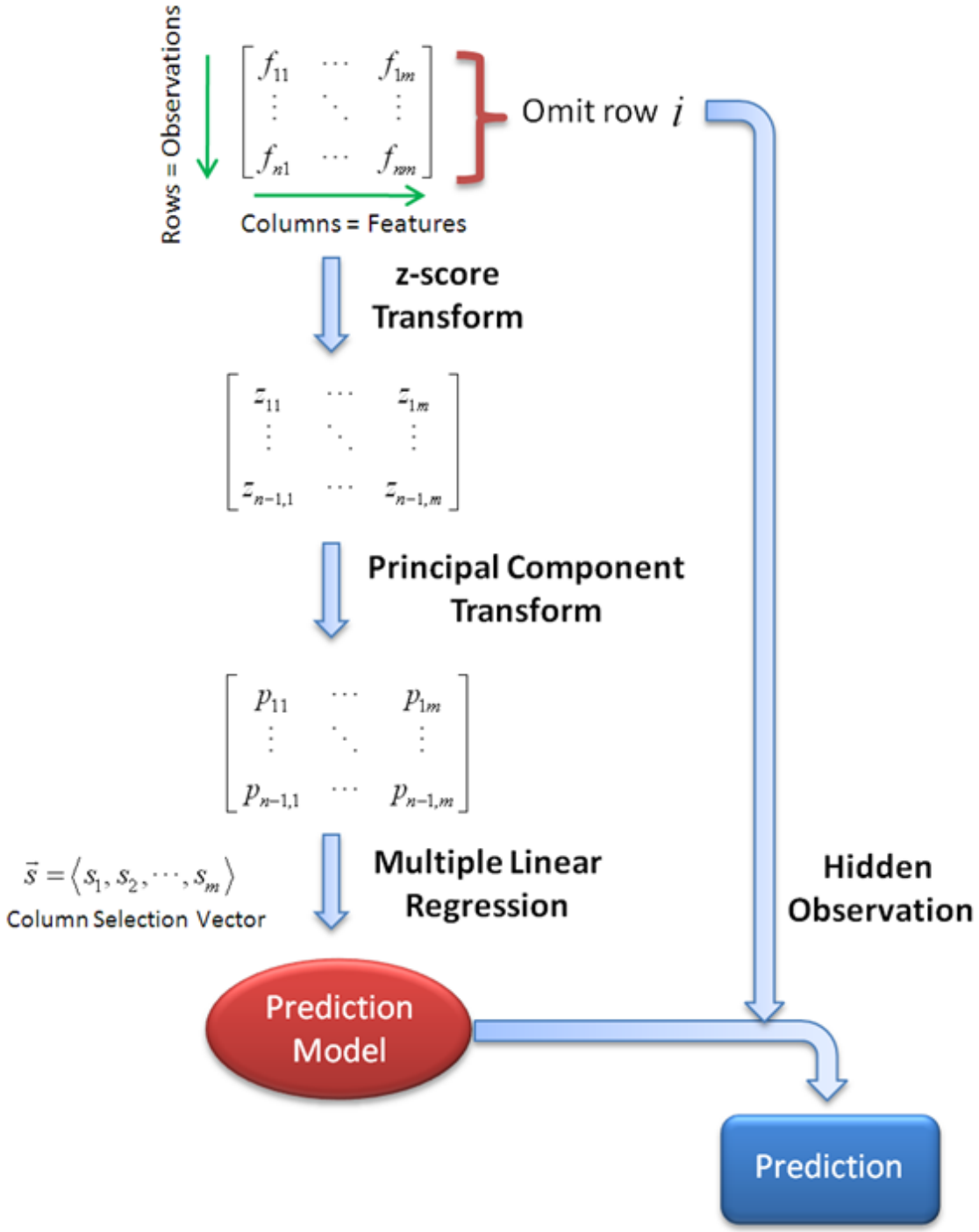

Figure 2.2: Principal component regression with leave-one-out cross-validation; the method for obtaining the mean squared error (MSE) of a feature matrix F with column selection vector $\vec{s}$ using leave-one-out cross-validation.

Optimization

The previous section described how to obtain MSE as a function of the feature matrix and the column selection vector (i.e. $MSE(F, \vec{s})$). This section will describe how the *simulated annealing principal component regression* (SA-PCR) technique was used to find the F and $\vec{s}$ which give the best $MSE(F, \vec{s})$. First, a feature set from Table 2.3 was selected. For each feature matrix F in this feature set, a simulated annealing process (57) is performed using a geometric cooling function with $T_i = 1 \times 10^{-2}$ and $T_f = 1 \times 10^{-4}$. The Boolean vector $\vec{s}$ is randomly initialized and then allowed to evolve according to the simulated annealing process through N = 1,000 steps using $MSE(F, \vec{s})$ as the objective function.

Out of all of the feature matrices in the feature set, the F and $\vec{s}$ that give the lowest MSE are taken to be the best F and $\vec{s}$. Since each feature matrix is generated by a unique set of feature extraction parameters, the best F then implies the best feature extraction parameters (i.e. HU cut-off and bit depths). The best $\vec{s}$ indicates which dimensions of the feature z-score PC space are best for multiple linear regression modeling. Together the best F and $\vec{s}$ define the best *model*.

Because simulated annealing is stochastic, for the same input feature set the SA-PCR algorithm may return different best MSEs for each execution. Therefore, in order to test the stability of the results, we repeated the entire SA-PCR process 100 times for each feature set in Table 2.3.

Validation

Even though the SA-PCR process implements leave-one-out cross-validation, because it considers many feature matrices and many feature z-score PC space selection vectors, there is a possibility that the best MSE and associated F and $\vec{s}$ that it finds are due to coincidence or spurious correlations. To test for this possibility, we performed a negative control study. We again performed the entire SA-PCR process 100 times for each feature

set in Table 2.3 as was done to study the stability of the SA-PCR results. However, in this case, for each run the tumor response vector was randomly permuted with respect to the associated feature matrix (independently for each run). This obliterated any prognostic relationship between the feature matrix and the tumor response vector.

2.3 Results

Algorithm stability

The histogram bins in Figure 2.3 indicate the number of times out of 100 runs that the best MSE returned by the SA-PCR algorithm fell into the indicated MSE range using the specified feature set. The Experiment bars indicate the stability test results where the feature matrices are correctly paired with the tumor response vector. The dispersion of these distributions can be quantified using the coefficient of variation (CV). The GEO+COM feature set had $CV_{MSE} = 3.77 \times 10^{-2}$, the GEO+RLM feature set had $CV_{MSE} = 2.32 \times 10^{-2}$, the GEO+INT+GRAD feature set had $CV_{MSE} = 4.98 \times 10^{-2}$, and the ALL feature set had $CV_{MSE} = 6.01 \times 10^{-2}$.

These small coefficients of variation indicate that the SA-PCR algorithm is stable and returns similar MSEs for the same input, despite being a stochastic algorithm. Thus, if the SA-PCR algorithm was deployed as a part of a clinical workflow, 100 repetitions would not be needed in order to generate a good prediction model. A single run would suffice and would take approximately 18 minutes on a single core machine but is easily parallelizable and runs in about 1.5 minutes on our 12 core machine.

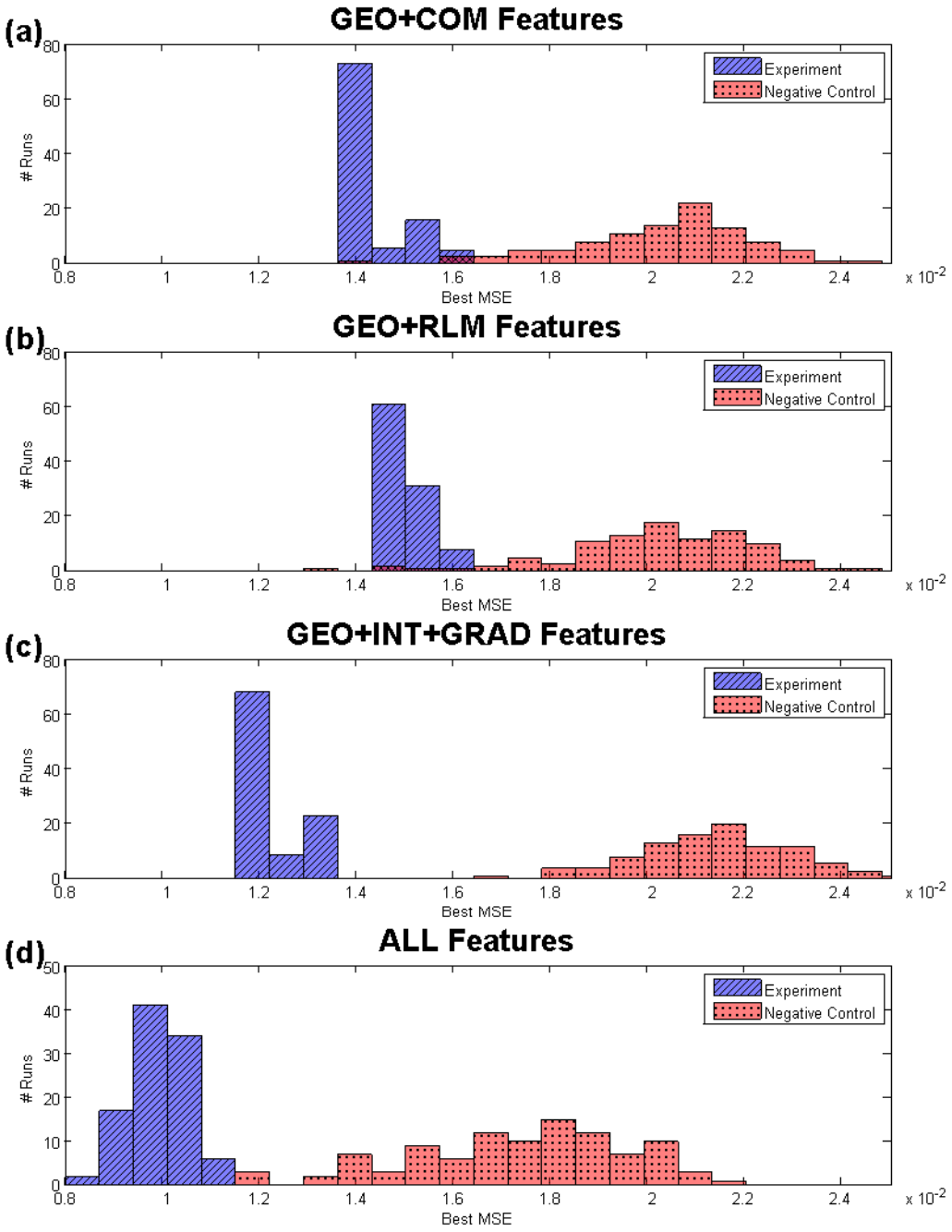

Figure 2.3: SA-PCR algorithm stability and validation. Histogram bins indicate the number of times out of 100 runs that the best MSE returned by the SA-PCR algorithm fell into the indicated MSE range using either a) geometric and COM features, b) geometric and RLM features, c) geometric, intensity histogram, and gradient features, or d) using all features (geometric, intensity histogram, gradient, RLM, and COM). Experiment indicates that the tumor response vector was correctly ordered with respect to the feature matrix. Negative control indicates that the tumor response vector was randomly permuted with respect to the feature matrix (independently for each run).

Algorithm validation

The Negative Control bars in Figure 2.3 indicate the negative control results when the feature matrices are paired with independently randomly permuted tumor response vectors. These negative control MSE distributions are well separated from the experiment distributions for the GEO+INT+GRAD and ALL feature sets. However, there is an overlap between the experiment and negative control distributions for the GEO+COM and GEO+RLM feature sets. The probability of an experiment MSE being due to spurious correlation can be approximated by calculating the one-sided p-value of the mean MSE returned by experiment runs given negative control distribution. Using a normal approximation for the negative control distributions gives $p = 1.09 \times 10^{-3}$ for the GEO+COM feature set, $p = 4.87 \times 10^{-3}$ for the GEO+RLM feature set, $p = 4.36 \times 10^{-9}$ for the GEO+INT+GRAD feature set, and $p = 5.02 \times 10^{-4}$ for the ALL feature set.

If the "good" MSEs found by the SA-PCR stability tests were due to spurious correlations, then the negative control runs should generate similarly "good" MSEs or even better MSEs (some tumor response vector permutations may be more susceptible to spurious correlations). However, this is not what was observed. Figure 2.3 shows that the SA-PCR stability test MSEs are significantly and systematically lower than the negative control MSEs (p-values estimated above). This indicates that there is a valid prognostic relationship between the feature matrix and the tumor response vector that is not due to spurious correlations within the data.

Algorithm Findings

For each of the four feature sets listed in Table 2.3, the best F and $\vec{s}$ (i.e. the best model) was found using 100 repetitions of the SA-PCR method. Figure 2.4 shows these results graphically. The best model predictions from each feature set are displayed with the observed tumor response on the x-axis versus the leave-one-out predicted tumor response on the y-axis. Table 2.4 displays the optimal feature extraction parameters found for each feature set. The overall best model was found using the ALL feature set with a Hounsfield

unit cut-off of -250 and bit depths of 5 for the RLM, 7 for the COM, and 5 for the absolute gradient image.

