

12-2015

Investigation Of Quatitative Image Features From Pretreatment Ct And Fdg-Pet Scans In Stage Iii Nsclc Patients Undergoing Defintive Radiation Therapy

David Fried

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



Part of the [Medical Biophysics Commons](#), and the [Oncology Commons](#)

Recommended Citation

Fried, David, "Investigation Of Quatitative Image Features From Pretreatment Ct And Fdg-Pet Scans In Stage Iii Nsclc Patients Undergoing Defintive Radiation Therapy" (2015). *Dissertations and Theses (Open Access)*. 641.

https://digitalcommons.library.tmc.edu/utgsbs_dissertations/641

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

INVESTIGATION OF QUATITATIVE IMAGE FEATURES FROM PRETREATMENT CT
AND FDG-PET SCANS IN STAGE III NSCLC PATIENTS UNDERGOING DEFINITIVE
RADIATION THERAPY

by

David Vincent Fried, B.S.

APPROVED:

Laurence Court, Ph.D.
Advisory Professor

Zhongxing Liao, M.D.

Geoffrey Ibbott, Ph.D.

Osama Mawlawi, Ph.D.

Shouhao Zhou, Ph.D.

APPROVED:

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

INVESTIGATION OF QUATITATIVE IMAGE FEATURES FROM PRETREATMENT CT
AND FDG-PET SCANS IN STAGE III NSCLC PATIENTS UNDERGOING DEFINITIVE
RADIATION THERAPY

A

DISSERTATION

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

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

by

David Vincent Fried, B.S.
Houston, Texas

December 2015

Dedicated to my wife, Erica. I cannot wait to see what life has in store for us next.

Acknowledgements

I would like to start by thanking my advisor, Dr. Laurence Court, for his never-ending optimism and support of my work. His enthusiasm and willingness to let the project be free to develop along the way was essential to its success. Dr. Court's ability to be involved in numerous graduate students' projects simultaneously is a testament to his skills as a researcher and mentor.

I would also like to thank Dr. Susan Tucker and Dr. Shouao Zhou for their patience and help with matters of a statistical/biostatistical nature. I would also like to thank Dr. Zhongxing Liao, Dr. Geoffrey Ibbott, and Dr. Osama Mawlawi for their encouragement and guidance in steering my research and work performed on various abstracts and manuscripts.

I would like to thank Dr. Lifei Zhang and Luke Hunter for their massive contributions to my work, particularly their development of IBEX.

I would like to thank my classmates and research group-mates for all their help along the way. I would particularly like to thank Xenia Favè for putting up with me sitting next to her in the office over the past few years. Being able to talk with her whenever I had a random question made a tremendous difference.

Most importantly, I would like to thank my family and friends. Their unwavering support really made this possible.

INVESTIGATION OF QUANTITATIVE IMAGE FEATURES FROM PRETREATMENT CT AND FDG-PET SCANS IN STAGE III NSCLC PATIENTS UNDERGOING DEFINITIVE RADIATION THERAPY

David Vincent Fried, Ph.D.

Advisory Professor: Laurence Court, Ph.D.

The purpose of this work was to determine if quantitative image features (QIFs) extracted from computed tomography (CT) and fluorodeoxyglucose (FDG) positron emission tomography (PET) could provide prognostic information to improve outcome models. Our goal for this work was to determine if it may one day be feasible to incorporate QIFs into personalized cancer care. QIFs were used to quantitatively characterize patient disease as seen on imaging. A leave-one-out cross-validation procedure was used to assess the prognostic ability of QIFs extracted from CT and PET in addition to conventional prognostic factors (CPFs). QIFs were found to improve model fit for overall survival in contrast enhanced CT (CE-CT) ($p = 0.027$) and FDG-PET ($p = 0.007$).

Correlations/associations were observed between QIFs from CE-CT, FDG-PET, and CPFs. However, our results indicate that while correlations/associations exist, QIFs provided *additional* prognostic information. QIFs from FDG-PET improved models using CPFs including GTV in terms of patient stratification, c-index, and log-likelihood more than QIFs from CE-CT alone. Various studies were performed assessing the reproducibility of FDG-PET based QIFs and found that reconstruction methods certainly impact the obtained QIF values. However, features maintain a reasonable reproducibility (mean CCC = 0.78) that may be improved when using similar reconstructions (e.g., 3D OSEM) (CCC = 0.93). The two FDG-PET features found to be prognostic were also able to isolate sub-cohorts of patients that demonstrated survival differences based on radiation dose.

QIFs were found to provide additional prognostic information beyond that found from CPFs. Initial evidence suggests that the examined FDG-PET based QIFs may have utility across cohorts and could potentially determine which patients may benefit from dose escalation.

Table of Contents

| | |
|--|----|
| Dedication | 3 |
| Acknowledgements | 4 |
| Table of Contents | 6 |
| List of Figures | 10 |
| List of Tables | 14 |
| Chapter 1 Introduction..... | 16 |
| Chapter 2 Principal Hypothesis and Specific Aims..... | 20 |
| Principal Hypothesis: | 20 |
| Specific Aim 1: Analysis of CT-based Quantitative Image Features..... | 20 |
| Specific Aim 2: Analysis of FDG-PET-based Quantitative Image Features | 20 |
| Specific Aim 3: Assess relationships between CT-based quantitative image features, PET-based quantitative image features, conventional features, and morphologic features | 21 |
| Specific Aim 4: Potential use of FDG-PET-based quantitative image features..... | 21 |
| Chapter 3 Methodology..... | 22 |
| 3.1 Conventional Prognostic Factors..... | 22 |
| 3.2 Patient Cohorts | 23 |
| Cohort 1: 91 Patients with Pretreatment T_{avg} , T_{50} , and CE-CT | 23 |
| Cohort 2: 249 Patients with Pretreatment CE-CT | 25 |
| Cohort 3: 195 Patients with Pretreatment PET..... | 27 |
| Cohort 4: 78 Patients with Pretreatment PET and Contrast-Enhanced CT | 29 |
| Cohort 5: 24 Patients with “Large” Tumors on PET | 29 |
| Cohort 6: 53 Patients with “Pseudo” Test/Retest PET | 29 |

| | |
|---|----|
| 3.3 Quantitative Image Features..... | 30 |
| 3.3.1 Histogram Features..... | 30 |
| 3.3.2 Co-Occurrence Matrix Features | 31 |
| 3.3.3 Nearest Gray Tone Difference Features | 35 |
| 3.3.4 Laplacian of Gaussian Filtration Features | 36 |
| 3.3.5 Contrast Enhanced CT Auto-segmentation of Morphologic Characteristics | 38 |
| 3.3.6 PET Necrosis Auto-segmentation | 40 |
| 3.4 Region of Interest Contouring on CT | 42 |
| 3.5 Region of Interest Contouring on PET | 42 |
| 3.6 Assessment of QIF Reproducibility using Phantom and Patient Data | 44 |
| 3.7 Statistical Methods | 47 |
| 3.7.1 Use of Cross-Validation for Assessment of Prognostic Value..... | 47 |
| 3.7.2 Permutation Test and Impact of Feature Reproducibility on Predictions (Cohort 1) | 49 |
| 3.7.3 Concordance Index at Multiple Time Points | 50 |
| 3.7.4 Analysis of Relationship between Quantitative Image Features, Conventional Features and Morphologic Characteristics | 50 |
| 3.7.5 K-Means Clustering of Predictions | 51 |
| 3.7.6 Concordance Correlation Coefficient..... | 52 |
| 3.7.7 Analysis of PET Tumor Resampling (Cohort 5) | 52 |
| 3.7.8 Sub-cohorts Based on FDG-PET QIFs to Determine Impact of Dose Escalation..... | 53 |
| Chapter 4 Results..... | 54 |
| 4.1 Results of Specific Aim 1: Analysis of CT-based Quantitative Image Features..... | 54 |

| | |
|--|----|
| 4.1.1 Results for Project 1.1: Quantify the impact of adding CT-based quantitative image features to outcome models containing only CPFs including and excluding GTV | 54 |
| 4.1.2 Results for Project 1.2: Quantify the reproducibility of FDG-PET-based quantitative image features using “pseudo” test-retest scans | 64 |
| 4.1.3 Results for Project 1.3: Quantify the prognostic value of adding CE-CT-based quantitative image features to outcome models containing only CPFs | 66 |
| 4.2 Results of Specific Aim 2: Analysis of FDG-PET-based Quantitative Image Features | 71 |
| 4.2.1 Results for Project 2.1: Quantify the impact of adding FDG-PET-based quantitative image features to outcome models containing only CPFs | 71 |
| 4.2.2 Results for Project 2.2: Quantify the reproducibility of FDG-PET-based quantitative image features using “pseudo” test-retest scans | 76 |
| 4.2.3 Results for Project 2.3: Quantify the reproducibility of FDG-PET-based quantitative image features using retrospective reconstructions of phantom and patient data | 78 |
| 4.3 Results of Specific Aim 3: Assess relationships between CT-based quantitative image features, PET-based quantitative image features, conventional features, and morphologic features | 80 |
| 4.3.1 Results for Project 3.1: Quantify correlations between prognostic FDG-PET-based and CECT-based quantitative image features | 80 |
| 4.3.2 Results for Project 3.2: Quantify if relationships exist between CE-CT-based and FDG-PET-based quantitative image features with tumor volume and TNM staging | 83 |
| 4.3.3 Results for Project 3.3: Quantify if there are correlations between FDG-PET-based quantitative image features, CECT-based quantitative image features, and morphologic characteristics (vessels, necrosis, air cavities, etc.) | 87 |
| 4.4 Results of Specific Aim 4: Potential use of FDG-PET-based quantitative image features | 93 |

| | |
|--|-----|
| 4.4.1 Results for Project 4.1: Assess whether significant PET-based quantitative image features relate to a difference in patient survival for those treated with an escalated radiation dose..... | 93 |
| Chapter 5 Discussion..... | 101 |
| Discussion Specific Aim 1 | 101 |
| Discussion Specific Aim 2 | 106 |
| Discussion Specific Aim 3 | 112 |
| Discussion Specific Aim 4 | 117 |
| Discussion Overall | 120 |
| References | 123 |
| Appendix A: Matlab Code for CE-CT Autosegmentation | 129 |
| Appendix B: Matlab Code for CE-CT Necrosis Identification | 132 |
| Appendix C: R Code for Cross-Validation Technique..... | 134 |
| Appendix D: Relationship of Cardiothoracic Dosimetry with Disease Solidity | 138 |
| Appendix E: Comparison of FDG-PET Delineation Methods | 143 |
| Appendix F: Sequential FDG-PET Analysis..... | 145 |
| Appendix G: Assessment of Volumetric Stability | 149 |
| VITA | 152 |