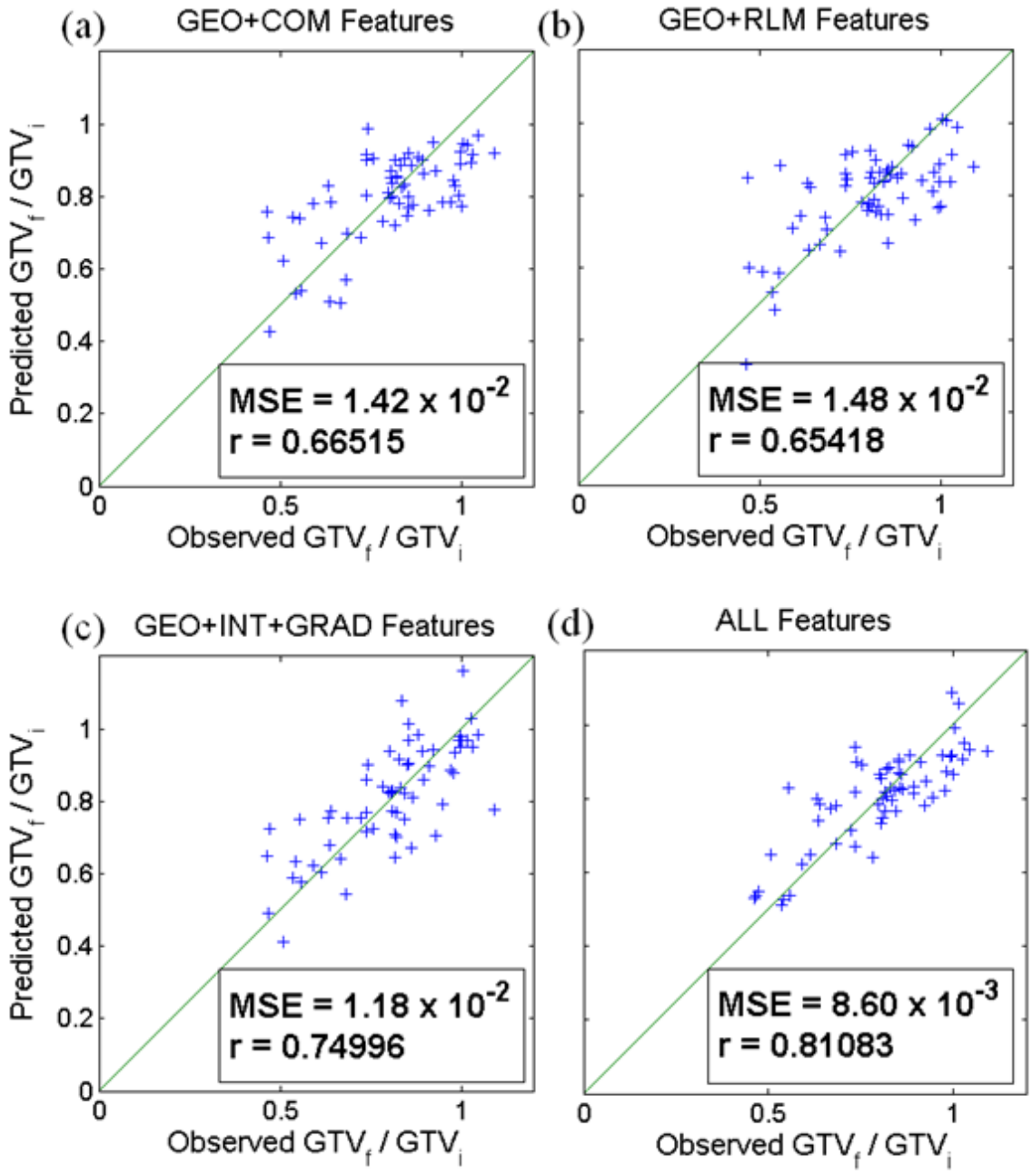

Figure 2.4: Model predictions. Each panel plots the observed tumor responses vs. the leave-one-out tumor response predictions from the best model found in 100 runs of the SA-PCR algorithm operating on one of the following feature sets: a) geometric and COM features, b) geometric and RLM features, c) geometric, intensity histogram, and gradient features, and d) all features. Note: tumor response = the final GTV volume (GTV_f) divided by the initial GTV volume (GTV_i); the MSE and Pearson's *r* value are indicated for each panel; a line with slope = 1 (i.e. perfect prediction) is shown as a reference.

Table 2.4: Optimal feature extraction parameters. A bit depth of 12 was used for all intensity histogram derived features.

Feature Set	HU_{cutoff}	BD_{RLM}	BD_{COM}	BD_{GRAD}	BD_{INT}
GEO + COM	-250	-	7	-	12
GEO + RLM	-250	5	-	-	12
GEO + INT + GRAD	-225	-	-	5	12
ALL	-250	5	7	5	12

To put the MSE values of Figure 2.4 in context, the best model found using the ALL feature set is shown in Figure 2.5 alongside two alternative prediction models. Figure 2.5b shows the predictions of the simplistic mean model where the "left out" patient is predicted to have a tumor response equal to the mean of all of the observed tumor responses. This gives an MSE of 2.51×10^{-2} which is 2.92 times larger than the best MSE found by the SA-PCR algorithm on the ALL feature set (Figure 2.5a). The negative control model shown in Figure 2.5c was generated from the ALL feature set using a random permutation of the tumor response vector. Without the true feature-response (input-output) pairings, this model is unable to produce a linear trend and instead clusters predictions around the mean tumor response value.

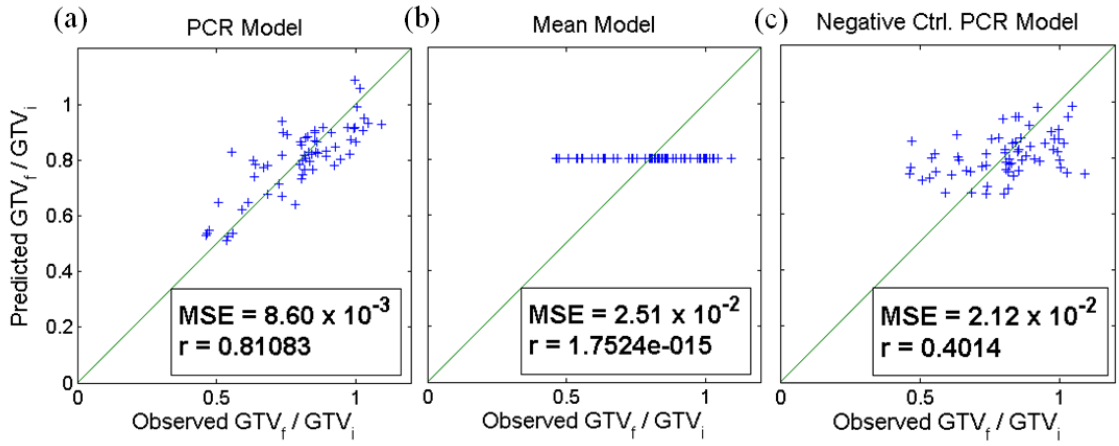

Figure 2.5: Comparison between tumor response models. Panels show the observed tumor response vs. leave-one-out tumor response predictions from a) the optimal SA-PCR model using the ALL feature set, b) the mean model (i.e. predicted response = mean response), and c) an SA-PCR model using the ALL feature set with a random permutation of the tumor responses with respect to the feature matrix.

For each leave-one-out cross-validation step illustrated in Figure 2.2, PCR generates regression coefficients for each of the principal components indicated by $\vec{s}$. By averaging these coefficients across all cross-validations, the coefficient with the largest average absolute value can be identified. The principal component associated with this coefficient is then the most dominant principal component for determining the prediction. Figure 2.6 shows the projection of such a dominant principal component into the feature z-score space for the best model found using the ALL feature set. From this projection it is clear that no particular image feature dominates the dominant principal component. That is, no single image feature seems to be a strong tumor response predictor; rather, it appears that the interaction of multiple features gives rise to the best model.

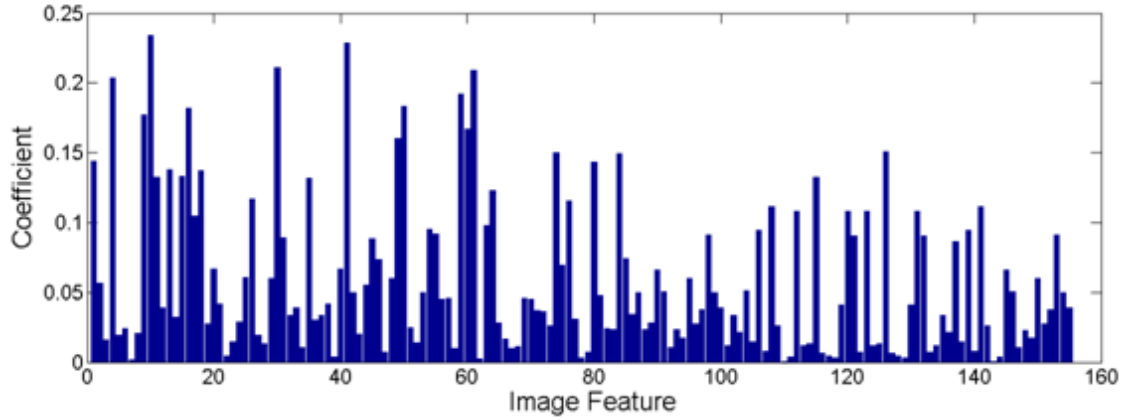

Figure 2.6: Projection of the dominant principal component into the original image feature z-score space for the best model found using the ALL feature set.

2.4 Discussion

Tumor response as a prediction outcome

When deciding treatment options, predicting patient outcome regarding survival and normal tissue toxicity is paramount. However, as our patients are currently part of an ongoing study these data are not yet available. Therefore, tumor response was used as a substitute. We believe that tumor response (i.e. tumor shrinkage) is an acceptable proxy for patient outcome due to the current RECIST guidelines which state:

> *"The use of tumor regression as the endpoint for phase II trials… is supported by years of evidence suggesting that… tumor shrinkage… demonstrates an improvement in overall survival or other time to event measures in randomized phase III studies." (58)*

When the data is available, we plan to determine the ability of quantitative image feature models to predict normal tissue complications and survival.

Pooling photon and proton patients

Although the photon and proton patients technically received different treatments, we believe that for the outcome of tumor response they can be approximated as receiving the same treatment. This can be justified from a theoretical standpoint since the GTVs of each modality were planned to receive *exactly* the same RBE-adjusted dose, and this argument is supported by the empirical findings of (55) (see Table 2.1).

As per the study's design, the only dosimetric difference between the two treatment groups is for normal tissue dose. Thus, in the future, for the outcome of patient survival, there may be differences between the two groups arising from differences in normal tissue dose between the two treatment modalities. In such a scenario, pooling the two patient groups would not be appropriate for prediction of the alternative outcome of survival.

Principal component regression as a model choice

Inspection of the image feature covariance matrices showed that many features were highly correlated. This can make standard multiple linear regression unreliable because small changes in the data can cause large, erratic changes in the estimated coefficients. This is known as the problem of multicolinearity (59). By creating a basis space of orthogonal principal component vectors, PCR is a way to address this problem and also provides dimensionality reduction. However, (60) has shown that the principal components with the most variation are not always optimal for modelling. Thus, we introduced the simulated annealing algorithm as a method to find which principal components should be included for modelling.

Reproducibility and robustness against spurious correlation

The low coefficients of variation for the experiment distributions of Figure 2.3 demonstrate the reproducibility of the SA-PCR algorithm on our dataset. Absent or minimal overlap between the experiment and negative control distributions in Figure 2.3 also demonstrate robustness to spurious correlations. With a p-value of 4.36×10^{-9} for the separation between its experiment and negative control distributions, the GEO+INT+GRAD

feature set results are the least likely to be due to spurious correlation. This implies that in cases where there are few patient observations, it may prudent to omit RLM and COM features for model construction and only use geometric, intensity histogram, and absolute image gradient features. Although this may slightly decrease the MSE compared to using all features, it will give a better guarantee that the model found truthfully identifies a relationship between the image features and the tumor shrinkage. Omitting the RLM and COM features would also make the parameter space easier to search since RLM and COM bit depths are no longer required.

The optimal feature extraction parameters shown in Table 2.4 indicate that the optimal bit depth for each feature source (e.g. RLM, COM, etc.) appears to be independent of the other bit depths. That is, the optimal RLM bit depth found with the GEO+RLM feature set was the same as the optimal RLM bit depth found using the ALL feature set. This was also true for the COM and absolute gradient image bit depths. Thus, in the future it may be prudent to search for each optimal bit depth individually and then combine them when a model using all features is desired. As for the optimal HU cut-off, it appears to be approximately the same for each feature set and is always at or near the lowest HU cut-off that was explored. This indicates that future studies should consider lower HU cut-offs and also hints that the air-tissue boundary may contain important prognostic information (otherwise the optimal HU cut-offs would have been higher).

Model success and applicability

Our findings indicate that a quantitative image feature model can use existing CT images to successfully predict tumor shrinkage (see Figure 2.4). In fact, Figure 2.5 shows that, relative to the simplistic mean model, the tumor shrinkage uncertainty can be reduced by nearly a factor of three. This is very important because tumor shrinkage is prognostic, and thus our predictions are also prognostic. For example, a patient predicted to have a poor response to a particular treatment can be identified before the treatment begins. From here

he can either be labelled as high risk and watched closely or perhaps be transferred to a different treatment strategy.

At the present, however, it is not known whether or not the models developed in this study can be applied to similar patients at other institutions. The model could be invalidated by different treatments, different imaging, and/or different patient demographics. We suggest a group of institutions work together to perform a study similar to (52) across multiple CT scanners to identify robust quantitative image features. Once these features are identified, standardized treatments and demographics could be agreed on, and robust predictive models for patient outcomes could be explored.

2.5 Conclusion

Quantitative image feature models derived from existing pre-treatment CT images were successfully able to predict NSCLC tumor shrinkage, an indicator of treatment efficacy and future survival. This supports the findings of (41) and (46) which showed that binary classifiers operating on quantitative image features could be used to predict survival outcomes for patients with NSCLC. When survival data is available for our patients we will develop similar classifiers and compare AUCs to their findings. Since our cohort's imaging and treatment are both relatively homogenous compared to these studies, such future work could indicate how much imaging and treatment heterogeneity can affect the prognostic power of quantitative image feature models.

Chapter 3: Identification of multi-machine reproducible and non-redundant NSCLC CT image features

3.1 Introduction

In computed tomography (CT) imaging, non-small cell lung cancer (NSCLC) tumors are often described using qualitative descriptors (spiculated, cavitated, heterogeneous, etc.)(61, 62). However, recently there has been a move towards generating additional quantitative, objective features to describe these tumors (17). Traditional quantitative measures have included the one-dimensional tumor size, mean voxel intensity, intensity standard deviation, etc. (50, 58, 63). Newer quantitative measures are usually statistical, model, or transform based and reflect variations in tumor morphology, heterogeneity, and/or texture (24). Motivated by the hypothesis that these quantitative image features are related to gene expression and phenotype, the field of radiomics aims to achieve their automated, high-throughput extraction from standard of care images (23). Ganeshan *et al.* have shown that quantitative measures of NSCLC CT heterogeneity are associated with tumor stage, metabolism (39), hypoxia, and angiogenesis (40). Other quantitative image features have been applied to classify NSCLC as adenocarcinoma or squamous cell carcinoma (36), and lung tumor image fractal dimension has been shown to correlate with tumor stage (38) and can classify pulmonary nodules (37). Thus, several lines of evidence support the radiomics hypothesis for NSCLC.

For NSCLC radiomics to be applied clinically, the intra- and inter-machine reproducibility of image features must be addressed. Towards this goal, the Quantitative Imaging Biomarker Alliance has created a technical committee dedicated towards multi-vender CT scan standardization for the measurement of lung nodules (42). In a CT phantom, Ganeshan *et al.* (43) showed that quantitative measures of entropy and uniformity have lower coefficient of variation values than the mean intensity. For human study, public data is available from 32 NSCLC patients each scanned twice (15 minutes apart) using the same CT machine and protocols (44) (Figure 3.1). Kumar *et al.* (23) recently used this so-called

"test-retest" dataset to identify 39 out of 327 image features that were reproducible, informative, and non-redundant.

The purpose of our study is to expand upon the work of Kumar *et al.* (23) by adding test-retest datasets from additional machines and image types. By studying different image types (cine 4D-CT end-exhale phase vs. average cine 4D-CT vs. breath-hold helical 3D-CT) we can determine which image type will give the most reproducible image features. By studying different machines we can investigate how much machine variation affects reproducibility. We also propose a way to integrate reproducibility and redundancy findings across multiple machines to identify a cross-machine reproducible and non-redundant image feature set. This robust set of NSCLC image features could be useful for the development and validation of multi-machine and multi-institutional NSCLC radiomics models.

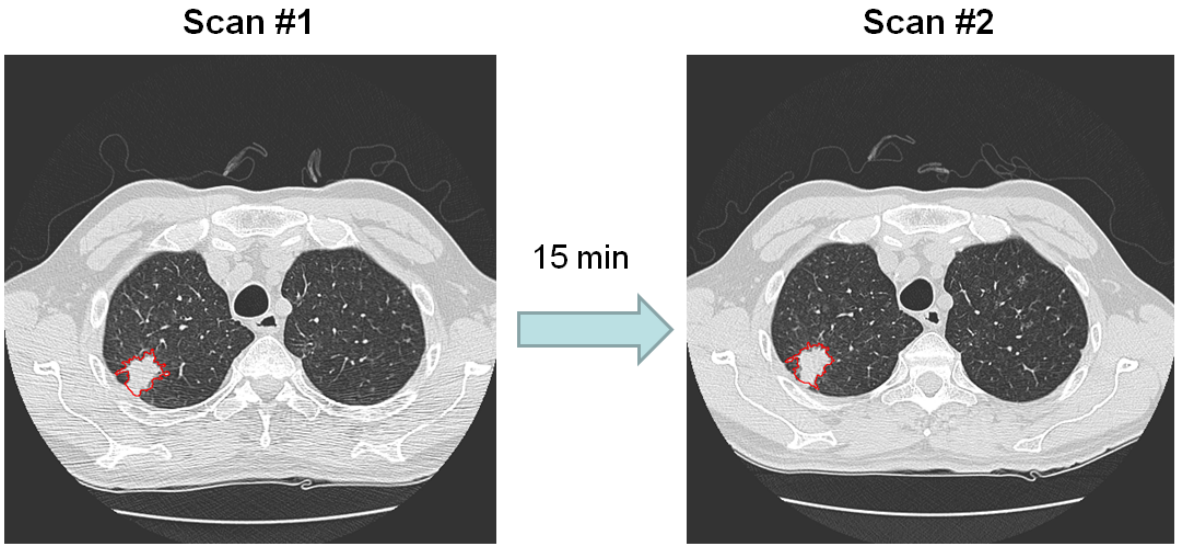

Figure 3.1: Publically available test-retest images taken 15 minutes apart.

Materials and methods

Test-retest datasets

We obtained non-contrast-enhanced cine 4D-CT test-retest scans from 31 patients with NSCLC who were imaged at the University of Texas M.D. Anderson (MDA) Cancer Center in Houston, TX (64). 16 of these test-retest pairs were acquired on a GE (General

Electric Healthcare, Milwaukee, WI) Discovery ST model machine ("M1") and 15 were acquired on a GE LightSpeed RT 16 model machine ("M2"). Imaging parameters are shown in Table 3.1. We also obtained breath-hold, non-contrast-enhanced helical 3D-CT test-retest scans from 25 patients with NSCLC who were imaged at the Memorial Sloan-Kettering Cancer Center in New York, NY (65). This is the dataset used by Kumar *et al.*(23) and is publically available as part of the Reference Image Database to Evaluate Therapy Response (RIDER) (44) collection hosted by the Cancer Imaging Archive (TCIA) at Washington University in Saint Louis, MO.

The complete RIDER dataset has 32 test-retest patients and was collected with two machines: a GE LightSpeed 16 and a GE LightSpeed VCT 64. However, to isolate machine variation we omitted test-retest pairs from the GE LightSpeed VCT 64 machine; we also removed scans with GTVs that had image artifacts or appeared too difficult to contour accurately. This resulted in 25 test-retest pairs all acquired on the same machine, a GE LightSpeed 16 ("M3"). These scans had variable axial pixel dimensions because the field of view was independently set for each scan based on its scout image. Therefore, a "virtual" fourth machine ("M4") was created by taking each M3 scan and performing 3D cubic spline interpolation to generate a new scan with the same voxel size as the MDA images. This was done to test if features could be extracted more reproducibly from scans with a uniform voxel size.

Table 3.1: CT machine parameters. MDACC indicates M.D. Anderson Cancer Center; MSKCC, Memorial Sloan-Kettering Cancer Center.

	Machines		
Parameter	M1	M2	M3
Test-Retest Pairs	16	15	25
Institution	MDACC	MDACC	MSKCC
GE Machine Model	Discovery ST	LightSpeed RT 16	LightSpeed 16
Detector Rows	8	16	16
Scan Mode	Cine 4D	Cine 4D	Helical 3D
Breathing	Free	Free	Insp. hold
kVp	120	120	120
Exposure Time (ms)	500	500	493 ± 42
Tube Current (mA)	100	200	339 ± 63
Filter Type	BODY	BODY	BODY
Convolution Kernel	STANDARD	STANDARD	LUNG
Axial Pixel Size (mm)	0.98	0.98	0.68 ± 0.10
Slice Thickness (mm)	2.5	2.5	1.25
Focal Spot Size (cm)	0.7	0.7	1.2

Feature Extraction

To minimize inter-operator contour variation, a single operator (the primary author) was responsible for contouring the 174 GTVs used in our study. All contouring was performed with the Pinnacle^3 TPS (treatment planning system; Philips Medical Systems, Andover, MA) using its slice by slice auto-contour function with a lower bound of 0 and an upper bound of 500 (values selected to increase automation and reproducibility). For 4D-CT scans, GTVs were contoured on the T50 phase (end-exhale) images and the average (AVG)

images. Clinical treatment plans with physician-generated GTVs were used to identify lesion locations. For the 3D-CT scans, lesions were located and contoured using dataset annotations from the TCIA website. In the interpolated dataset, the binary mask from original voxel size scans underwent a 3D cubic spline interpolation and thresholding to determine the new GTV mask.

For quantitative image feature extraction, we developed the in-house Imaging Biomarker Extractor (IBEX) software package. IBEX was developed using MATLAB (Mathworks Inc., Natick, MA) and has a graphical user interface which can directly access images and regions of interest (ROIs) from the Pinnacle^3 TPS file system. Using this information, it can extract image features from the following 3D feature sources: geometry, intensity histogram, absolute gradient image, co-occurrence matrix (COM) (25), and run-length matrix (RLM) (26). For a complete list of features extracted, refer to Table 3.2. Additionally, for NSCLC GTVs it is common to remove voxels from the GTV region of interest that are below a certain Hounsfield unit (HU) cutoff before performing feature extraction (39, 40). This is thought to refine the ROI, making it more reproducible and focusing feature extraction on tissues other than air. IBEX allows for the specification of an HU cutoff to support this capability.

Table 3.2: Image Features Extracted with the Imaging Biomarker Extractor (IBEX) Software.

Geometry	Intensity Histogram	Absolute Gradient	Run-Length Matrix (RLM)	Co-Occurrence Matrix (COM)
Volume (V)	Mean	Mean	Run Length Nonuniformity	Angular 2nd Moment
Area (A)	Median	Variance	Grey Level Nonuniformity	Contrast
V/A Ratio	Variance	Skewness	Long Run Emphasis	Correlation
Pruned Volume	Skewness	Kurtosis	Short Run Emphasis	Sum of Squares
Pruned Area	Kurtosis	% Non-Zero	Fx. of Image in Runs	Inv. Diff. Moment
Pruned V/A Ratio	Minimum			Sum Average
Fx. Volume Pruned	1th percentile			Sum Variance
	10th percentile			Sum Entropy
	90th percentile			Entropy
	99th percentile			Diff. Variance
	Maximum			Diff. Entropy
	Entropy			

In our analysis, we used IBEX to extract image features from each GTV using the four 3D sources described. RLM features were extracted for 13 different 3D directions, and COM features were extracted for 3 different 3D displacements. For many features, the resulting values can vary substantially depending on the bit depth used to represent the image. Therefore, in order to get a representative sampling of feature values, all absolute gradient image, RLM, and COM features were extracted with three different image bit depths: 2, 6, and 10. The intensity histogram features were extracted using a bit depth of 12, and the geometric features are unaffected by bit depth. In total, for 11 different, equally spaced HU cutoffs between -1000 and 0, 328 image features were extracted for every GTV.

Quantifying reproducibility

For a particular feature, let x be a vector of that feature's values across the patients' first scan. Let y be a vector of that feature's values across the patients' second scan. The concordance correlation coefficient (CCC) (45), a commonly used measure of reproducibility and inter-rater reliability, is then defined as:

$$CCC \equiv 1 - \frac{\left\langle (x-y)^2 \right\rangle}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{3.1}$$

As suggested by McBride (66), we defined reproducible features as those features with CCC ≥ 0.90. Like Kumar *et al.* (23), we also considered selecting features based on their dynamic ranges (DRs), but we noticed a strong overlap between the feature sets that passed a DR cutoff and the feature sets that passed a CCC. Plots of CCC vs. DR^{-2} were very linear (Figure 3.2), so we suspected an approximate relationship between the two metrics existed. Therefore, using reasonable assumptions (verified by our feature data values), we searched for an explanation.

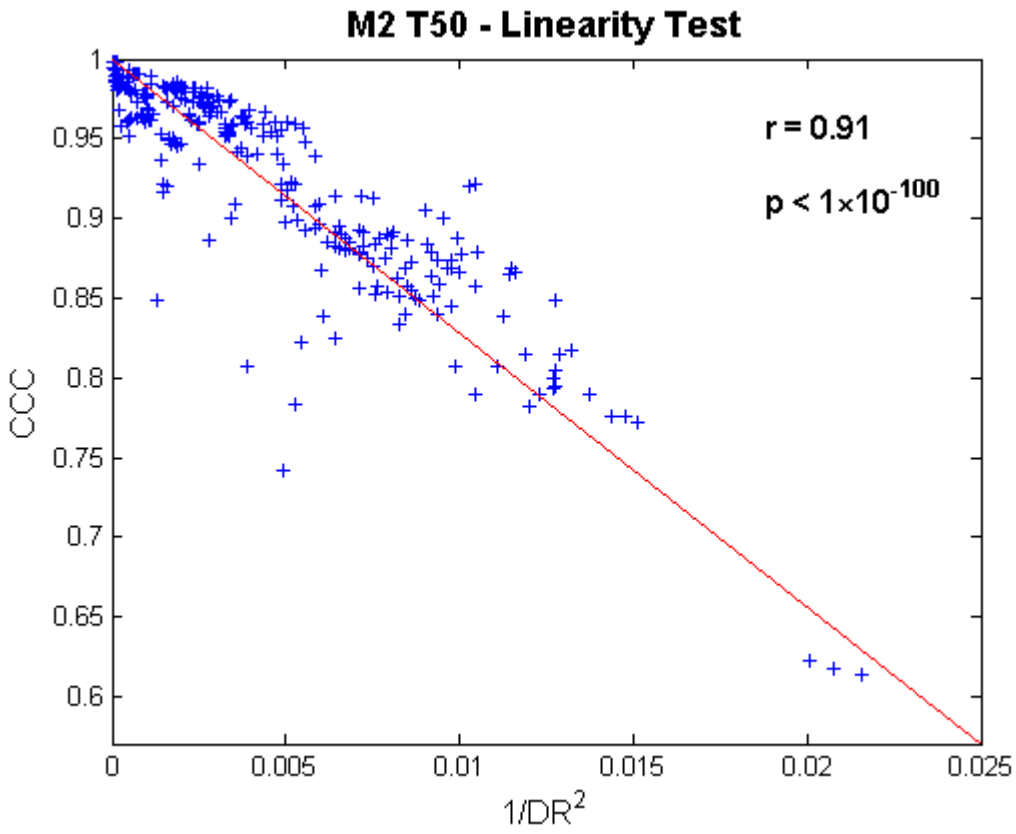

Figure 3.2: Feature CCC vs. 1/DR^2 for 328 features extracted from M2 T50 test-retest scans.