List of Figures

| | |
|--|----|
| Figure 1. Laplacian of Gaussian Example. Original Tumor (left) and Results of LOG Filtration (right) ($\sigma = 1$) | 37 |
| Figure 2. CE-CT Results of Auto-segmentation of Air (gray), Necrosis (red), Tissue (green), and Vessels (blue) | 40 |
| Figure 3. PET Results of Auto-segmentation of Necrosis (blue) | 41 |
| Figure 4. (A) Original FDG-PET Contour/Image (B) Analyzed Voxels Using a 50% Cutoff | 44 |
| Figure 5. Example Range Image of NEMA IEC Phantom..... | 46 |
| Figure 6. Overall Survival Comparing High Risk versus Low Risk Patients Using Models Incorporating CPFs Excluding GTV. | 55 |
| Figure 7. Overall Survival Comparing High Risk versus Low Risk Patients Using Models Incorporating CPFs Including GTV. | 56 |
| Figure 8. Overall Survival Comparing High Risk versus Low Risk Patients Using Models Incorporating CPFs Including GTV and CT-Based QIFs. | 56 |
| Figure 9. Local-Regional Control Comparing High Risk versus Low Risk Patients Using Models Incorporating CPFs Excluding GTV. | 57 |
| Figure 10. Local-Regional Control Comparing High Risk versus Low Risk Patients Using Models Incorporating CPFs Including GTV. | 57 |
| Figure 11. Local-Regional Control Comparing High Risk versus Low Risk Patients Using Models Incorporating CPFs Including GTV And CT-Based QIFs. | 58 |
| Figure 12. Freedom from Distant Metastases Comparing High Risk versus Low Risk Patients Using Models Incorporating CPFs Excluding GTV. | 58 |
| Figure 13. Freedom from Distant Metastases Comparing High Risk versus Low Risk Patients Using Models Incorporating CPFs Including GTV. | 59 |

| | |
|--|----|
| Figure 14. Freedom from Distant Metastases Comparing High Risk versus Low Risk Patients Using Models Incorporating CPFs Including GTV And CT-Based Qifs. | 59 |
| Figure 15. Concordance Indices for Overall Survival Predictions Using Minimum Outcome Differences of 6, 12, 18, and 24 Months. | 61 |
| Figure 16. Concordance Indices for Local-Regional Control Predictions Using Minimum Outcome Differences Of 6, 12, 18, And 24 Months. | 61 |
| Figure 17. Concordance Indices for Freedom from Distant Metastases Predictions Using Minimum Outcome Differences of 6, 12, 18, And 24 Months. | 62 |
| Figure 18. Example from Single Simulation of the Impact of Texture Feature Reproducibility on FFDM Estimates. Outcome Prediction from Original Model (X-Axis) Compared to Prediction Incorporation of the Variation in QIFs From Test/Retest Scans (Y-Axis)..... | 65 |
| Figure 19. Overall Survival Comparing High Risk, Medium Risk, and Low Risk Patients Using Models Incorporating CPFs Excluding GTV | 67 |
| Figure 20. Overall Survival Comparing High Risk, Medium Risk, and Low Risk Patients Using Models Incorporating CPFs Including GTV | 67 |
| Figure 21. Overall Survival Comparing High Risk, Medium Risk, and Low Risk Patients Using Models Incorporating CPFs Including GTV And CE-CT Based QIFs | 68 |
| Figure 22. Concordance Indices for Overall Survival in Cohort 2 (CE-CT QIFs Only) Using Minimum Outcome Differences of 6, 12, 18, and 24 Months..... | 69 |
| Figure 23. Overall Survival Comparing Various Risk Groups (Defined Using K-Means Clustering from Low Risk to High Risk) Using Models Incorporating CPFs Excluding GTV. | 72 |
| Figure 24. Overall Survival Comparing Various Risk Groups (Defined Using K-Means Clustering from Low Risk to High Risk) Using Models Incorporating Cpfes Including GTV..... | 73 |
| Figure 25. Overall Survival Comparing Various Risk Groups (Defined Using K-Means Clustering from Low Risk to High Risk) Using Models Incorporating Cpfes Including GTV And QIFs. | 73 |

| | |
|--|----|
| Figure 26. Concordance Indices for Overall Survival in Cohort 3 (FDG-PET Based QIFs) Using Minimum Outcome Differences of 6, 12, 18, And 24 Months..... | 74 |
| Figure 27. Assessment of Correlations between Prognostic CE-CT QIFs (LOG_Average and Uniformity) and Prognostic FDG-PET QIFs (COM Energy, Solidity, and Uniformity) | 82 |
| Figure 28. Comparison of LOG_Average and Uniformity from CE-CT Versus Tumor Volume and Staging..... | 84 |
| Figure 29. Comparison of COM Energy, Uniformity, and Solidity from FDG-PET versus Tumor Volume and Staging | 86 |
| Figure 30. Comparison of LOG_Average and Uniformity on CE-CT versus The Presence/Absence of Various Tissue Types | 88 |
| Figure 31. Comparison of Necrosis Volumes Determined By FDG-PET Vs CE-CT..... | 89 |
| Figure 32. Comparison of Necrosis Percentage Determined By FDG-PET Vs CE-CT..... | 89 |
| Figure 33. Comparison of CE-CT Feature Value for Entire Tumor (X-Axis) with Feature Value Excluding a Particular Tissue Type or Types (Only Tissue: Excludes Air, Necrosis, and Vessels). (A) Comparison of LOG_Average Values from Tumors with Enhancing Vessels when the Vessels Are Present (X-Axis) or Excluded (Y-Axis). (B) Comparison of Uniformity Values from Tumors with Enhancing Vessels when the Vessels are Present (X-Axis) or Excluded (Y-Axis). (C) Comparison of LOG_Average Values from Tumors with Necrosis when the Necrosis Is Present (X-Axis) or Excluded (Y-Axis). (D) Comparison of Uniformity Values from Tumors with Necrosis when the Necrosis Is Present (X-Axis) or Excluded (Y-Axis). (E) Comparison of LOG_Average Values from Tumors with Cavitation when the Cavitation Is Present (X-Axis) or Excluded (Y-Axis). (F) Comparison of Uniformity Values from Tumors With Cavitation when the Cavitation Is Present (X-Axis) or Excluded (Y-Axis). | 91 |
| Figure 34. Stratification of Overall Survival and Progression-Free Survival by Dose Level in All Patients within Cohort 3 | 94 |

| | |
|---|-----|
| Figure 35. Log-Rank P-Values from Sub-Cohorts Based on High Values FDG-PET QIFs In Terms Of Overall Survival (A) and Progression-Free Survival (B) | 95 |
| Figure 36. Kaplan-Meier Plots Stratified by Dose Level for The Sub-Cohort with High Values Of FDG-PET QIFs In Terms of Overall Survival and Progression-Free Survival | 96 |
| Figure 37. Log-Rank P-Values from Sub-Cohorts Based on High Values of FDG-PET QIFs In Terms of Overall Survival(A) and Progression-Free Survival(B) | 97 |
| Figure 38. Kaplan-Meier Plots Stratified by Dose Level for the Sub-Cohort with Low Values of FDG-PET QIFs In Terms of Overall Survival and Progression Free Survival | 98 |
| Figure 39. GTV versus solidity in terms of mean lung dose | 138 |
| Figure 40. GTV versus solidity in terms of lung V20 | 139 |
| Figure 41. GTV versus solidity in terms of mean heart dose | 139 |
| Figure 42. Mean lung dose stratified by risk score..... | 140 |
| Figure 43. Lung V20 stratified by risk score..... | 141 |
| Figure 44. Mean heart dose stratified by risk score..... | 141 |
| Figure 45. Primary volume changes during treatment | 145 |
| Figure 46. SUVmax changes during treatment | 146 |
| Figure 47. SUVmean changes during treatment..... | 146 |
| Figure 48. Uniformity changes during treatment | 147 |
| Figure 49. COM Energy changes during treatment..... | 147 |
| Figure 50. Plots of original versus resampled entropy values and associated CCC values..... | 149 |

List of Tables

| | |
|--|-----|
| Table 1. Cohort 1 – Patient CPFs and Treatment Characteristics | 24 |
| Table 2. Cohort 2 - Patient CPFs and Treatment Characteristics | 26 |
| Table 3. Cohort 3 - Patient CPFs and Treatment Characteristics | 28 |
| Table 4. Parameters Used for Phantom and Patient Retrospective Reconstructions..... | 45 |
| Table 5. Extracted Quantitative Image Features for Cohort 1 | 55 |
| Table 6. Outcome Models for Covariate Combinations in Cohort 1..... | 63 |
| Table 7. Extracted Quantitative Image Features for Cohort 2..... | 66 |
| Table 8. Outcome Models for Covariate Combinations in Cohort 2..... | 70 |
| Table 9. Extracted Quantitative Image Features for Cohort 3 (pretreatment FDG-PET)..... | 71 |
| Table 10. Overall Survival Models for Covariate Combinations in Cohort 3..... | 75 |
| Table 11. CCC Values from “Pseudo” Test-Retest PET Scans | 77 |
| Table 12. Percent of Sphere Voxels with a Maximum Change in SUV <1 or < 2..... | 78 |
| Table 13. Change in QIF Values due to Variation in Reconstruction Parameters and Comparison to Variation in Cohort 3 Patient QIF Values | 79 |
| Table 14. Change in QIF Values from Patient Scans due to Variation in Reconstruction Parameters and Comparison to Variation in Patient QIF Values..... | 80 |
| Table 15. Correlations between PET and CE-CT Features | 81 |
| Table 16. Comparison of Conventional Prognostic Factors..... | 99 |
| Table 17. CE-CT Test/Retest Scan Information..... | 103 |
| Table 18. Summary of Results for Specific Aim 1 Projects..... | 104 |
| Table 19. Summary of Results for Specific Aim 2 Hypotheses | 110 |
| Table 20. Summary of Results for Specific Aim 3 Hypotheses | 116 |

| | |
|---|-----|
| Table 21. Summary of Results for Specific Aim 4 Hypotheses | 119 |
| Table 22. CCC Values for Comparison of Delineation Methodologies | 143 |
| Table 23. Summary of Changes in Features between Pre, Mid, and Post Treatment..... | 148 |
| Table 24. CCC values of features with respect to the resampled number of voxels | 150 |

Chapter 1 Introduction

Non-small cell lung cancer (NSCLC) results in more deaths than any other type of cancer in the United States.¹ AJCC TNM staging, which classifies patients as stage I through IV, is a commonly used tool that dictates patient prognosis and treatment.² Patients with stages I-III are viewed as potentially curative and receive definitive treatment. Early stage patients (stages I/II) can achieve a 5-year survival rate between 45-50% and are predominantly treated with surgical resection.³ Locally-advanced, non-metastatic (stage III) patients have a 5-year survival rate between 5-15%.³ These patients are predominantly treated with a combination of radiation therapy and chemotherapy (chemoradiotherapy). Stage III NSCLC is a particularly diverse cohort because patients can have varying primary tumor size/extent (T stages: 1 through 4) and nodal involvement (N stages: 0 through 3). This diversity of patients yields similarly diverse outcomes. Some patients succumb to their disease only a few months after diagnosis while others are able to do remarkably well and live 5 or more years post treatment.

Currently, the predominant factor in assessing prognosis and treatment is the patient's TNM stage. Further individualization is performed in practice based on other conventional prognostic factors (CPFs), such as tumor volume, histology, age, gender, performance status, and smoking history.⁴ However, the impact of these CPFs is based purely on the experience/opinion of the treating physician(s) and is not standardized. Furthermore, quantitative models have been shown to have the ability to outperform physicians when it comes to predicting a patient's outcome to treatment.⁵ While there is a body of literature regarding CPFs in stage III NSCLC, there are no standardized tools that allow physicians to individually predict patient outcome in order to give a more personalized prognosis or aid in treatment decision making. Outcome nomograms exist for various other forms of cancer via the Memorial Sloan Kettering Cancer Center (<http://www.mskcc.org/nomograms>). However, these nomograms are not routinely used in the clinic, have yielded a wide range of observed results, and only utilize relatively generic CPFs similar to TNM staging.

The concept of personalized cancer care recognizes that each cancer patient is unique. The needs, tolerances, and outcomes of patients can vary widely even if they receive the same treatment/care and are classified as similar based on CPFs. Therefore, a major goal in cancer medicine is to eventually tailor each patient's care specifically to that individual rather than to utilize population-based data when determining prognosis, appropriate follow-up intervals, or treatment.^{6,7}

Medical imaging is a source of potentially prognostic information that is routinely obtained, non-invasive, and specific to each patient. Imaging is already a primary tool for determining TNM stage and is currently performed as part of routine standard of care for patients with NSCLC. Additionally, there is a growing body of evidence suggesting that additional prognostic information can be ascertained from quantitatively analyzing a patient's tumor using quantitative image features (QIFs). QIFs are commonly based on disease histograms, co-occurrence matrices, nearest gray tone difference matrices, filtration-based features, and shape/volume based features.⁸⁻¹¹ These QIFs have been shown to have prognostic abilities in a variety of settings using pretreatment computed tomography (CT) and fluorodeoxyglucose (FDG) positron emission tomography (PET) scans.^{8, 10, 12-25} This process is referred to as "radiomics" since it was motivated by other high-throughput analyses methods, such as genomics, proteomics, etc. However, many radiomics studies relate QIFs purely to patient outcome and not any sort of genomic, proteomic, or biological endpoint.

Multiple publications using QIFs extracted from CT scans have shown relationships between tumor heterogeneity and patient outcome.^{8, 10, 14-17, 19, 20, 25-30} Relationships have been observed using both non-contrast enhanced (NCE) and contrast enhanced (CE) scans. Furthermore, associations have also been shown relating QIFs from CT to tumor histology, genetic variations, and glucose metabolism. The largest and most comprehensive of these studies was performed by Aerts et al.⁸ They found that a four-feature signature developed from a cohort of 422 patients had prognostic power when applied to an independent data set of 225 patients. They also were able to demonstrate associations between the four prognostic QIFs and tumor gene expression in a cohort of 83 patients. All of these cohorts consisted of patients with NSCLC of varying stages (i.e., stages I through IV).

Other literature regarding prognostic value of CT-based QIFs is largely composed of small, retrospective studies that frequently utilize re-substitution statistics or optimal cut-off methods for assessing prognostic value. Proper validation techniques are needed when analyzing the potential impact of CT-based QIFs. Ideally, bootstrapping or cross-validation techniques may be utilized if independent or external sources of data are not available. Furthermore, there is tremendous uncertainty in the literature regarding whether information from CT-based QIFs yields added predictive accuracy compared to CPFs, such as staging, disease volume, performance status, histology, etc.

Similar observations and pitfalls exist regarding FDG-PET-based QIFs. A significant body of literature exists regarding “standard” FDG-PET measures, such as SUV_{max} , SUV_{mean} , metabolic tumor volume, etc.^{31–34} However, significantly fewer publications address more complex QIFs examining disease heterogeneity and shape. The literature regarding more complex QIFs is composed predominantly of small, retrospective studies lacking proper validation and/or multivariate analyses examining the added benefit of QIFs to currently known CPFs. Nonetheless, existing publications suggests a potential relationship between FDG-PET-based QIFs and patient outcome in NSCLC.^{12, 21,}

35, 36

While there is compelling evidence that additional prognostic information can be extracted from quantitatively analyzing CT and FDG-PET, additional evaluations are needed to thoroughly investigate the potential of QIFs in these modalities and to address gaps in existing data. The goal of this work is to expand upon findings from existing publications regarding CT and FDG-PET QIFs in an effort establish a foundation for assessing whether or not imaging-based QIFs may one day be used as part of personalized cancer care. Retrospective cohorts will be generated and used to extract QIFs from patient imaging alongside patients CPF and outcomes from patient medical records. This data will be used to assess the prognostic value of QIFs and CPFs, variability/reproducibility of QIFs, and to quantify the relationship(s) between QIFs, CPFs, and physical tumor characteristics (e.g.,

necrosis, vessels, and cavitation). Exploratory analysis will evaluate the possible modification of treatment based on QIFs.

Chapter 2 Principal Hypothesis and Specific Aims

Principal Hypothesis:

The addition of quantitative image features from CT and PET scans to models using only conventional prognostic factors can improve patient outcome models.

Specific Aim 1: Analysis of CT-based Quantitative Image Features

Specific Aim 1 Hypothesis: The addition of CT-based quantitative image features will significantly improve outcome models compared to models using conventional prognostic factors

Project 1.1 Quantify the impact of adding CT-based quantitative image features to outcome models containing only CPFs including and excluding GTV

Project 1.2 Quantify the reproducibility of CT-based quantitative image features and its impact on outcome models

Project 1.3 Quantify the prognostic value of adding CE-CT-based quantitative image features to outcome models containing only CPFs

Specific Aim 2: Analysis of FDG-PET-based Quantitative Image Features

Specific Aim 2 Hypothesis: The addition of FDG-PET-based quantitative image features will significantly improve outcome models compared to models using conventional prognostic factors

Project 2.1 Quantify the impact of adding FDG-PET-based quantitative image features to outcome models containing only CPFs

Project 2.2 Quantify the reproducibility of FDG-PET-based quantitative image features using “pseudo” test-retest scans

Project 2.3 Quantify the reproducibility of FDG-PET-based quantitative image features using retrospective reconstructions of phantom and patient data

Specific Aim 3: Assess relationships between CT-based quantitative image features, PET-based quantitative image features, conventional features, and morphologic features

Specific Aim 3 Hypothesis: There will be significant relationships between some quantitative image features between modalities and with tumor volume, staging, and morphologic characteristics.

Project 3.1 Quantify correlations between prognostic FDG-PET-based and CECT-based quantitative image features

Project 3.2 Quantify if relationships exist between CE-CT-based and FDG-PET-based quantitative image features with tumor volume and TNM staging

Project 3.3 Quantify if there are correlations between FDG-PET-based quantitative image features, CECT-based quantitative image features, and morphologic characteristics (vessels, necrosis, air cavities, etc.)

Specific Aim 4: Potential use of FDG-PET-based quantitative image features

Specific Aim 4 Hypothesis: Significant FDG-PET-based based quantitative image features found in Specific Aim 2 will allow for identification of sub-cohorts that will demonstrate a significant stratification of patients based on radiation dose.

Project 4.1 Assess whether significant PET-based quantitative image features relate to a difference in patient survival for those treated with an escalated radiation dose

Chapter 3 Methodology

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

Fried DV, Mawlawi O, Zhang L, Fave X, Zhou S, Ibbott G, Liao Z, Court LE. Stage III non-small cell lung cancer: Prognostic value of FDG PET quantitative image features combined with clinical prognostic factors. *Radiology* doi: 10.1148/radiol.2015142920. Published online July 15, 2015. ©Radiological Society of North America.

Fried DV, Tucker SL, Zhou S, Liao Z, Mawlawi O, Ibbott G, Court LE. Prognostic Value and Reproducibility of Pretreatment CT Texture Features in Stage III Non-Small Cell Lung Cancer. *International Journal of Radiation Oncology, Biology, and Physics* doi: 10.1016/j.ijrobp.2014.07.020. Volume 90, Issue 4, Pages 834-842. 2014. ©Elsevier.