For a particular feature, starting with Equation 3.1, assume that the first scans and the second scans have the same means and variances (i.e. $\mu_x = \mu_y \equiv \mu_p$ and $\sigma_x^2 = \sigma_y^2 \equiv \sigma_p^2$). The CCC then simplifies:

$$CCC \approx 1 - \left\langle (x-y)^2 \right\rangle / 2\sigma_p^2 \tag{3.2}$$

Define the range for the feature: $\Delta f \equiv \max(\max(x), \max(y)) - \min(\min(x), \min(y))$ (3.3)

Define the dynamic range of the feature: $DR \equiv \Delta f \,/\, \left\langle |x-y| \right\rangle$ (3.4)

Assume: $\delta_i \sim N\left(0, \sigma_\delta^2\right)$ and $y \equiv x + \delta \Rightarrow \delta = y - x$, then:

$$\left\langle |\delta| \right\rangle = \sqrt{2/\pi}\,\sigma_\delta \tag{3.5}$$

$$\left\langle \delta^2 \right\rangle = \sigma_\delta^2 \tag{3.6}$$

If $\sigma_p \gg \sigma_\delta$, then for uniform, Gaussian, Poisson, exponential, and other distributions:

$$\langle \Delta f \rangle \propto \sigma_p \tag{3.7}$$

Substituting Equation 3.6 into Equation 3.2 gives: $1 - CCC \propto \sigma_\delta^2 / \sigma_p^2$ (3.8)

Substituting Equations 3.5 and 3.7 into Equation 3.4 gives: $DR \propto \sigma_p / \sigma_\delta$ (3.9)

Therefore: $1 - CCC \propto DR^{-2}$ (3.10)

This explains the observations of Figure 3.2. It also implies that a cutoff on CCC implicitly implies a corresponding cutoff on DR. Therefore, to avoid redundant cutoffs, we chose to only use a CCC constraint since Δf is sensitive to outliers.

Thus, for each of the six machine / image type combinations (M1 T50, M2 T50, M1 AVG, M2 AVG, M3, and M4), we used the CCC constraint to identify which of the 328 features extracted were reproducible. This was done for each of the 11 HU cutoffs between -1000 and 0. In order to find features that were reproducible across multiple machines, the intersection of individual machine reproducible feature sets was taken.

To quantify how machine-sensitive feature reproducibility is, for two machines with identical image types, let A be the set of reproducible features on one machine and B be the set of reproducible features on the other. In terms of these variables, the Dice similarity coefficient (DSC) (67) and the Jaccard index (JI) (68) are defined as follows:

$$DSC(A,B) \equiv \frac{2|A \cap B|}{|A| + |B|} \tag{3.11}$$

$$JI(A,B) \equiv \frac{|A \cap B|}{|A \cup B|} \tag{3.12}$$

To quantify inter-machine feature selection agreement, each of these values was calculated using machine M1 and machine M2 for both T50 images and AVG images.

Finding reproducible, non-redundant features

Applying the CCC reproducibility constraint and taking the feature set intersection across multiple machines substantially reduced the total number of candidate features. However, many of the remaining features were strongly correlated with one another, and it would be useful to find a non-redundant subset of these features. Therefore, for a set of features $\mathcal{F}$ that is reproducible across N machines, let $\rho_k(i) =$ concordance correlation coefficient for feature i on machine k, and let $r_k(i,j)$ be the sample Spearman's rank correlation coefficient between feature i and feature j on machine k. Next, define:

$$\bar{\rho}(i) = \frac{1}{N}\sum_{k=1}^{N}\rho_k(i) \tag{3.13}$$

$$d_k(i,j) = 1 - \left|r_k(i,j)\right| \tag{3.14}$$

$$\bar{d}(i,j) = \frac{1}{N}\sum_{k=1}^{N} d_k(i,j) \tag{3.15}$$

Where $\bar{\rho}(i)$ is the mean CCC for feature i across all N machines, $d_k(i,j)$ is the similarity distance between features i and j on machine k, and $\bar{d}(i,j)$ is the mean similarity distance between features i and j across all N machines.

Using these definitions, the similarity distance between all of the feature pairs of $\mathcal{F}$ can be calculated for each machine, and the mean similarity distances between the feature pairs of $\mathcal{F}$ can be established. The features present in $\mathcal{F}$ can then be hierarchically clustered based on the $\bar{d}(i,j)$ distance function. By clustering until a threshold value is reached, several non-redundant, reproducible clusters can be identified. By picking the feature from each cluster with the highest average reproducibility across machines (Equation 3.13), a set of features that are non-redundant and reproducible across multiple machines can be identified. This process was performed for several different HU cutoffs and machine / image type combinations. In each case, the clustering algorithm used an *average* linkage function and an *average* mean similarity distance clustering threshold of 0.1.

3.3 Results

Single Machine Reproducibility

Figure 3.3 shows the percent of the 328 image features that had a certain CCC value or higher when no voxels are removed from the GTV prior to feature extraction. Intersections with CCC = 0.90 indicate the number of features that were considered reproducible. In the figure, the three image types track each other initially, but then diverge to have considerably different levels of reproducibility. The cine 4D-CT average images were the most reproducible and had curves that bent down the least; 90.5% of the features were reproducible for M1 AVG and 94.5% of the features were reproducible for M2 AVG. The cine 4D-CT T50 scans were the next most reproducible (M1 T50 = 75.0% pass; M2 T50 = 71.0% pass), and the helical 3D-CT breath-hold scans had the fewest number of reproducible features (M3 = 61.0% pass). Images interpolated with uniform voxel sizes did not improve the number of reproducible features (M4 = 57.3% pass). Figure 3.3 was generated multiple times using various HU cutoffs to generate Figure 3.4. In Figure 3.4, it is apparent that for all machines and image types, as the HU cutoff goes down the number of reproducible features tends to goes up. This trend implies that for maximum feature reproducibility, no HU cutoff should be used.

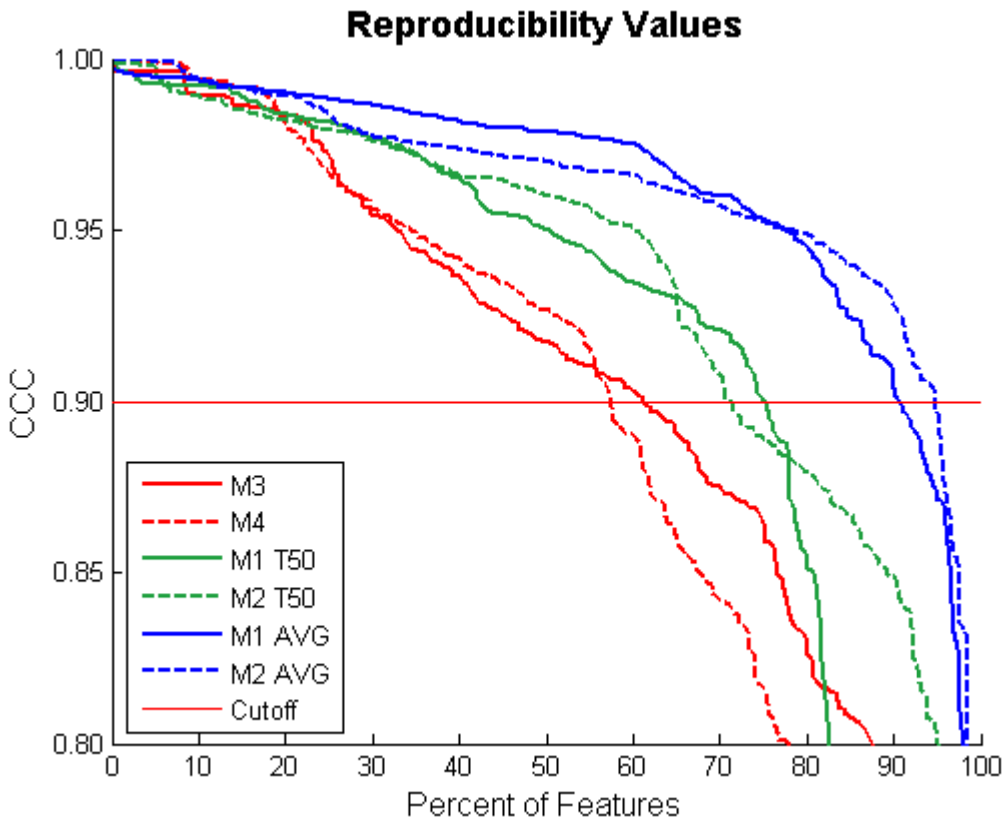

Figure 3.3: Percent of GTV image features with a CCC greater than or equal to the indicated value.

Multi-machine reproducibility

For each of the data points in Figure 3.4, each machine has a set of reproducible features. By taking the intersection of these sets we found features that were reproducible across multiple machines. Figure 3.5 shows the reproducible feature percentage for several different multi-machine combinations. The same trend is noted: the number of multi-machine reproducible features increases as the HU cutoff is decreased. At an HU cutoff of -1000 (i.e. no cutoff), for average scans, 86.3% of features were reproducible across both machines (M1 and M2), for T50 scans, 52.1% of features were reproducible across both machines, and for "single phase" scans (M1 T50, M2 T50, and M3), 42.1% of features were reproducible across all three machines. Exchanging M3 with M4 did not appreciably affect the number of multi-machine reproducible features (41.5%).

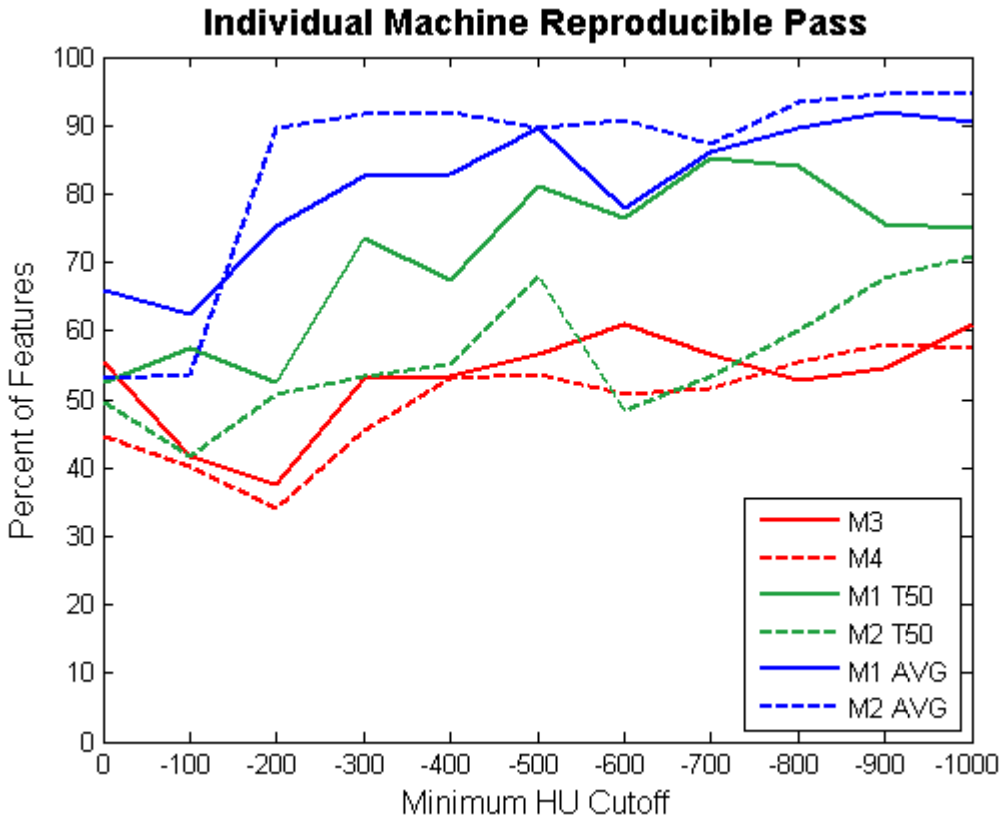

Figure 3.4: Percent of features that are reproducible (CCC ≥ 0.90) when voxels with CT numbers less than the specified HU cutoff are removed from the GTV before feature extraction.

To investigate if feature reproducibility is machine-sensitive, Figure 3.6 displays M1 CCC vs. M2 CCC for both T50 and AVG scans (no HU cutoff). Points falling to the right of the vertical lines correspond to reproducible features on M1, and points falling above the horizontal lines correspond to reproducible features on M2. Venn diagram areas indicate the relative number of reproducible features that were common or unique for the two machines. The Dice similarity coefficients and Jaccard indices and indicate that T50 feature reproducibility is relatively machine-sensitive (DSC = 0.71, JI = 0.55) while AVG feature reproducibility is relatively machine-insensitive (DSC = 0.93, JI = 0.87). However, if HU cutoffs are applied, both image types begin to show machine-sensitivity. For example, an HU cutoff of 0 gives DSC = 0.61 and JI = 0.44 for T50 images and DSC = 0.54 and 0.37 for AVG images.

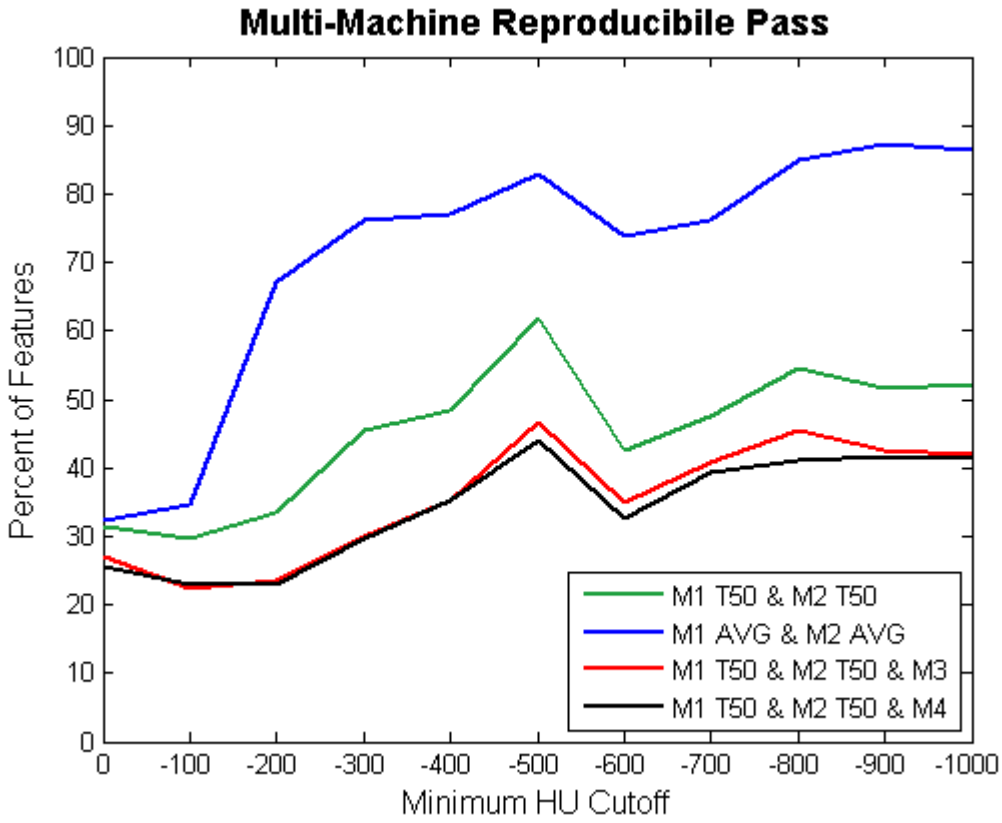

Figure 3.5: Percent of features that are reproducible across all of the machines specified given a particular HU cutoff.