The permissions for reuse of these materials were obtained from both the Radiological Society of North America and Elsevier B V.

3.1 Conventional Prognostic Factors

We extracted patient T stage (T1/T2 vs T3/T4), N stage (N0/N1 vs N2/N3), Overall Stage (3a vs 3b), age, gender, histology (squamous cell carcinoma vs other), Karnofsky performance status (KPS) (100-90 vs <90), smoking status (current, former, never), estimated pack years, use of induction chemotherapy, and gross tumor volume (GTV) from the medical record. These factors were included as they have all been suggested to be prognostic in stage III NSCLC.⁴ All TNM staging was performed according to the 7th edition American Joint Committee on Cancer staging manual.³⁷ GTV consisted of *both* the primary and nodal disease as defined by the treating radiation oncologist for definitive radiation therapy. The GTV was transformed in all cohorts except Cohort 1 (the different cohorts are described below) prior to modeling using the logarithm to the base 2 in order to reduce the influence of relatively extreme measurements during the modeling process. The GTV in Cohort 1 was not log transformed as the distribution of GTV was relatively free of outliers and was approximately normally distributed. CPFs were used to construct reference prediction models using factors previously thought to be prognostic in stage III NSCLC in order to have an appropriate assessment for the incremental benefit of QIFs. It is essential in biomarker research to demonstrate evidence that new biomarkers (i.e., QIFs) provide added prognostic value in addition to what is already known from CPFs. There is some debate regarding whether to treat GTV as a CPF or QIF. While GTV is quantitative in nature and not considered part of determining AJCC TNM stage, it has been cited as a prognostic factor in NSCLC in a variety of publications. To accommodate both sides

of this debate, results using CPFs excluding GTV, CPFs including GTV, and CPFs including GTV and QIFs were determined.

3.2 Patient Cohorts

Cohort 1: 91 Patients with Pretreatment T_{avg} , T_{50} , and CE-CT

We retrospectively reviewed the medical records of patients with stage III NSCLC treated at MD Anderson Cancer Center with definitive radiation therapy between July 2004 and January 2012. These dates were chosen in order to include patients receiving 4DCT, which our institution implemented in early 2004, and provide adequate follow-up time. We excluded all patients receiving induction chemotherapy, proton-based radiation therapy, <5 years post-treatment for another solid tumor, multiple primary lesions, non-platinum-based concurrent chemotherapy, and those not receiving a diagnostic contrast-enhanced scan prior to 4DCT treatment planning. The median follow-up for all living patients at time of analysis was 59 months (range, 17 – 97 months). CPFs and treatment characteristics of all patients are listed in Table 1. All patients received a diagnostic contrast-enhanced CT (CE-CT) and a non-contrasted 4DCT scan prior to treatment. For contrast-enhanced scans, patients were scanned using 120 kVp, 400-1160 mA, and an exposure time of 265-570 ms. All images were reconstructed using the standard reconstruction kernel. Axial images were 512 x 512 pixels with voxel dimensions of 0.059-0.090 cm x 0.059-0.090 cm x 0.25 cm. For the 4DCT scans, the average intensity projection (T_{AVG}) and expiratory phase (T_{50}) images were used in this study. Patients were scanned using 120 kVp, 100-200 mA, and an exposure time of 500-800 ms. All images were reconstructed using the standard reconstruction kernel. Axial images were 512 x 512 pixels with voxel dimensions of 0.096 cm x 0.096 cm x 0.25-0.30 cm.

Effort was made to generate a cohort that was as homogeneous as possible in terms of their clinical characteristics (all stage III NSCLC), treatment characteristics (all treated with definitive radiation therapy), and imaging characteristics (similar acquisition/reconstruction parameters).

The aim of this cohort was to test the improvement of using QIFs from CT in outcome models compared to models using only CPFs.

Table 1. Cohort 1 – Patient CPFs and Treatment Characteristics

| Conventional Prognostic Factors | N | % | Treatment Characteristics | N | % |
|--|----------|----------|----------------------------------|----------|----------|
| No. Patients | 91 | NA | Radiation Dose | | |
| Median Age (years) | 65 | NA | 1.8- 2.0 Gy/fx | 79 | 87 |
| Mean GTV (cc) | 132 | NA | Other | 12 | 13 |
| Gender | | | Radiation Type | | |
| Male | 55 | 60 | 3DCRT | 5 | 6 |
| Female | 36 | 40 | IMRT | 86 | 94 |
| T Stage | | | Concurrent Chemotherapy | | |
| T1/T2 | 43 | 47 | Carboplatin-based | 78 | 86 |
| T3/T4 | 48 | 53 | Cisplatin-based | 13 | 14 |
| N Stage | | | Adjuvant Chemotherapy | | |
| N0/N1 | 11 | 12 | Yes | 37 | 41 |
| N2/N3 | 80 | 88 | No | 54 | 59 |
| Overall Stage | | | | | |
| IIIa | 45 | 50 | | | |
| IIIb | 46 | 50 | | | |
| Histology | | | | | |
| Squamous cell carcinoma | 46 | 50 | | | |
| Other | 45 | 50 | | | |
| Smoking Status | | | | | |
| Never | 5 | 6 | | | |
| Former | 65 | 71 | | | |
| Current | 21 | 23 | | | |
| Pack Years | | | | | |
| 0-24 | 13 | 14 | | | |
| 25-49 | 37 | 41 | | | |
| 50-74 | 22 | 24 | | | |
| 75+ | 19 | 21 | | | |
| Performance Status (KPS) | | | | | |
| 100-90 | 37 | 41 | | | |
| 80-70 | 53 | 58 | | | |
| <70 | 1 | 1 | | | |

Abbreviations: No. = Number; cc = cubic centimeters; GTV = gross tumor volume; KPS = Karnofsky performance status; Gy = gray; fx = fraction; 3DCRT = 3 dimensional conformal radiation therapy; IMRT = intensity modulated radiation therapy

This table has been reused with the permission of the original publisher from the following publication: Fried DV, Tucker SL, Zhou S, Liao Z, Mawlawi O, Ibbott G, Court LE. Prognostic Value and Reproducibility of Pretreatment CT Texture Features in Stage III Non-Small Cell Lung Cancer. *International Journal of Radiation Oncology, Biology, and Physics* doi: 10.1016/j.ijrobp.2014.07.020. Volume 90, Issue 4, Pages 834-842. 2014. ©Elsevier.

Cohort 2: 249 Patients with Pretreatment CE-CT

We retrospectively reviewed the medical records of patients with stage III NSCLC treated with definitive radiation therapy between August 2004 and December 2012. These dates were chosen in order to include many patients receiving contrast-enhanced CT scans during the time of use of 4DCT at MD Anderson Cancer Center and provide adequate follow-up time. We excluded all patients <5 years post-treatment for another solid tumor, multiple primary lesions and those not receiving a diagnostic contrast-enhanced scan prior treatment. These criteria yielded a cohort of 249 patients for analysis. The median follow-up for all patients living at the time of analysis was 53 months (range, 15 – 106 months). CPFs and treatment characteristics of all patients are listed in Table 2. All patients received a diagnostic contrast-enhanced CT (CE-CT) and a non-contrasted 4DCT scan prior to treatment. For contrast-enhanced scans, patients were scanned using 120 kVp, 400-1160 mA, and an exposure time of 265-570 ms. All images were reconstructed using the standard reconstruction kernel. Axial images were 512 x 512 pixels with voxel dimensions of 0.059-0.090 cm x 0.059-0.090 cm x 0.25 cm.

The purpose of this cohort was similar to cohort 1. Effort was made to identify a homogeneous cohort of patients in terms of clinical characteristics, treatment characteristics, and imaging characteristics. The predominant difference being the increased number of patients (91 versus 249) and that these were only required to have a pretreatment CE-CT for analysis. We found that CE-CT features appeared to be the most prognostic in section 4.1.1 Results for Project 1.1: Quantify the impact of adding CT-based quantitative image features to outcome models containing only CPFs including and excluding GTV. Therefore, this cohort was assembled to test if a model could be developed using only CE-CT derived features and test whether using QIFs from CE-CT improved outcome models compared to models using only CPFs.

Table 2. Cohort 2 - Patient CPFs and Treatment Characteristics

| Conventional Prognostic Factors | N | % | Treatment Characteristics | N | % |
|--|----------|----------|----------------------------------|----------|----------|
| No. Patients | 249 | NA | Fractionation | | |
| Median Age (years) | 66 | NA | 1.8- 2.0 Gy/fx | 227 | 91 |
| Mean GTV (cc) | 156 | NA | Other | 22 | 9 |
| Gender | | | Radiation Type | | |
| Male | 138 | 55 | 3DCRT | 9 | 4 |
| Female | 11 | 44 | IMRT | 187 | 75 |
| T Stage | | | Protons | 53 | 21 |
| T1/T2 | 145 | 58 | Chemotherapy Sequence | | |
| T3/T4 | 104 | 42 | Concurrent | 105 | 42 |
| N Stage | | | Induction-Concurrent | 60 | 24 |
| N0/N1 | 24 | 10 | Concurrent-Adjuvant | 69 | 28 |
| N2/N3 | 225 | 90 | Other | 8 | 3 |
| Overall Stage | | | None | 7 | 3 |
| IIIa | 131 | 53 | Concurrent Type | | |
| IIIb | 118 | 47 | Platin Doublet | 212 | 85 |
| Histology | | | Platin Doublet + Erlotinib | 13 | 5 |
| Squamous Cell Carcinoma | 104 | 42 | Single Agent Platin | 11 | 4 |
| Other | 145 | 58 | | | |
| Smoking Status | | | | | |
| Never | 16 | 6 | | | |
| Former | 182 | 74 | | | |
| Current | 51 | 20 | | | |
| Pack Years | | | | | |
| 0-24 | 44 | 18 | | | |
| 25-49 | 88 | 35 | | | |
| 50-74 | 63 | 25 | | | |
| 75+ | 54 | 22 | | | |
| Performance Status (KPS) | | | | | |
| 100-90 | 73 | 29 | | | |
| 80-70 | 171 | 69 | | | |
| <70 | 5 | 2 | | | |

Abbreviations: No. = Number; cc = cubic centimeters; GTV = gross tumor volume; KPS = Karnofsky performance status; Gy = gray; fx = fraction; 3DCRT = 3 dimensional conformal radiation therapy; IMRT = intensity modulated radiation therapy

Cohort 3: 195 Patients with Pretreatment PET

We retrospectively reviewed the medical records of patients with stage III NSCLC treated definitively with external beam radiation therapy between January 2008 and January 2013. These dates were chosen for two reasons: 1) to ensure patients' PET scans were acquired and reconstructed in 3D, which MD Anderson Cancer Center implemented in 2008, and 2) to ensure patients had a minimum potential follow-up of one year at the time of analysis. We excluded patients that were <5 years post-treatment for another solid tumor, had multiple primary lesions, or had primary lesions <5mL as measured on their PET scan. This yielded 195 patients for analysis. The median follow-up for all patients living at the time of analysis was 37 months (range, 3-70 months). Three patients were lost to follow-up prior to one year. CPFs and treatment characteristics of all patients are listed in Table 3.

All patients received a PET/CT scan prior to initiation of treatment. Scans were taken using either a GE Discovery RX or STE scanner at MD Anderson Cancer Center. Patients with PET scans taken at any outside institutions were excluded. All images were reconstructed using 3D-ordered subset expectation maximization using 2 iterations, 20-21 subsets, and a 6mm post-processing Gaussian blurring filter. All images were comprised of 128 x 128 pixels with voxel dimensions of 5.47 x 5.47 x 3.27 mm. Patients fasted for at least 6 hours prior to administration of an average injected dose of 381 Mbq (range, 255 – 540). The average duration from injection to scan was 78 minutes (range, 50 – 124). A low-dose non-contrasted CT was acquired for attenuation correction using 120 kVp, automated mA modulation, 1.35 pitch, and 3.75 mm slice thickness.

The aim of this cohort was similar to cohorts 1 and 2. We wanted to test the improvement of using QIFs from FDG-PET in outcome models compared to models using only CPFs.

Table 3. Cohort 3 - Patient CPFs and Treatment Characteristics

| Conventional Prognostic Factors | N | % | Treatment Characteristics | N | % |
|--|----------|----------|----------------------------------|----------|----------|
| No. Patients | 195 | NA | Fractionation | | |
| Median Age (years) | 66 | NA | 1.8- 2 Gy/fx | 160 | 82 |
| Mean GTV (cc) | 183 | NA | Other | 35 | 18 |
| Gender | | | Radiation Type | | |
| Male | 125 | 64 | 3DCRT | 1 | <1 |
| Female | 70 | 36 | IMRT | 126 | 66 |
| T Stage | | | Protons | 64 | 33 |
| T1/T2 | 97 | 50 | Chemotherapy Sequence | | |
| T3/T4 | 98 | 50 | Concurrent | 80 | 41 |
| N Stage | | | Induction-Concurrent | 56 | 29 |
| N0/N1 | 31 | 16 | Concurrent-Adjuvant | 46 | 23 |
| N2/N3 | 164 | 84 | Other | 11 | 6 |
| Overall Stage | | | None | 2 | 1 |
| IIIa | 107 | 55 | Concurrent Type | | |
| IIIb | 88 | 45 | Platin Doublet | 176 | 90 |
| Histology | | | Platin Doublet + Erlotinib | 13 | 7 |
| Squamous Cell Carcinoma | 89 | 46 | Single Agent Platin | 6 | 3 |
| Other | 106 | 54 | | | |
| Smoking Status | | | | | |
| Never | 19 | 10 | | | |
| Former | 130 | 66 | | | |
| Current | 46 | 24 | | | |
| Pack Years | | | | | |
| 0-24 | 47 | 24 | | | |
| 25-49 | 55 | 28 | | | |
| 50-74 | 49 | 25 | | | |
| 75+ | 44 | 23 | | | |
| Performance Status (KPS) | | | | | |
| 100-90 | 58 | 30 | | | |
| 80-70 | 131 | 67 | | | |
| <70 | 6 | 3 | | | |

Abbreviations: No. = Number; cc = cubic centimeters; GTV = gross tumor volume; KPS = Karnofsky performance status; Gy = gray; fx = fraction; 3DCRT = 3 dimensional conformal radiation therapy; IMRT = intensity modulated radiation therapy

This table has been reused with the permission of the original publisher from the following publication: Fried DV, Mawlawi O, Zhang L, Fave X, Zhou S, Ibbott G, Liao Z, Court LE. Stage III non-small cell lung cancer: Prognostic value of FDG PET quantitative image features combined with clinical prognostic factors. *Radiology* doi: 10.1148/radiol.2015142920. Published online July 15, 2015. ©Radiological Society of North America.

Cohort 4: 78 Patients with Pretreatment PET and Contrast-Enhanced CT

This cohort contained the 78 patients that were included in both Cohorts 2 and 3. The parameters regarding the patient's scans are thus the same as described in Cohorts 2 and 3. The primary function of this cohort was to compare features extracted from CECT versus PET/CT as well as assess auto-segmentation concordance in terms of necrosis volume.

Cohort 5: 24 Patients with “Large” Tumors on PET

This cohort contained 24 patients from Cohort 3 who had primary tumors ranging in size between 77 and 309 cc as measured on FDG-PET. The primary tumors of these patients were resampled within MIM version 6.2 (MIM Software Inc., Cleveland, OH) using trilinear interpolation. After each resampling, the PETedge tool was used to recontour the primary. This was done in an effort to determine how robust features were to changing the size of the region of interest (ROI). Some findings have indicated that the ROI volume can influence the obtained quantitative values due to the nature of some of the feature calculations.(Xenia Fave, unpublished). This approach was primarily used to determine how reproducible the mathematical formulas of the underlying features were to changes in volume. In addition, we used these resampled tumors in order to determine an approximate threshold where QIF reproducibility breaks down in our patients. The aim was to establish a cut-off for tumor volumes that are too small to be adequately assessed using our metrics.

Cohort 6: 53 Patients with “Pseudo” Test/Retest PET

This cohort contained 53 patients with NSCLC who received a PET/CT at an outside institution followed by a PET/CT at MD Anderson Cancer Center. No treatment was delivered to these patients between the two scans. There were no requirements regarding treatment, stage, etc., as was the case in other cohorts. The average time between scans was 48 days (range: 8 – 111). The purpose of this cohort was to analyze the reproducibility of PET-based QIFs by calculating the concordance correlation coefficient (CCC) (see 3.7.6 Concordance Correlation Coefficient).

3.3 Quantitative Image Features

3.3.1 Histogram Features

Histogram features used a first-order histogram that represents a particular region of interest by tabulating the number of pixels within a particular value. For QIFs requiring a histogram (standard deviation, uniformity, entropy), i.e., CT scans in Hounsfield units (HU), the images are first transformed into 8-bit images across the entire hypothetical CT range resulting in 16 HU bins (4096 HU values/ $2^8 = 16$ HU bins). This transformation was done in order to de-noise the image and supply a more appropriate bin size for histogram calculations (i.e., not 1 HU per bin). From a histogram, metrics, such as the mean, median, variance, entropy, skewness, kurtosis, and uniformity, can be calculated. The mean, median, and variance of a distribution are commonplace in mathematics/statistics; however, the concepts of entropy, skewness, kurtosis, and uniformity are often times less well-known. Entropy is generally thought of as the associated “randomness.” We define entropy below in equation 1.

$$\text{ENTROPY} = - \sum \left(\frac{H}{n} \right) * \log_2 \left(\frac{H}{n} \right)$$

$$\mathbf{H} = \text{histogram}, \quad \mathbf{n} = \text{total number of values}$$

Skewness is a measure of the asymmetry of the histogram where a positive value is when the distribution is skewed towards lower values and negative value is when a distribution is skewed towards higher values. We define skewness below in equation 2.

$$\text{SKEWNESS} = \frac{E(x - \mu)^3}{\sigma^3}$$

$$E = \text{expected value}, x = \text{values within histogram}, \mu = \text{mean}, \sigma = \text{standard deviation}$$

Kurtosis is a measure of the “peakedness” of the histogram where a positive value is when the distribution peaks more than the normal distribution and a negative value is when a distribution peaks less than the normal distribution. We define kurtosis below in equation 3.

$$\text{KURTOSIS} = \frac{E(x - \mu)^4}{\sigma^4}$$

$E = \text{expected value}, x = \text{values within histogram}, \mu = \text{mean}, \sigma = \text{standard deviation}$

Uniformity (or Energy as it is referred to in some publications) is a measure of how much variation of values is present in the histogram. A value of 1 means there is only one value within the ROI and the smaller the uniformity the more variation in the values within the ROI. We define uniformity below in equation 4.

$$\text{UNIFORMITY} = \sum \left(\frac{H}{n} \right)^2$$

$H = \text{histogram}, \quad n = \text{total number of values}$

The bin width used for histograms was different between CT and PET images. For CT, the features requiring a histogram (entropy and uniformity) used a bin width of 16 HU. For PET, the features requiring a histogram used a bin width of 1 SUV (i.e. entropy and uniformity). The remaining PET histogram features used the native floating point SUV values from the ROI (i.e. mean, maximum, peak, standard deviation, coefficient of variation).

3.3.2 Co-Occurrence Matrix Features

Co-occurrence matrix (COM) features were first proposed by Haralick et al. in 1973.⁹ These features expand upon information contained in histograms by also containing information regarding the spatial relationships between voxels. Traditionally, COM features are calculated by generating a matrix relating voxel displacement and directions. In this work, a voxel displacement of 1 was always used along with averaging across the unique directions (13 in 3D; 4 in 2D). This allowed the features calculated to be non-directional in nature. Once the average is performed, the COM is normalized by the total number of voxels to express the matrix in terms of probabilities rather than raw counts. Averaging across the 13 unique 3D directions was used in the analysis of Cohort 1. Subsequent analyses averaged across the 4-unique 2D directions. This transition was made in order to address the concern of non-isocentric displacements that arise when voxel x-y dimensions are not equivalent to the slice spacing within an image.