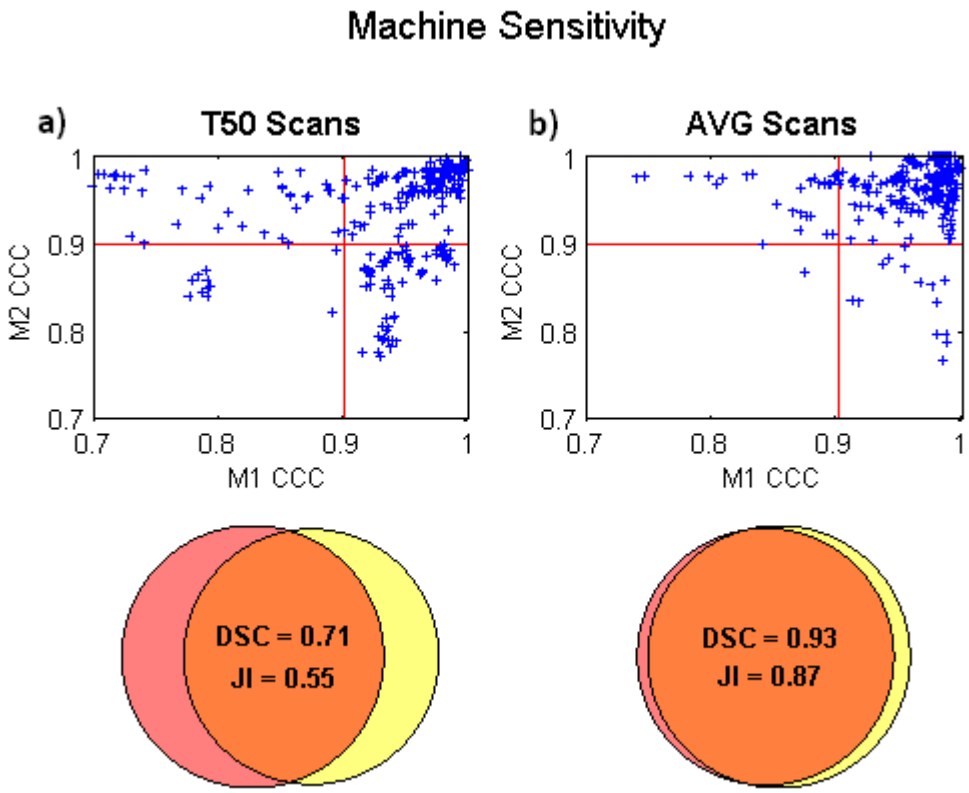

Figure 3.6: M1 CCC vs. M2 CCC. Panel a) depicts the relationship for T50 scans and panel b) depicts the relationship for AVG scans. Venn diagrams, Dice similarity coefficients (DSC), and the Jaccard indices (JI) are displayed. The CCC ≥ 0.90 cutoff is shown for both axes. No HU cutoff was used.

Reproducible, non-redundant feature sets

Across single phase scans (M1 T50, M2 T50, and M3), 42.1% or 138 out of 328 image features were reproducible on all machines (no HU cutoff). The cluster heat map (69) in Figure 3.7 shows the values of these 138 reproducible features for the 25 M3 test-retest pairs (50 scans). Repeated horizontal patterns present in the cluster heat map indicate that that there is considerable feature redundancy, especially for the RLM grey level nonuniformity (GLNU) features and RLM run length nonuniformity (RLNU) features.

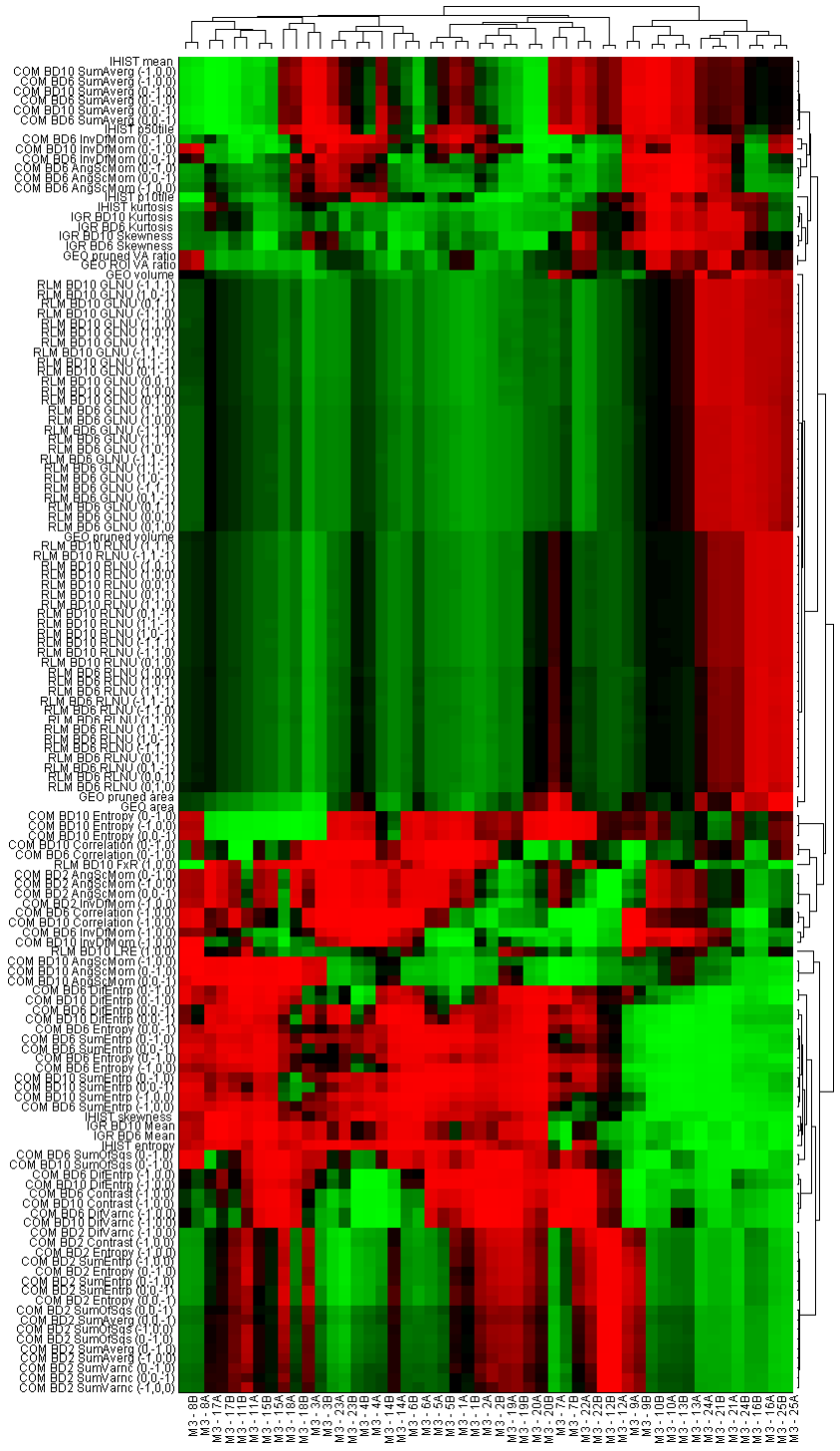

Figure 3.7: Cluster heat map representation of the M3 scans using the 138 features that passed the reproducibility cutoff on single phase scans. No HU cutoff was used. Green cells correspond to a feature value above the mean, red cells correspond to feature values below the mean, and the color intensity indicates the deviation magnitude. Image features (rows) were clustered using the Pearson's distance metric, and patient scans (columns) were clustered using a Euclidean distance metric. Both used an average linkage function. Scans are labeled such that the number indicates the test-retest pair; 'A' and 'B' represent scan #1 and scan #2, respectively. Dendrograms indicate the hierarchical relationships both between the features and between the scans

To address this, Equation 3.14 was used to quantify how redundant two features are on a particular machine. The distribution of feature similarity distances ($d_k(i, j)$) for different machines is shown in Figure 3.8 (no HU cutoff). Feature pairs with distances near zero are very redundant; feature pairs with distances near one are non-redundant. The different shape and increased number of non-redundant feature pairs for Figure 3.8b (M2 T50) compared to Figure 3.8a (M1 T50) indicate that feature redundancy is moderately machine-sensitive. Similar histograms (data not shown) show different feature similarity distance distributions both for M1 T50 vs. M1 AVG and M2 T50 vs. M2 AVG, indicating that feature redundancy is also image type sensitive. In Figure 3.8c (M3 data) both the machine and image type are changed compared to Figure 3.8a and Figure 3.8b. Thus, its feature pair distance distribution appears quite different; relatively more feature pairs are either very redundant or very non-redundant, with fewer feature pairs in between. Figure 3.8d uses Equation 3.15 and shows the distribution of the mean similarity distances ($\bar{d}(i, j)$) across the three machines. Its final (0.9 to 1.0) distance bin is relatively empty, implying that it is rare for a feature to be strongly non-redundant across all three machines.

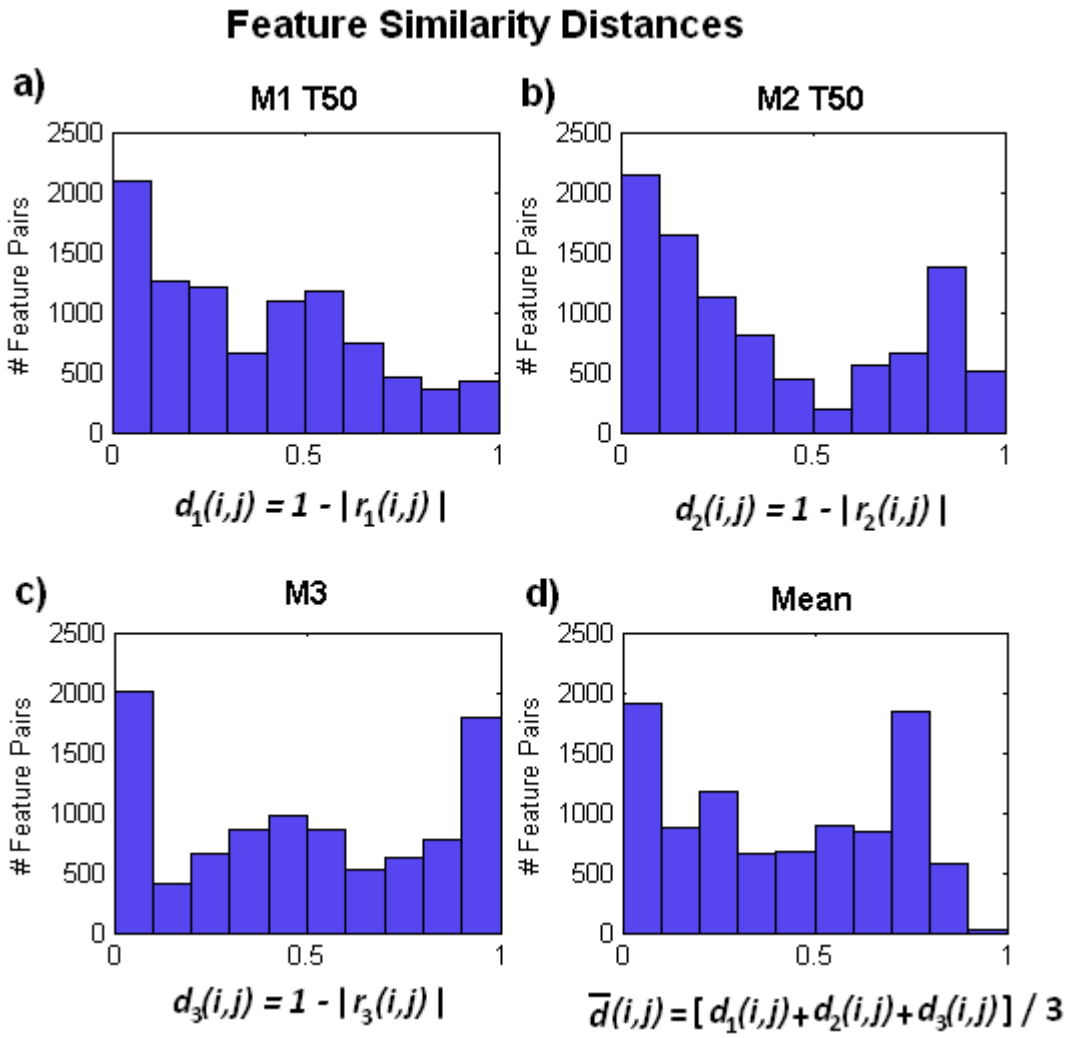

Figure 3.8: Panels a), b), and c) show histograms of the feature pair similarity distances for the 9453 feature pairs possible from the 138 features that passed the reproducibility cutoff for all single phase scans with no HU cutoff. Panel d) shows the distribution of mean feature similarity distances.

Figure 3.9 shows a dendrogram generated from the hierarchical clustering of the 138 features that were reproducible across all single phase scans (no HU cutoff). It used the mean similarity distance of a feature pair as a distance function (distribution shown in Figure 3.8d). It also used an *average* linkage function, so the distance between two feature clusters is the *average* mean similarity distance, and a cutoff value of 0.1 was selected to find a finite number of relatively non-redundant feature clusters (in this case 23).

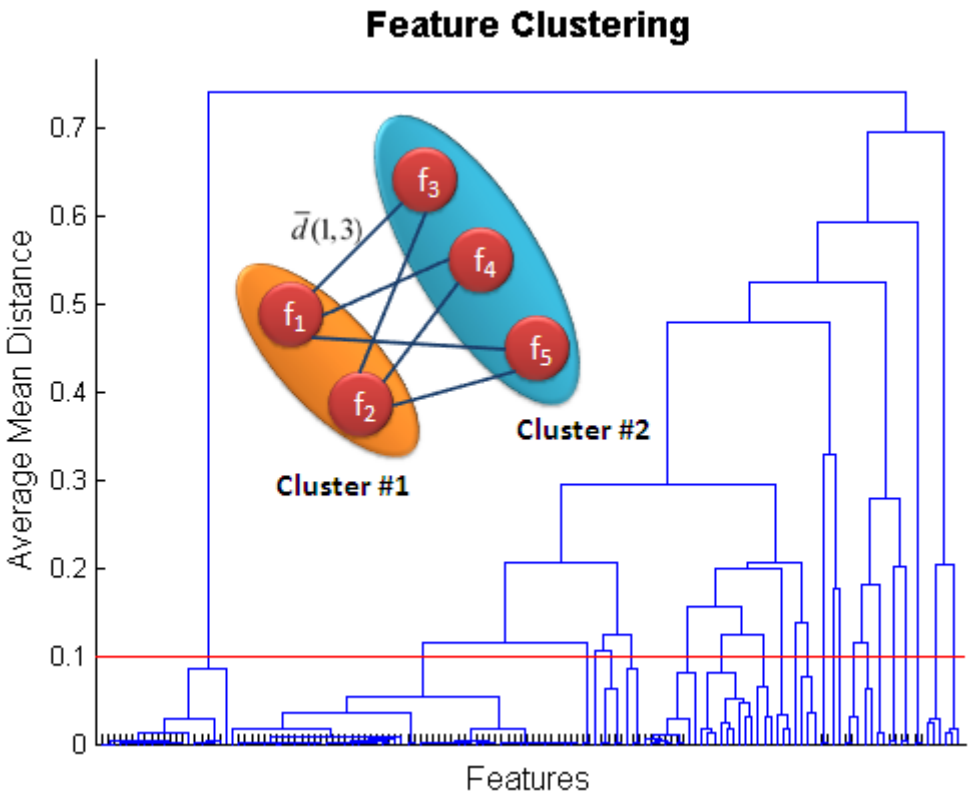

Figure 3.9: Dendrogram illustrating hierarchical feature clustering using the *average* mean similarity distance between feature clusters as a distance metric. The *average* mean similarity distance between two clusters is the *average* of the mean similarity distances between the elements of each cluster (see inset). Two clusters merge (horizontal bar) when the average mean similarity distance between their constituent features becomes greater than the y-value indicated in the diagram. There are 23 clusters at the average mean similarity distance cutoff of 0.1 (red line). Feature values came from the 138 features that passed the reproducibility cutoff on all single phase scans (no HU cutoff).

This whole process (identifying multi-machine reproducible features and clustering to a threshold) was done for 4 different multi-machine combinations and 11 different HU cutoffs. All used the same 0.1 clustering threshold. The results are displayed in Figure 3.10, and show a similar trend to those in Figure 3.5: as the HU cutoff goes down, the number of non-redundant feature clusters goes up. Therefore, we used HU cutoff $= -1000$ and generated a list of reproducible, non-redundant features for each of the 4 multi-machine combinations in Figure 3.10. See Table 3.3. Each entry in the table is a representative feature from one of the reproducible, non-redundant clusters of the multi-machine combination indicated. For cine 4D-CT average images and for cine 4D-CT T50 images, we

recommend columns 1 and 2, respectively. For “single phase” mixed cine 4D-CT T50 images and helical 3D-CT breath-hold images, we recommend column 3. For helical 3D-CT breath-hold images we recommend column 4.

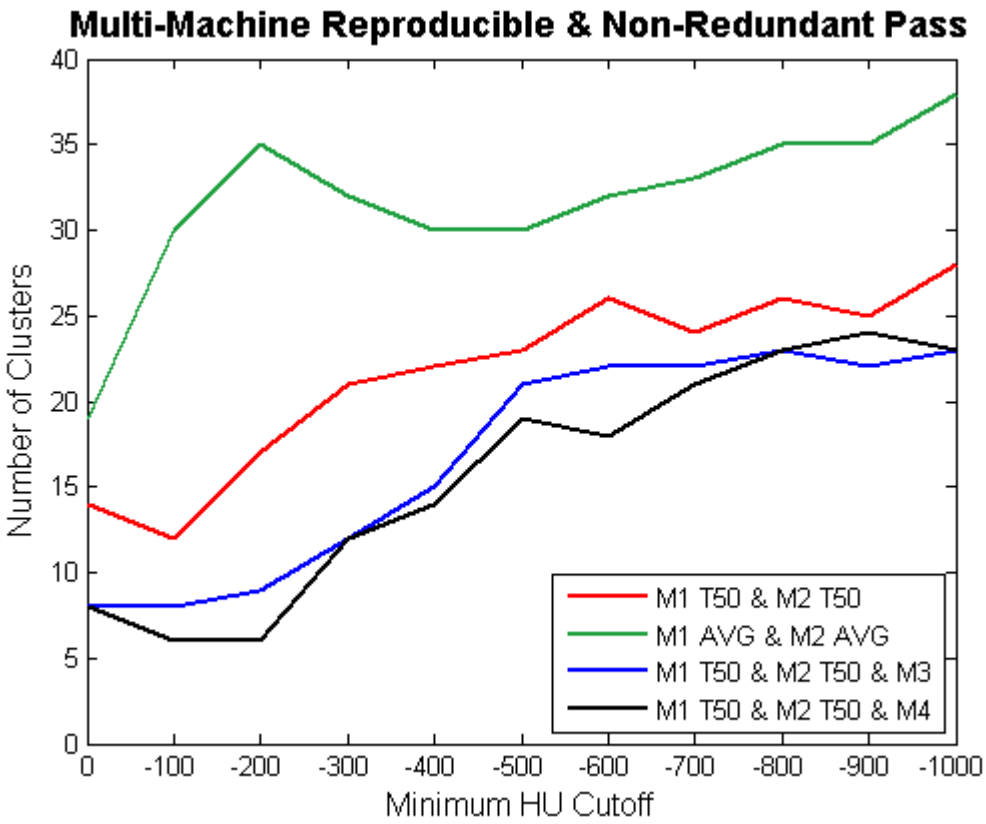

Figure 3.10: Number of non-redundant feature clusters (average mean similarity distance > 0.1) generated by hierarchically clustering features that are reproducible across all of the machines specified.

Table 3.3: Multi-machine reproducible and non-redundant feature lists. Coordinates indicate either a direction (RLM) or a displacement (COM); BD indicates bit depth; VA ratio, volume to area ratio; GEO, geometric; IHIST, intensity histogram; IGR, absolute gradient image; RLM, run length matrix; COM, co-occurrence matrix; p10tile, 10th percentile; FxR, fraction in runs; SRE, short run emphasis; LRE, long run emphasis; RLNU, run length nonuniformity; GLNU, grey level nonuniformity; PNonZero, probability non-zero.

M1 AVG & M2 AVG (38 features)	M1 T50 & M2 T50 (28 features)	M1 T50 & M2 T50 & M3 (23 features)	M3 (33 features)
IHIST entropy	GEO VA ratio	GEO VA ratio	GEO volume
IHIST kurtosis	IHIST entropy	GEO volume	IHIST kurtosis
IHIST max	IHIST kurtosis	IHIST kurtosis	IHIST p10tile
IHIST mean	IHIST p10tile	IHIST p10tile	IHIST skewness
IHIST min	IHIST p1tile	IHIST skewness	IGR BD10 PNonZero
IHIST p10tile	IHIST skewness	IGR BD10 Skewness	IGR BD6 Mean
IHIST p1tile	IGR BD2 Variance	IGR BD6 Mean	RLM BD10 FxR (1,0,1)
IHIST p99tile	IGR BD6 Mean	RLM BD10 LRE (1,0,0)	RLM BD10 FxR (1,0,-1)
IHIST variance	RLM BD10 RLNU (0,1,0)	COM BD10 AngScMom (-1,0,0)	RLM BD10 FxR (-1,1,1)
IGR BD10 Variance	RLM BD10 SRE (1,0,0)	COM BD10 Correlation (0,-1,0)	RLM BD10 FxR (1,1,-1)
IGR BD2 Kurtosis	COM BD10 AngScMom (-1,0,0)	COM BD10 Correlation (-1,0,0)	RLM BD10 LRE (0,1,0)
IGR BD2 Variance	COM BD10 Correlation (-1,0,0)	COM BD10 DifEntrp (-1,0,0)	RLM BD10 LRE (1,0,0)
IGR BD6 Skewness	COM BD10 DifEntrp (0,-1,0)	COM BD10 Entropy (-1,0,0)	RLM BD10 LRE (1,1,0)
RLM BD10 LRE (0,1,-1)	COM BD10 DifEntrp (-1,0,0)	COM BD10 InvDfMom (0,-1,0)	RLM BD10 LRE (-1,1,0)
RLM BD10 RLNU (0,0,1)	COM BD10 Entropy (-1,0,0)	COM BD10 InvDfMom (-1,0,0)	RLM BD6 SRE (0,1,0)
RLM BD2 FxR (1,1,0)	COM BD10 InvDfMom (0,-1,0)	COM BD10 SumEntrp (-1,0,0)	RLM BD6 SRE (1,1,0)
RLM BD2 FxR (-1,1,-1)	COM BD10 InvDfMom (-1,0,0)	COM BD10 SumOfSqs (0,-1,0)	RLM BD6 SRE (1,1,-1)
RLM BD2 RLNU (1,0,-1)	COM BD10 SumEntrp (-1,0,0)	COM BD2 SumVarnc (0,0,-1)	COM BD10 AngScMom (-1,0,0)
RLM BD2 SRE (1,1,0)	COM BD10 SumOfSqs (0,-1,0)	COM BD6 DifEntrp (-1,0,0)	COM BD10 Correlation (0,-1,0)
RLM BD2 SRE (-1,1,0)	COM BD2 SumOfSqs (0,0,-1)	COM BD6 DifVarnc (-1,0,0)	COM BD10 DifEntrp (0,0,-1)
RLM BD6 GLNU (0,0,1)	COM BD2 SumVarnc (0,0,-1)	COM BD6 Entropy (0,-1,0)	COM BD10 DifEntrp (-1,0,0)
COM BD10 AngScMom (0,0,-1)	COM BD6 Contrast (0,0,-1)	COM BD6 InvDfMom (0,0,-1)	COM BD10 Entropy (0,-1,0)
COM BD10 Correlation (0,-1,0)	COM BD6 Correlation (0,0,-1)	COM BD6 InvDfMom (0,-1,0)	COM BD10 InvDfMom (0,0,-1)
COM BD10 DifVarnc (-1,0,0)	COM BD6 DifEntrp (0,-1,0)		COM BD10 InvDfMom (0,-1,0)
COM BD10 Entropy (0,-1,0)	COM BD6 DifEntrp (-1,0,0)		COM BD10 InvDfMom (-1,0,0)
COM BD10 InvDfMom (0,-1,0)	COM BD6 DifVarnc (0,-1,0)		COM BD10 SumEntrp (0,0,-1)
COM BD10 SumEntrp (0,0,-1)	COM BD6 DifVarnc (-1,0,0)		COM BD10 SumEntrp (-1,0,0)
COM BD2 Contrast (0,0,-1)	COM BD6 InvDfMom (0,0,-1)		COM BD2 SumVarnc (0,0,-1)
COM BD2 Contrast (-1,0,0)			COM BD6 Correlation (-1,0,0)
COM BD6 AngScMom (0,-1,0)			COM BD6 Entropy (-1,0,0)
COM BD6 DifEntrp (0,0,-1)			COM BD6 InvDfMom (0,-1,0)
COM BD6 DifEntrp (0,-1,0)			COM BD6 InvDfMom (-1,0,0)
COM BD6 DifEntrp (-1,0,0)			COM BD6 SumVarnc (0,0,-1)
COM BD6 DifVarnc (0,0,-1)			
COM BD6 DifVarnc (0,-1,0)			
COM BD6 Entropy (0,0,-1)			
COM BD6 InvDfMom (0,-1,0)			
COM BD6 InvDfMom (-1,0,0)			

To confirm that the recommended feature sets are non-redundant, we generated a cluster heat map (Figure 3.11) for all single phase scans (32 M1 T50 scans, 30 M1 T50 scans, and 50 M3 scans) using the 23 features from Table 3.3, column 3. The absence of repeated vertical patterns in the heat map indicates that the selected features are non-redundant. Furthermore, blinded to the fact that test-retest pairs were present, 55 out of 56 test-retest pairs were correctly placed adjacent to one another in the figure. This indicates that the selected features were sufficiently informative to accurately match and distinguish patients.