ROIs from CT and PET images were normalized prior to COM matrix feature calculation. In CT, the values within the ROI were scaled to 8 bits (256 values) over the standard digital CT representation range (4096 values) in the same manner as previously described. This effectively rounded the values within the ROI to the nearest 16 HU. In PET, images were first scaled to the number of gray levels between the minimum and maximum of the tumor SUV using the minimum and maximum as the gray level limits. This effectively rounded the SUV values within the contour to the nearest whole number and subtracted the minimum SUV. For example, a lesion with a minimum SUV of 3.2 and a maximum of 17.8 would be first scaled to be comprised of values ranging from 3 to 18 and then the minimum value (3) subtracted resulting in values from 0 – 15. This allowed for the analyses to have a finite number of gray levels and ensure that the new scaled values had a consistent relationship to the underlying SUV values (i.e., a difference of one between scaled values represented an SUV change of one). This methodology was recommended by Leijenaar et al. as this methodology allows for a more meaningful comparison of texture values between images.³⁸ By subtracting the minimum SUV value, the COM features were calculated using variability in uptake regardless of underlying amplitude. Other metrics, such as SUVmax and SUVmean, were used to quantify amplitude of uptake.

Numerous COM features exist within the literature. A variety of these features were investigated during the course of multiple projects; however, the four features that were consistently used in all analyses are explained below. $COM(i,j)$ corresponds to the COM for an arbitrary displacement and direction.

COM contrast quantifies the amount of discrepancy in values seen within the ROI. COM contrast increases when there are voxels within a displacement region that differ greatly in terms of their value (this is expressed in the $|i-j|^2$ term below).

$$CONTRAST = \sum_{i=1}^N \sum_{j=1}^N |i-j|^2 COM(i,j)$$

$N = \text{total number of values}$

COM correlation quantifies the joint probability occurrence of the specified pixel pairs (i.e., the dependency of values on those of the neighboring pixels).

$$\mathbf{CORRELATION} = \sum_{i,j=0}^{N-1} \mathbf{COM}(i,j) \left[\frac{(i - \mu_i)(j - \mu_j)}{\sigma_i^2 \sigma_j^2} \right]$$

$$\begin{aligned} \mu_i &= \text{mean of COM rows}, & \mu_j &= \text{mean of COM columns} \\ \sigma_i &= \text{variance of COM rows}, & \sigma_j &= \text{variance of COM columns} \end{aligned}$$

COM energy quantifies the variation in values seen within the ROI and is very similar to histogram uniformity. The main difference is that COM energy is based on values within the specified displacement (1 in this work) whereas in histogram uniformity, location of voxels plays no role. If the probability of finding adjacent voxels with different values is high, the COM energy will decrease, and if the probability of finding adjacent voxels with similar values or patterns of values is low, the COM energy will be closer to 1. A uniform image has a COM energy value of 1.

$$\mathbf{ENERGY} = \sum_{i=1}^N \sum_{j=1}^N \mathbf{COM}(i,j)^2$$

$$N = \text{total number of values}$$

COM homogeneity quantifies how consistent values are within the ROI. This can also be seen as the opposite of contrast only with a linear relationship between value differences rather than an exponential one.

$$\mathbf{HOMOGENEITY} = \sum_{i=1}^N \sum_{j=1}^N \frac{\mathbf{COM}(i,j)}{1 + |i - j|}$$

Additional COM features were only assessed in section 4.1 Results of Specific Aim 1. The formulas for these features are shown below.

$$\mathbf{SUM\ OF\ SQUARES:VARIANCE} = \sum_{i=1}^N \sum_{j=1}^N (i - \mu)^2 \mathbf{COM}(i,j)$$

$$\mu = \text{Mean value}$$

$$\mathbf{INVERSE\ DIFFERENCE\ MOMENT} = \sum_{i=1}^N \sum_{j=1}^N \frac{1}{1 + (i - j)^2} \mathbf{COM}(i, j)$$

$$\mathbf{SUM\ AVERAGE} = \sum_{i=2}^N i p_{x+y}(i) \mathbf{COM}(i, j)$$

$p_{x+y}(i)$ = probability of co – occurrence matrix corrdinates suming to $x + y$

$$\mathbf{SUM\ VARIANCE} = \sum_{i=2}^N (i - \mathbf{SUM\ ENTROPY})^2 p_{x+y}(i)$$

$$\mathbf{SUM\ ENTROPY} = - \sum_{i=2}^N p_{x+y}(i) \mathbf{LOG}\{p_{x+y}(i)\}$$

$$\mathbf{ENTROPY} = - \sum_{i=1}^N \sum_{j=1}^N \mathbf{COM}(i, j) \mathbf{LOG}\{\mathbf{COM}(i, j)\}$$

$$\mathbf{DIFFERENCE\ ENTROPY} = - \sum_{i=0}^{N-1} p_{x-y}(i) \mathbf{LOG}\{p_{x-y}(i)\}$$

$p_{x-y}(i)$ = probability of co – occurrence matrix corrdinates suming to $x - y$

$$\mathbf{INFORMATION\ MEASURE\ OF\ CORRELATION\ 1} = \frac{HXY - HXY1}{\mathbf{MAX}(HX, HY)}$$

$HXY = \mathbf{ENTROPY}$; HX and HY = entropies of p_x and p_y

$$HXY1 = - \sum_{i=1}^N \sum_{j=1}^N \mathbf{COM}(i, j) \mathbf{LOG}\{p_x(i)p_y(j)\}$$

$$HXY2 = - \sum_{i=1}^N \sum_{j=1}^N p_x(i)p_y(j) \mathbf{LOG}\{p_x(i)p_y(j)\}$$

$$\mathbf{INFORMATION\ MEASURE\ OF\ CORRELATION\ 2} = 1 - e^{\sqrt{-2(HXY2-HXY)}}$$

3.3.3 Nearest Gray Tone Difference Features

Nearest gray tone difference features were introduced by Amadasum and King.¹¹ Their aim was to design texture features corresponding to visual properties due to their wide applicability and promise in feature selection. We calculated these features after converting CT scans into 8-bit images in the same manner as described above in section 3.3.2 Co-Occurrence Matrix Features. These features are calculated by first constructing a 1-D matrix (NGTDM) where “the i^{th} entry is a summation of the differences between the gray level of all pixels with gray level I , and the average gray level of their surrounding neighbors.” Four features were extracted from this matrix: coarseness, contrast, busyness, and complexity using a neighborhood distance of 1 in all three dimensions.

Coarseness measures the size of the primitive (basic pattern) making up the texture. For instance, an image of static would have a small coarseness value; however, a checkerboard with large square sizes would have a large coarseness value.

$$COARSENESS = \left[\epsilon + \sum_{i=0}^{Gh} p_i s(i) \right]^{-1}$$

p_i = probability of occurrence of gray – tone (i), $s(i)$ = i^{th} entry in NGTDM,
 $\epsilon = 1E - 6$, Gh = highest gray – tone

Contrast is a measure of visible of different intensity levels within the ROI. For instance, an image of black and white stripes has adjacent regions of high and low values and thus has very high contrast.

$$CONTRAST = \left[\frac{1}{Ng(Ng - 1)} \sum_{j=0}^{Gh} \sum_{i=0}^{Gh} p_i p_j (i - j)^2 \right] \left[\frac{1}{Ng^2} \sum_{i=0}^{Gh} s(i) \right]$$

$p_{i/j}$ = probability of occurrence of gray – tone (i/j), $s(i)$ = i^{th} entry in NGTDM,
 Ng = total number of gray – tones, Gh = highest gray – tone

Busyness measures the amount of rapid changes of intensity from pixel to its neighbor (i.e., the spatial frequency of intensity changes).

$$BUSYNESS = \left[\sum_{i=0}^{Gh} p_i s(i) \right] / \sum_{j=0}^{Gh} \sum_{i=0}^{Gh} i p_i - j p_j$$

$p_{i/j}$ = probability of occurrence of gray – tone (i/j), $s(i)$ = i th entry in NGTDM,
 Gh = highest gray – tone

Complexity measures the information content within an ROI. ROIs with many primitives with varying average intensities are viewed to have high content and thus high complexity.

$$\sum_{i=0}^{Gh} \sum_{j=0}^{Gh} \left[\frac{|i-j|}{n^2(p_i + p_j)} \right] [p_i s(i) + p_j s(j)]$$

$p_{i/j}$ = probability of occurrence of gray – tone (i/j), $s(i)$ = i th entry in NGTDM,
 Gh = highest gray – tone

3.3.4 Laplacian of Gaussian Filtration Features

Laplacian of Gaussian filtration (LOG) features rely on the Laplacian of Gaussian filter, which is commonly used for edge and/or blob detection. An LOG filter first convolves a given image with a Gaussian kernel with a specified scale (σ) and then the result is computed with a Laplacian operator. The scales examined were 1, 1.5, 1.8, 2.0 and 2.5. These scales along with a filter size of 11-voxels were chosen as they have been used extensively in publications by Ganeshan et al.^{10, 14, 15, 19} Since the voxels within our images were approximately 1mm in the x-y dimension, edges/objects larger than approximately 4mm ($\sigma = 1$) to 12mm ($\sigma = 2.5$) were not blurred out by the Gaussian portion of the LOG filter. This allowed the selection of a particular scale by which to blur out potential image noise and selectively focus on areas of interest of differing sizes (e.g. regions of heterogeneity between 4 and 12mm). A filter size of 11-voxels was used as this was found to be sufficient to include the entirety of the filter at the various scales yet an increase in this size did not drastically alter the quantified feature values. Modifications to a pure LOG filtering process need to be made when applying this filter to a ROI and not an entire image. The steps used in our work are described below:

1. A threshold is applied to the original ROI to exclude values below -50HU (i.e., air). These values are replaced with the value 5000 (or arbitrary “high” number in relation to the HU values within the image). This is done to remove edge effects later in the calculation.
2. The LOG kernel is applied to each 2D slice with the designated σ with a size of 11x11.
3. The result is converted into an unsigned integer so that all negative values are converted to zeros. The replacement of values with 5000 in Step 1 is performed so that influences of the edge of the ROI are negative values and are thus removed.