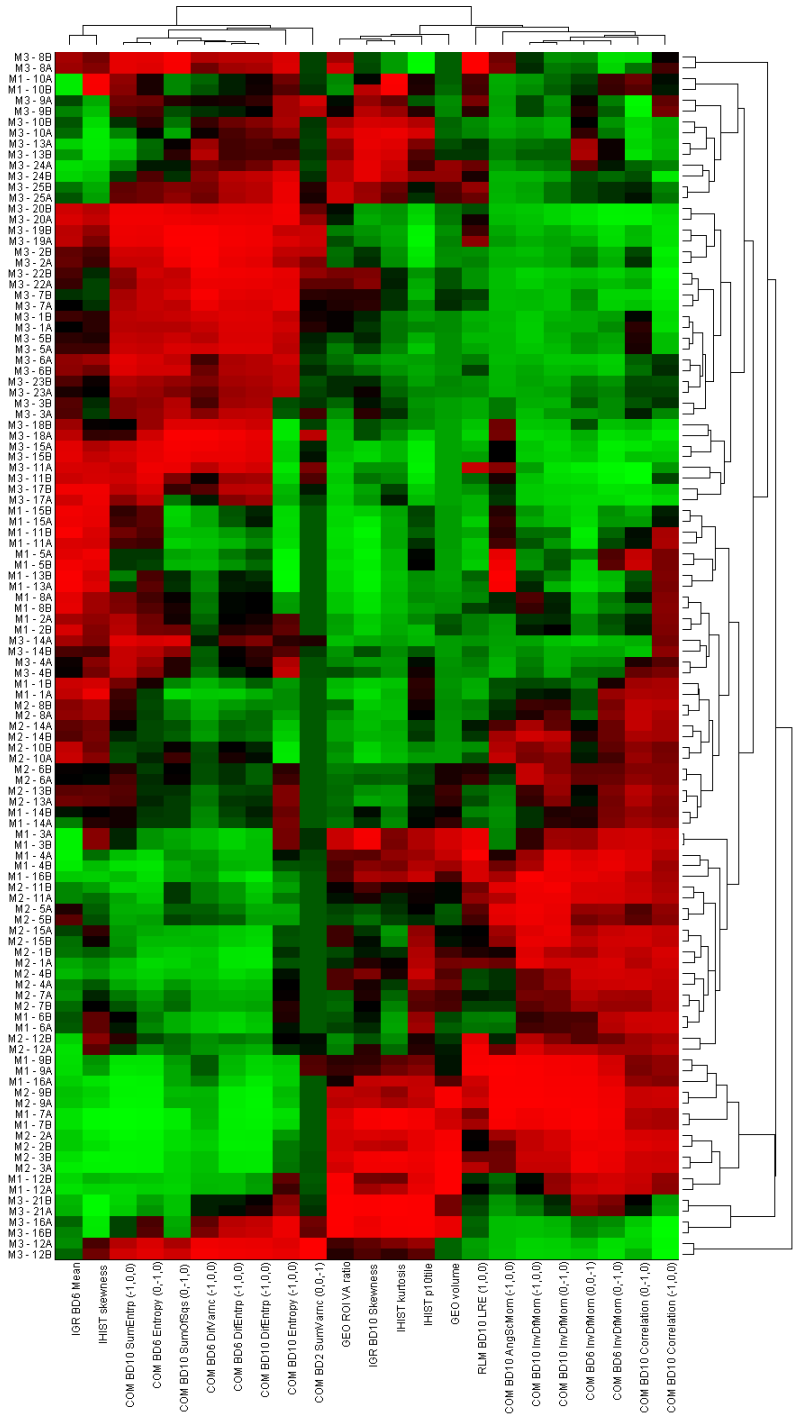

Figure 3.11: Cluster heat map representation of all scans from M1 T50, M2 T50, and M3 (no HU cutoff). Green cells correspond to a feature value above the mean, red cells correspond to feature values below the mean, and the color intensity indicates the deviation magnitude. For readability, orientations are reversed from Figure 3.6. Each of the 23 features shown had the highest average CCC from one of the 23 independent clusters shown in Figure 3.9. Features (columns) were clustered using the Pearson's distance metric, and patient scans (rows) were clustered using a Euclidean distance metric. Both used an average linkage function. Scans are labeled such that the number indicates the test-retest pair; 'A' and 'B' represent scan #1 and scan #2, respectively. Dendrograms indicate the hierarchical relationships between the features and between the scans.

3.4 Discussion

Our study was strongly inspired by Kumar *et al.* (23), however, there are several key differences. We applied their techniques to several different machines and image types and studied how different machines and image types affected feature reproducibility and redundancy. We also studied feature reproducibility as a function of HU cutoff (i.e. GTV pruning) and omitted the feature dynamic range requirement. Additionally, we developed a strategy to integrate the findings to produce a multi-machine reproducible and non-redundant feature set. We also omitted some of the patients used by Kumar *et al.* (23), used a different set of image features, and had a different semi-automated contouring process. These methodological differences could explain why our CCC values were relatively higher, leading us to chose to use a reproducibility cutoff of 0.90 instead of 0.85 (23).

CT image type was found to strongly affect feature reproducibility (Figure 3.3, Figure 3.4, and Figure 3.5). Features from cine 4D-CT average images were substantially more reproducible than those associated with T50 images, and T50 associated features were somewhat more reproducible than breath-hold helical 3D-CT features. This may seem counter-intuitive since T50 images are intended to “freeze” motion. However, because average images are an average of 10 image phase reconstructions, they are less susceptible to photon noise and various 4D-CT artifacts (70-72). As for the breath-hold CTs, it is difficult to draw conclusions since the machine, image type, and convolution kernel varied simultaneously compared to the other datasets. However, M4 did not appear to have better reproducibility than M3. That is, no significant reproducibility improvements were observed when scans were interpolated to have a uniform voxel size. If raw CT data were available it may have been possible to reconstruct images with a standard voxel size and achieve better results.

As the HU cutoff goes down (i.e. as fewer low-intensity GTV voxels are pruned), we found 1) reproducibility becomes less machine-sensitive (Figure 3.6), 2) individual machine

reproducibility goes up (Figure 3.4), 3) multi-machine reproducibility goes up (Figure 3.5), and 4) the number of multi-machine, non-redundant clusters goes up (Figure 3.10). This last finding is particularly interesting since it implies that the low intensity voxels contain non-redundant information. Since the vast majority of the low intensity voxels appear at the tumor periphery, these findings indicate that the periphery contains valuable non-redundant information which can be used to discriminate tumors. This is understandable, since non-contrast-enhanced CT has limited soft tissue contrast; if the periphery is removed, there are fewer intensity variations to extract independent features from. Thus, we recommend against removing voxels based on their intensity values as a pre-processing step to feature extraction. This is in opposition to the NSCLC CT protocol of Ganeshan *et al.* (39, 40), but we believe that it is likely due to methodological differences. In their studies, directionally independent first-order statistics (e.g. entropy and uniformity) were extracted from the filtered 2D image slice that had the largest transverse tumor length. In our work, features are extracted from the intensity histogram, absolute gradient image, RLM, and COM values evaluated over the entire unfiltered GTV. Therefore, as a caveat, we should emphasize that our tumor pre-processing (i.e. pruning) protocols and feature set recommendations are only applicable to similar datasets (non-contrast-enhanced NSCLC CT images) using similar image features (Table 3.2).

One limitation of our study is that all semi-automated contouring was done by a single individual (the primary author). Ideally, multiple individuals could contour each GTV and operator sensitivity could be studied. Alternatively, one of several automated segmentation techniques could be used to possibly increase reproducibility (73). In the future it would also be useful to study the reproducibility of additional image features and the effect of convolution kernels. Another limitation is that all of the machines used in our study are GE machines. Ideally, multiple machines from various manufacturers should be tested to see if there are any systematic or random feature value variations between them. Finally, a key point to note is that our study only assessed intra-machine reproducibility. Multi-machine

reproducible features were defined as features that have good intra-machine reproducibility on multiple machines. Because patients were not test-retest imaged on multiple scanners, we cannot directly assess inter-machine reproducibility (i.e. agreement of feature values between machines). For ethical reasons, this necessitates a phantom, and there are several good candidates for future work. Court *et al.* (74) created a model of a real lung tumor using rapid prototyping, placed it into an anthropomorphic phantom, and moved it to match recorded tumor motion trajectories. This phantom would be very useful to study how motion affects feature reproducibility across multiple machines. In a separate application of rapid prototyping, another group has developed a way to independently control the CT number of every voxel of a phantom (75). Unlike a solid phantom, this model could be used to study image feature fidelity in the presence of subtle changes in voxel intensity and indistinct tissue boundaries.

3.5 Conclusion

This study integrated multiple NSCLC test-retest CT datasets to identify informative, non-redundant image features with high intra-machine reproducibility on multiple machines. Image feature quality was best for average 4D-CT images, and the tumor periphery was found to play an important role in tumor discrimination. Further advanced phantom studies are needed to investigate inter-machine image feature reproducibility.

Chapter 4: Discussion

4.1 Summary and Conclusions

Chapter 2 showed that quantitative image feature models derived from existing pre-treatment CT images could successfully predict NSCLC tumor shrinkage, an indicator of treatment efficacy and future survival.

Chapter 3 showed that image feature reproducibility and redundancy depended on both the CT machine and the CT image type. For each of the image types (end-exhale 4D-CT, average 4D-CT, and helical breath-hold 3D-CT) multiple NSCLC test-retest CT datasets were integrated to identify informative, non-redundant image features with high intra-machine reproducibility on multiple machines. Pruning of low intensity voxels showed that the tumor periphery plays an important role in tumor discrimination. Compared to end-exhale 4D-CT and breath-hold 3D-CT, average 4D-CT derived image features showed superior multi-machine reproducibility and are the best candidates for clinical correlation.

4.2 Evaluation of Hypotheses and Specific Aims

Specific aims 1 and 2 were successfully completed by the work presented in Chapter 2 and resulted in the confirmation of the first hypothesis: CT image features extracted from pre-treatment NSCLC tumors can be used to predict tumor shrinkage in response to therapy.

Specific aims 3 and 4 were successfully completed by the work presented in Chapter 3 and resulted in the confirmation of the second hypothesis: various selection metrics can identify a small subset of “ideal” image features for each of several different CT image types.

4.3 Future Research and Applications

Chapter 2 supports the findings of (41) and (46) which showed that quantitative image features could be used to make clinically relevant predictions for NSCLC patients. As our survival outcome data is currently unavailable, we are unable to develop a binary classifier to predict survival at present. However, in the future when this data is available, we could develop such a model. Since our dataset had more uniform treatment and imaging,

comparing its AUC to their findings could indicate how important treatment uniformity and imaging uniformity are for accurate radiomics survival prediction. This could have important applications when planning future prospective radiomics studies.

A limitation of our study in Chapter 3 is that a test-retest patient is only scanned on one machine. Therefore, we only have information to assess intra-machine reproducibility. To simultaneously assess both intra- and inter-machine reproducibility, a patient needs to be imaged at least twice on two machines (i.e. four total scans). This is not ethically justifiable, so a phantom is required. To examine how tumor motion affects inter-machine reproducibility, a 3D printed moveable phantom could be used (74). To how much fine tumor detail affects inter-machine reproducibility, another 3D printed phantom with finely controlled voxel CT numbers could be used (75). This later approach is a better simulation of tumor visual complexity, but it does not include tumor motion. One possible solution to this is to print test-retest image pairs directly into a CT number accurate phantom. By scanning this phantom on multiple machines, human test-retest scans on multiple machines could be simulated. Additionally, since the printed images already have motion-induced changes, this static phantom would allow for the simultaneous investigation of fine tumor detail and motion. This could be applied in future radiomics studies to find image feature sets that simultaneously optimize intra- and inter-machine reproducibility.

References

1. Siegel, R., D. Naishadham, and A. Jemal. Cancer statistics, 2012. CA Cancer J Clin 62:10-29.
2. Jemal, A., R. Siegel, E. Ward, Y. Hao, J. Xu, T. Murray, and M. J. Thun. 2008. Cancer statistics, 2008. CA Cancer J Clin 58:71-96.
3. Molina, J. R., P. Yang, S. D. Cassivi, S. E. Schild, and A. A. Adjei. 2008. Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship. Mayo Clin Proc 83:584-594.
4. Weiss, W. 1997. Cigarette smoking and lung cancer trends. A light at the end of the tunnel? Chest 111:1414-1416.
5. Zhang, H., and B. Cai. 2003. The impact of tobacco on lung health in China. Respirology 8:17-21.
6. Chansky, K., J. P. Sculier, J. J. Crowley, D. Giroux, J. Van Meerbeeck, and P. Goldstraw. 2009. The International Association for the Study of Lung Cancer Staging Project: prognostic factors and pathologic TNM stage in surgically managed non-small cell lung cancer. J Thorac Oncol 4:792-801.
7. Valk, P. E., T. R. Pounds, D. M. Hopkins, M. K. Haseman, G. A. Hofer, H. B. Greiss, R. W. Myers, and C. L. Lutrin. 1995. Staging non-small cell lung cancer by whole-body positron emission tomographic imaging. Ann Thorac Surg 60:1573-1581; discussion 1581-1572.
8. Poullis, M., J. McShane, M. Shaw, S. Woolley, M. Shackcloth, R. Page, and N. Mediratta. 2012. Lung cancer staging: a physiological update. Interact Cardiovasc Thorac Surg 14:743-749.
9. Guan, X., M. Yin, Q. Wei, H. Zhao, Z. Liu, L. E. Wang, X. Yuan, M. S. O'Reilly, R. Komaki, and Z. Liao. 2010. Genotypes and haplotypes of the VEGF gene and survival in locally advanced non-small cell lung cancer patients treated with chemoradiotherapy. BMC Cancer 10:431.

10. Xu, T., Q. Wei, J. L. Lopez Guerra, L. E. Wang, Z. Liu, D. Gomez, M. O'Reilly, S. H. Lin, Y. Zhuang, L. B. Levy, R. Mohan, H. Zhou, and Z. Liao. 2012. HSPB1 Gene Polymorphisms Predict Risk of Mortality for US Patients After Radio(chemo)therapy for Non-Small Cell Lung Cancer. Int J Radiat Oncol Biol Phys 84:e229-235.

11. Yin, M., Z. Liao, Y. J. Huang, Z. Liu, X. Yuan, D. Gomez, L. E. Wang, and Q. Wei. 2011. Polymorphisms of homologous recombination genes and clinical outcomes of non-small cell lung cancer patients treated with definitive radiotherapy. PLoS One 6:e20055.

12. Rinewalt, D., D. D. Shersher, S. Daly, C. Fhied, S. Basu, B. Mahon, E. Hong, G. Chmielewski, M. J. Liptay, and J. A. Borgia. 2012. Development of a serum biomarker panel predicting recurrence in stage I non-small cell lung cancer patients. J Thorac Cardiovasc Surg.

13. Chung, C. H., P. S. Bernard, and C. M. Perou. 2002. Molecular portraits and the family tree of cancer. Nat Genet 32 Suppl:533-540.

14. Segal, E., N. Friedman, N. Kaminski, A. Regev, and D. Koller. 2005. From signatures to models: understanding cancer using microarrays. Nat Genet 37 Suppl:S38-45.

15. Chen, X., S. T. Cheung, S. So, S. T. Fan, C. Barry, J. Higgins, K. M. Lai, J. Ji, S. Dudoit, I. O. Ng, M. Van De Rijn, D. Botstein, and P. O. Brown. 2002. Gene expression patterns in human liver cancers. Mol Biol Cell 13:1929-1939.

16. Subramanian, J., and R. Simon. Gene expression-based prognostic signatures in lung cancer: ready for clinical use? J Natl Cancer Inst 102:464-474.

17. Lambin, P., E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. G. van Stiphout, P. Granton, C. M. Zegers, R. Gillies, R. Boellard, A. Dekker, and H. J. Aerts. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 48:441-446.