Once these steps are performed, the resulting filtered image is used to calculate the average, standard deviation, entropy, and uniformity of the results. An example of the result can be seen in Figure 1. It can be seen that the filtered image highlights the edges seen within the original image but does not take into consideration the edge of the air cavity within the center of the tumor. It can also be seen that the “tissue” in the superior portion of the tumor has a larger number and more intense edges than the necrotic fluid present under the air cavity. The edges seen within the necrotic fluid are most likely due to noise from the reconstruction while the more intense edges in the tumor tissue are most likely due to vessel and tissue contrast enhancement.

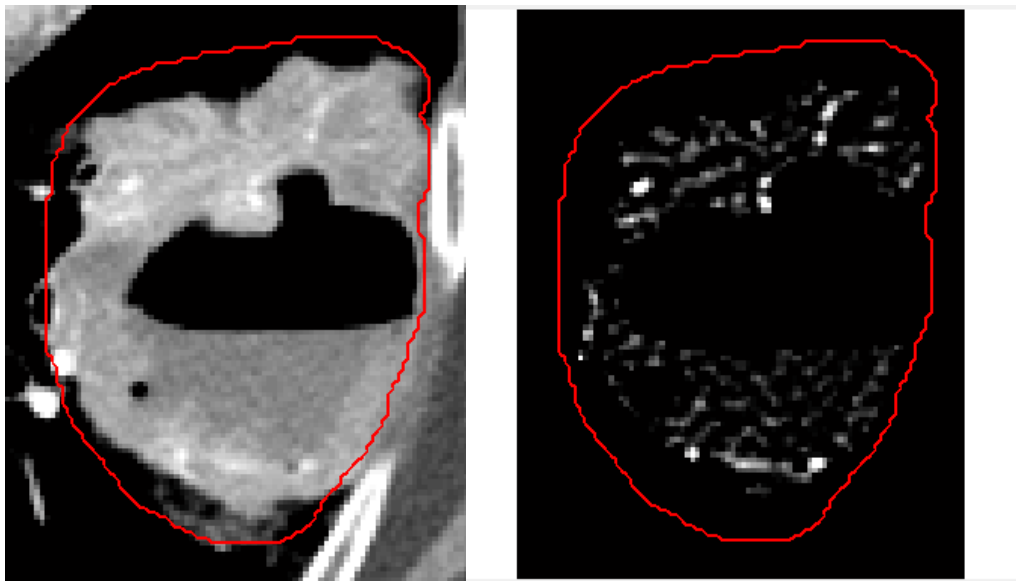


Figure 1. Laplacian of Gaussian Example. Original Tumor (left) and Results of LOG Filtration (right) ($\sigma = 1$)

3.3.5 Contrast Enhanced CT Auto-segmentation of Morphologic Characteristics

We developed an auto-segmentation algorithm to separate tumors into air, necrosis, tissue, and enhancing vessels components (see Figure 2). Physician-delineated gross tumor volumes (GTV) were drawn from patients' treatment CTs (non-contrasted). The GTV contour was transferred to the diagnostic CE-CT after deformable registration of the two images. The cutoffs used within the auto-segmentation algorithms were defined by taking manually drawn regions of interest from 10+ patients. Tumor tissue, enhancing vessels, and necrosis were delineated and used to determine histograms for each tissue type. Values that maximally separated the histograms from each tissue type were used as cutoffs. The auto-segmentation allows for the calculation of the volume of each category along with the percentage each category represents of a particular tumor. The algorithm follows the series of steps described below:

1. Air outside the tumor is removed via thresholding below the value of -50 HU.

Any portions that are removed but are surrounded by non-excluded voxels are filled with the value of 0 HU. The voxels are filled to minimize the impact of air in the following step (blurring).

2. The resulting region of interest (ROI) is blurred with a Gaussian kernel of size 5x5 pixels and a sigma of 1.5. The blurring leads to more accurate thresholding segmentation that is less influenced by image noise.
3. An initial guess of what regions constitute necrosis is made by finding voxels with values between -25 HU and 20 HU. Values surrounded by voxels deemed to be part of the initial necrosis guess are also deemed part of the initial necrosis guess.
(Steps 4-9 deal exclusively with identifying necrosis).
4. The initial necrosis guess is then eroded two pixels and dilated two pixels. The purpose of this is to remove small regions.

5. If a necrotic guess does not exist, then the tumor is deemed to have no voxels that constitute necrosis, and the algorithm proceeds to Step 10. If a necrotic guess does exist, the algorithm continues to Step 6.
6. The largest 3D continuous region of the initial necrosis guess is identified and the centroid found.
7. From this centroid, a 3D region growing algorithm is performed using an inclusion criterion of being within $\pm 35\text{HU}$ of the initial centroid seed point.
8. Post region growing, 2D regions containing less than 50 pixels are removed. This is done to avoid small regions of necrosis that are not usually observed (i.e., a vast majority of necrotic regions are quite sizable).
9. The resulting voxels are deemed “necrosis.”
10. Regions having values less than -50Hu are deemed “air.”
11. Voxels with values greater than 120HU are the initial enhancing vessel guess. Regions that are less than 3 voxels are removed and the result is deemed “enhancing vessels”.
12. Voxels with values between 20 HU and 120 HU are the initial tissue guess. Regions smaller than 20 voxels are removed from the guess.
13. Regions belonging to air, necrosis, or enhancing vessels are removed from the tissue guess.
14. The results are deemed “tissue.” Unlabeled regions can exist; however, these regions are usually located at a tumor/air border and have a very small volume.

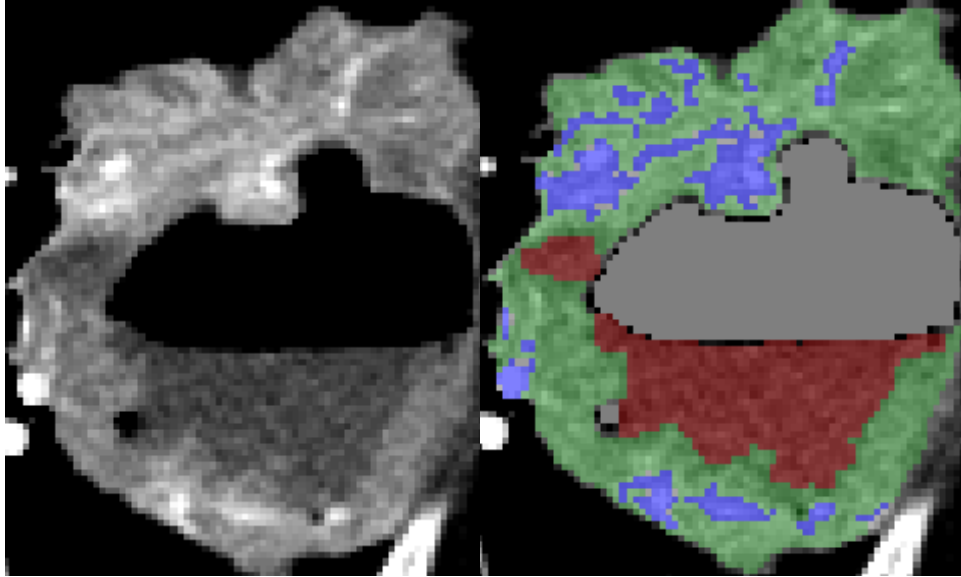


Figure 2. CE-CT Results of Auto-segmentation of Air (gray), Necrosis (red), Tissue (green), and Vessels (blue)

This auto-segmentation allows for the calculation of the volume of each category along with the percentage each category represents of a particular tumor.

3.3.6 PET Necrosis Auto-segmentation

The PET necrosis auto-segmentation algorithm first must have an ROI that encompasses the tumor. These contours were first made using the PETedge tool in MIMvista as described in previous work.³⁹ The algorithm then follows the series of steps below:

1. Voxels within 5 voxels of the edge vertically or horizontally or 3 voxels diagonally are deemed not eligible to be labeled as necrosis.
2. If no voxels are eligible or the max SUV within the eligible voxels is less than 8, the tumor is designated to not have any necrosis. Otherwise, proceed to Step 3. The value of 8 was used as it was observed during algorithm development that necrotic regions rarely had SUV values higher than 8.
3. A k-means clustering is performed with the number of means equal to one-fifth the range of the eligible voxels plus one. The lowest k-means centroid from this process is used as the

- threshold for necrosis unless this value is greater than an SUV of 5. If the threshold from the k-means process is greater than 5, then 5 is used as the guess. The purpose of this is to adjust the level of the threshold based on the range of the SUV values within the tumor. Tumors with lower overall SUV values tend to also have lower SUV within necrotic regions. This adaptive step helps to facilitate picking a representative cutoff.
4. The voxels that are eligible and greater than the threshold cutoff are deemed not necrosis while the remaining values are still eligible.
 5. The voxels still eligible to be deemed necrotic are then eroded using the same process as described in Step 1.
 6. Remaining eligible voxels are removed if they are not at least connected to two other voxels (i.e., the minimum size criteria for a necrotic region is 3 voxels).
 7. Voxels that remain eligible after these steps are deemed necrotic.

An example of the result of this process is shown below in Figure 3.