18. Lambin, P., S. F. Petit, H. J. Aerts, W. J. van Elmpt, C. J. Oberije, M. H. Starmans, R. G. van Stiphout, G. A. van Dongen, K. Muylle, P. Flamen, A. L. Dekker, and D. De

Ruysscher. The ESTRO Breur Lecture 2009. From population to voxel-based radiotherapy: exploiting intra-tumour and intra-organ heterogeneity for advanced treatment of non-small cell lung cancer. Radiother Oncol 96:145-152.

19. Rubin, D. L. 2008. Creating and curating a terminology for radiology: ontology modeling and analysis. J Digit Imaging 21:355-362.
20. Opulencia, P., D. S. Channin, D. S. Raicu, and J. D. Furst. Mapping LIDC, RadLex, and lung nodule image features. J Digit Imaging 24:256-270.
21. Burton, A. 2007. REGIST: right time to renovate? Eur J Cancer 43:1642.
22. Choi, H. 2008. Response evaluation of gastrointestinal stromal tumors. Oncologist 13 Suppl 2:4-7.
23. Kumar, V., Y. Gu, S. Basu, A. Berglund, S. A. Eschrich, M. B. Schabath, K. Forster, H. J. Aerts, A. Dekker, D. Fenstermacher, D. B. Goldgof, L. O. Hall, P. Lambin, Y. Balagurunathan, R. A. Gatenby, and R. J. Gillies. Radiomics: the process and the challenges. Magn Reson Imaging 30:1234-1248.
24. Davnall, F., C. S. Yip, G. Ljungqvist, M. Selmi, F. Ng, B. Sanghera, B. Ganeshan, K. A. Miles, G. J. Cook, and V. Goh. Assessment of tumor heterogeneity: an emerging imaging tool for clinical practice? Insights Imaging 3:573-589.
25. Haralick, R. M., Shanmuga.K, and I. Dinstein. 1973. Textural Features for Image Classification. Ieee Transactions on Systems Man and Cybernetics Smc3:610-621.
26. Galloway, M. M. 1975. Texture analysis using gray level run lengths. Computer Graphics and Image Processing 4:172-179.
27. Jackson, A., J. P. O'Connor, G. J. Parker, and G. C. Jayson. 2007. Imaging tumor vascular heterogeneity and angiogenesis using dynamic contrast-enhanced magnetic resonance imaging. Clin Cancer Res 13:3449-3459.
28. Rose, C. J., S. J. Mills, J. P. O'Connor, G. A. Buonaccorsi, C. Roberts, Y. Watson, S. Cheung, S. Zhao, B. Whitcher, A. Jackson, and G. J. Parker. 2009. Quantifying

spatial heterogeneity in dynamic contrast-enhanced MRI parameter maps. Magn Reson Med 62:488-499.

29. Gibbs, P., and L. W. Turnbull. 2003. Textural analysis of contrast-enhanced MR images of the breast. Magn Reson Med 50:92-98.

30. Canuto, H. C., C. McLachlan, M. I. Kettunen, M. Velic, A. S. Krishnan, A. A. Neves, M. de Backer, D. E. Hu, M. P. Hobson, and K. M. Brindle. 2009. Characterization of image heterogeneity using 2D Minkowski functionals increases the sensitivity of detection of a targeted MRI contrast agent. Magn Reson Med 61:1218-1224.

31. Segal, E., C. B. Sirlin, C. Ooi, A. S. Adler, J. Gollub, X. Chen, B. K. Chan, G. R. Matcuk, C. T. Barry, H. Y. Chang, and M. D. Kuo. 2007. Decoding global gene expression programs in liver cancer by noninvasive imaging. Nat Biotechnol 25:675-680.

32. Diehn, M., C. Nardini, D. S. Wang, S. McGovern, M. Jayaraman, Y. Liang, K. Aldape, S. Cha, and M. D. Kuo. 2008. Identification of noninvasive imaging surrogates for brain tumor gene-expression modules. Proc Natl Acad Sci U S A 105:5213-5218.

33. Ganeshan, B., O. Strukowska, K. Skogen, R. Young, C. Chatwin, and K. Miles. Heterogeneity of focal breast lesions and surrounding tissue assessed by mammographic texture analysis: preliminary evidence of an association with tumor invasion and estrogen receptor status. Front Oncol 1:33.

34. Skogen, K., B. Ganeshan, C. Good, G. Critchley, and K. Miles. Measurements of heterogeneity in gliomas on computed tomography relationship to tumour grade. Journal of Neuro-Oncology 111:213-219.

35. Ng, F., B. Ganeshan, R. Kozarski, K. A. Miles, and V. Goh. Assessment of primary colorectal cancer heterogeneity by using whole-tumor texture analysis: contrast-enhanced CT texture as a biomarker of 5-year survival. Radiology 266:177-184.

36. Basu, S., L. O. Hall, D. B. Goldgof, G. Yuhua, V. Kumar, C. Jung, R. J. Gillies, and R. A. Gatenby. Developing a classifier model for lung tumors in CT-scan images. In

Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on. 1306-1312.

37. Kido, S., K. Kuriyama, M. Higashiyama, T. Kasugai, and C. Kuroda. 2002. Fractal analysis of small peripheral pulmonary nodules in thin-section CT: evaluation of the lung-nodule interfaces. J Comput Assist Tomogr 26:573-578.

38. Al-Kadi, O. S., and D. Watson. 2008. Texture analysis of aggressive and nonaggressive lung tumor CE CT images. IEEE Trans Biomed Eng 55:1822-1830.

39. Ganeshan, B., S. Abaleke, R. C. Young, C. R. Chatwin, and K. A. Miles. 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.

40. Ganeshan, B., V. Goh, H. C. Mandeville, Q. S. Ng, P. J. Hoskin, and K. A. Miles. Non-Small Cell Lung Cancer: Histopathologic Correlates for Texture Parameters at CT. Radiology 266:326-336.

41. Basu, S. 2012. 'Developing Predictive Models for Lung Tumor Analysis'. MS thesis, University of South Florida.

42. ctte, C. V. t. 2009. Profile: CT Lung Nodule Volume Measurement for Primary/Regional Nodes and Metastatic Sites. QIBA Web site.

43. Ganeshan, B., K. Burnand, R. Young, C. Chatwin, and K. Miles. Dynamic contrast-enhanced texture analysis of the liver: initial assessment in colorectal cancer. Invest Radiol 46:160-168.

44. Armato, S. G., 3rd, C. R. Meyer, M. F. McNitt-Gray, G. McLennan, A. P. Reeves, B. Y. Croft, and L. P. Clarke. 2008. The Reference Image Database to Evaluate Response to therapy in lung cancer (RIDER) project: a resource for the development of change-analysis software. Clin Pharmacol Ther 84:448-456.

45. Lin, L. I. K. 1989. A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics 45:255-268.

46. Aerts, H., E. Rios-Velazquez, R. Leijenaar, S. Carvalho, and P. Lambin. 2012. Radiomics: Extracting Advanced Features from Medical Imaging. Radiotherapy and Oncology 103:S70-S71.

47. Gillies, R. J., and R. A. Gatenby. 2012. QIN Progress Report: Radiomics of NSCLC. Moffitt Cancer Center, Tampa, FL.

48. Lambin, P., E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. G. van Stiphout, P. Granton, C. M. Zegers, R. Gillies, R. Boellard, A. Dekker, and H. J. Aerts. 2012. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 48:441-446.

49. Marusyk, A., and K. Polyak. 2010. Tumor heterogeneity: causes and consequences. Biochim Biophys Acta 1805:105-117.

50. Therasse, P., S. G. Arbuck, E. A. Eisenhauer, J. Wanders, R. S. Kaplan, L. Rubinstein, J. Verweij, M. Van Glabbeke, A. T. van Oosterom, M. C. Christian, and S. G. Gwyther. 2000. New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada. J Natl Cancer Inst 92:205-216.

51. Jaffe, C. C. 2006. Measures of response: RECIST, WHO, and new alternatives. J Clin Oncol 24:3245-3251.

52. Kumar, V., Y. Gu, K. Jongphil, R. A. Gatenby, and R. J. Gillies. 2012. Test Retest Reproducibility of Image Features Extracted from CT Images of Lung Tumors. In Radiology.

53. Court, L. E., and L. Dong. 2003. Automatic registration of the prostate for computed-tomography-guided radiotherapy. Medical Physics 30:2750-2757.

54. Zhang, L., L. Dong, L. Court, H. Wang, M. Gillin, and R. Mohan. 2005. Validation of CT-Assisted targeting (CAT) software for soft tissue and bony target localization. Medical Physics 32:2106-2106.

55. Chen, Y., L. Zhang, Z. Liao, R. Komaki, J. Cox, P. Balter, R. Mohan, and L. Dong. 2011. Comparison of Tumor Shrinkage in Proton and Photon Therapy of Lung Cancer. Medical Physics 38.

56. Jolliffe, I. T. 1982. A Note on the Use of Principal Components in Regression. Applied Statistics-Journal of the Royal Statistical Society Series C 31:300-303.

57. Kirkpatrick, S., C. D. Gelatt, Jr., and M. P. Vecchi. 1983. Optimization by simulated annealing. Science 220:671-680.

58. Eisenhauer, E. A., P. Therasse, J. Bogaerts, L. H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe, and J. Verweij. 2009. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer 45:228-247.

59. Farrar, D. E. 1982. Citation Classic - Multicollinearity in Regression-Analysis - the Problem Revisited. Current Contents/Social & Behavioral Sciences:22-22.

60. Hadi, A. S., and R. F. Ling. 1998. Some Cautionary Notes on the Use of Principal Components Regression. The American Statistician 52:15-19.

61. Lindell, R. M., T. E. Hartman, S. J. Swensen, J. R. Jett, D. E. Midthun, H. D. Tazelaar, and J. N. Mandrekar. 2007. Five-year lung cancer screening experience: CT appearance, growth rate, location, and histologic features of 61 lung cancers. Radiology 242:555-562.

62. Gimenez, A., T. Franquet, R. Prats, P. Estrada, J. Villalba, and S. Bague. 2002. Unusual primary lung tumors: a radiologic-pathologic overview. Radiographics 22:601-619.

63. Dehing-Oberije, C., H. Aerts, S. Yu, D. De Ruysscher, P. Menheere, M. Hilvo, H. van der Weide, B. Rao, and P. Lambin. 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 NCT00572325). Int J Radiat Oncol Biol Phys 81:360-368.

64. Starkschall, G., P. Balter, K. Britton, M. F. McAleer, J. D. Cox, and R. Mohan. Interfractional Reproducibility of Lung Tumor Location Using Various Methods of Respiratory Motion Mitigation. International Journal of Radiation Oncology*Biology*Physics 79:596-601.

65. Zhao, B., L. P. James, C. S. Moskowitz, P. Guo, M. S. Ginsberg, R. A. Lefkowitz, Y. Qin, G. J. Riely, M. G. Kris, and L. H. Schwartz. 2009. Evaluating variability in tumor measurements from same-day repeat CT scans of patients with non-small cell lung cancer. Radiology 252:263-272.

66. McBride, G. 2005. A proposal for strength-of-agreement criteria for Linâ€™s concordance correlation coefficient. NIWA Client Report: HAM2005-062. Report to Ministry of Health.

67. Dice, L. R. 1945. Measures of the Amount of Ecologic Association Between Species. Ecology 26:297-302.

68. Jaccard, P. 1901. Distribution de la flore alpine dans le bassin des Dranses et dans quelques rÃgions voisines. Bulletin de la SociÃtÃ Vaudoise des Sciences Naturelles 37:241-272.

69. Eisen, M. B., P. T. Spellman, P. O. Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95:14863-14868.

70. Watkins, W. T., R. Li, J. Lewis, J. C. Park, A. Sandhu, S. B. Jiang, and W. Y. Song. Patient-specific motion artifacts in 4DCT. Medical Physics 37:2855-2861.

71. Keall, P. J., G. Starkschall, H. Shukla, K. M. Forster, V. Ortiz, C. W. Stevens, S. S. Vedam, R. George, T. Guerrero, and R. Mohan. 2004. Acquiring 4D thoracic CT scans using a multislice helical method. Physics in Medicine and Biology 49:2053.

72. Wolthaus, J. W., C. Schneider, J. J. Sonke, M. van Herk, J. S. Belderbos, M. M. Rossi, J. V. Lebesque, and E. M. Damen. 2006. Mid-ventilation CT scan construction from four-dimensional respiration-correlated CT scans for radiotherapy planning of lung cancer patients. International journal of radiation oncology, biology, physics 65:1560-1571.

73. Sharma, N., and L. M. Aggarwal. Automated medical image segmentation techniques. J Med Phys 35:3-14.

74. Court, L. E., J. Seco, X. Q. Lu, K. Ebe, C. Mayo, D. Ionascu, B. Winey, N. Giakoumakis, M. Aristophanous, R. Berbeco, J. Rottman, M. Bogdanov, D. Schofield, and T. Lingos. Use of a realistic breathing lung phantom to evaluate dose delivery errors. Med Phys 37:5850-5857.

75. Yoo, T. S., T. Hamilton, D. E. Hurt, J. Caban, D. Liao, and D. T. Chen. Toward quantitative X-ray CT phantoms of metastatic tumors using rapid prototyping technology. In Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on. 1770-1773.

Vita

Luke Aaron Hunter was born in Oklahoma City, Oklahoma on January 19, 1986, the son of Scott and Carla Hunter. After completing high school at Longview High School, Longview, Texas in 2004, he entered Texas A&M University in College Station, Texas. He spent his final year of undergraduate education as a Visiting Student at Harvard University, Cambridge, Massachusetts. In 2008, he graduated as Valedictorian with University Honors from Texas A&M University and received a Bachelor of Science in Physics, a Bachelor of Science in Biochemistry, Minors in Chemistry and Computer Science, and an Honors Minor in Mathematics. Afterwards, he served for one year as an emergency room scribe at the St. Joseph Regional Health Center in Bryan, Texas. In 2009, he entered the M.D. Program at the Baylor College of Medicine in Houston, Texas. In 2010, he took a leave of absence to pursue a Master's of Science in Medical Physics through The University of Texas Health Science Center at Houston Graduate School of Biomedical Sciences. In 2013, he will re-enter the Baylor College of Medicine M.D. Program. Luke is married to Rachel A. Hunter, and they enjoy traveling, scuba diving, and hiking.