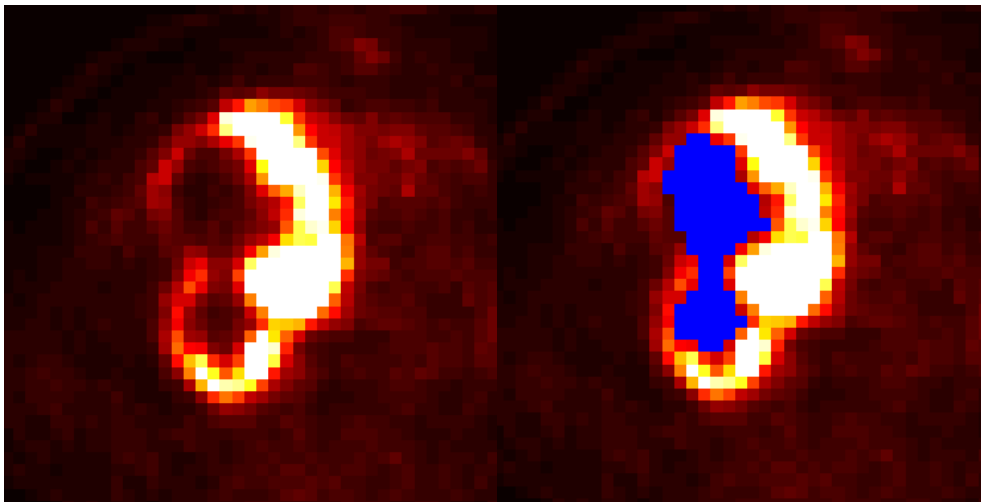


Figure 3. PET Results of Auto-segmentation of Necrosis (blue)

This auto-segmentation allows for the calculation of the volume of necrosis and the percentage of the tumor that is necrotic. The segmentation of necrosis is simpler on FDG-PET than on

CE-CT due to its functional nature and higher contrast between necrotic and non-necrotic tissue. The volume of necrosis and the percentage of the tumor that is necrotic were compared between autosegmentation methods as a measure of robustness of the CE-CT segmentation.

3.4 Region of Interest Contouring on CT

Tumor texture analysis on CT scans was conducted using the primary gross tumor volume (GTV) contour delineated by each patient's treating physician. The nodal tumor volumes were excluded from texture analysis. In some cases, further contour modification was performed. The reason for further modification is due to the goal of analyzing tissue with an *extremely high* likelihood of representing tumor, whereas clinically physicians routinely include any and all tissue with a *reasonable* likelihood of representing tumor. Therefore, overly generous portions of the contour, such as invasion into bone and other normal tissue structures such as the aorta, were modified. Contours were extracted and analyzed directly from the treatment planning system using in-house software (IBEX) developed by Luke Hunter and Dr. Lifei Zhang built using a commercial software package (Matlab version 8.1.0. Natick, Massachusetts: The MathWorks Inc., 2013).^{40, 41} For the AVG-CT and T50-CT, a lower and upper threshold of -100 to 200 HU was implemented to exclude lung tissue, air, and/or bone in order to determine our final ROI. A lower threshold of -100 HU was used for the contrast enhanced images in Cohort 1 and -50 HU threshold in subsequent cohorts. In CE-CT images, no upper threshold was used. Only voxels within the defined threshold bounds were included in the texture analysis.

3.5 Region of Interest Contouring on PET

Patient's primary and nodal tumor volumes were delineated using the PETedge feature from MIM version 6.2 (MIM Software Inc., Cleveland, OH). This method was chosen as it was found to be the most accurate and consistent technique for target volume contouring for lung cancer lesions on FDG-PET in an extensive review by Werner-Wasik et al.⁴² This study found that PETedge was the most accurate for both segmenting spheres at multiple source-to-background ratios and multiple sizes. Spheres of a known volume that were > 20mm and < 20mm in diameter were found to have lower

mean absolute percent error (10.99% error) using PET edge compared to thresholding (17.5% error) (25% to 50% at 5% increments of SUVmax) and manual segmentation (19.5% error). It was also observed that PETedge had the least systematic bias (-0.05% error) among the segmentation methods tested. PETedge works, in principal, by first placing the cursor towards the center of the lesion of interest. Upon clicking, four “spokes” emanate from the central point in orthogonal directions and dragging the cursor extends them until they reach what the user identifies as a reasonable edge of the tumor. The software then uses the gradient of the SUV values to determine where the maximum descent is located and contours accordingly. This algorithm is purely quantitative and is therefore not influenced by the user’s preference of window/level, which can generate significant variation in the apparent size of the lesion. For heterogeneous and/or necrotic tumors, this sequence sometimes had to be repeated more than once to adequately cover the entire tumor. The PETedge algorithm is semi-automated and thus is capable of higher throughput than manual contouring. When contouring the primary and nodal volumes on the FDG-PET, the radiation treatment plan and the diagnostic radiology notes were consulted to determine location of primary and nodal disease. Once the contours were finished, they were exported along with the FDG-PET image into IBEX.

One issue that arose during this process was that the RT structure is stored as a series of spatial points based on a resolution higher than that of PET scans. IBEX is programmed to convert these line/point contours into binary masks for analysis. IBEX considers any voxel containing or within the contour as part of the binary mask. This was problematic for the voxels at the edge of the contour due to the relatively low-resolution of PET. The issue initially observed was that voxels could be included in the binary mask when the majority of the voxel was not even within the contour. In order to remedy this, Dr. Lifei Zhang and I developed a resampling algorithm that determined the fraction of the edge PET voxels that were included in the delineated contour. This was then used to determine how much of the voxel should be included in the contour in order to be analyzed (i.e., included in the binary mask). The results of this process can be seen in Figure 4. For this work, a 50% cutoff was used to determine which voxels would be analyzed (i.e., if a majority of the voxel was

inside the contour, it was analyzed). This algorithm helped ensure that unnecessary low SUV voxels from the tumor edge were not included in the analyses.

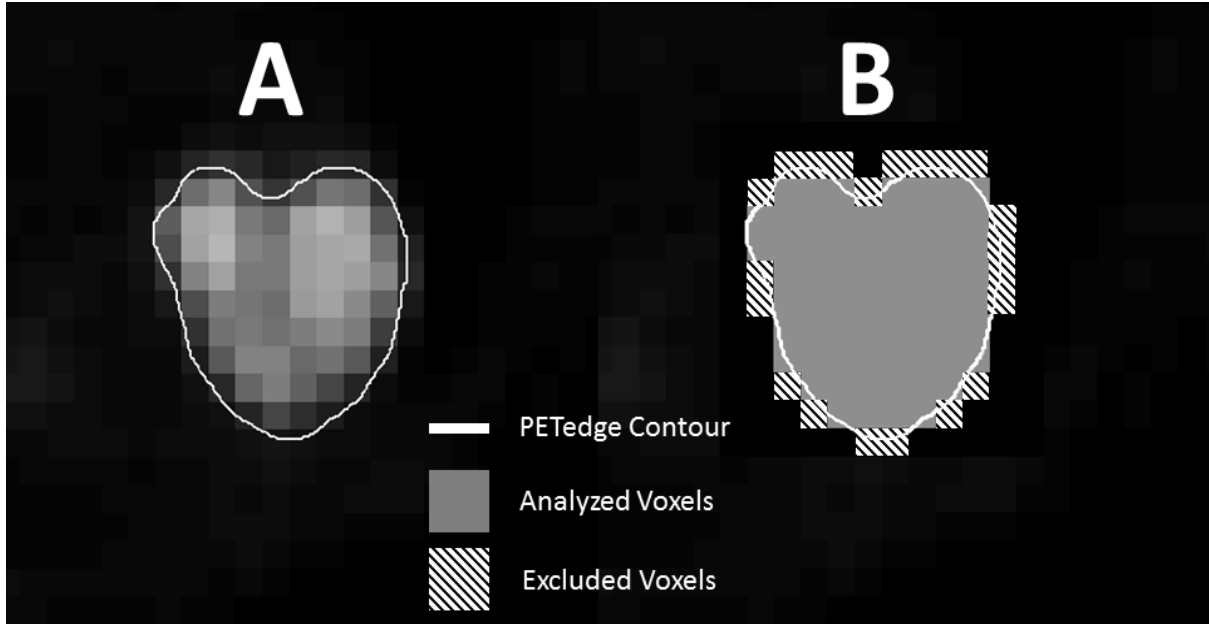


Figure 4. (A) Original FDG-PET Contour/Image (B) Analyzed Voxels Using a 50% Cutoff

3.6 Assessment of QIF Reproducibility using Phantom and Patient Data

A National Electrical manufacturers Association (NEMA) International Electrotechnical Commission (IEC) PET phantom was used to assess the reproducibility of QIFs on 3 different scanners (GE Discovery VCT, GE Discovery 710, and Siemens mCT). Phantom preparation and initial scanning was performed by Dr. Osama Mawlawi and Joe Meier. Acquisitions were made using a source-to-background ratio of approximately 10:1. The phantom contains 6 spheres with inner diameters of 10, 13, 17, 22, 28 and 37mm suspended in the background of ^{18}F -FDG water. The phantom was positioned at the center of the FOV of the PET scanner and data were acquired in 3D mode. Acquisitions lengths were varied in an effort to provide equal integrated disintegrations (i.e. count statistics). Initial acquisitions were made on three PET scanners (GE Discovery VCT, GE Discovery 710, and Siemens mCT) using standard clinical protocols by Dr. Osama Mawlawi and Joe

Meier. I then retrospectively reconstructed using different values of iteration, subsets, Gaussian filter width, and matrix size. Table 4 below illustrates the reconstructions acquired per scanner. These values were chosen as the seemed representative of what was performed clinically in a publication analyzing PET/CT scanner performance characterization.⁴³ The differences in parameters between scanners (i.e., 21 versus 24 subsets) were due to differences in manufacturer options. Both the 710 and mCT were reconstructed using time-of-flight whereas the VCT was not as it is not time-of-flight capable.

Table 4. Parameters Used for Phantom and Patient Retrospective Reconstructions

| <i>GE Discovery VCT</i> | | | | <i>GE Discovery 710</i> | | | | <i>Siemens mCT</i> | | | |
|-------------------------|----|----|-------------|-------------------------|----|----|-------------|--------------------|----|----|-------------|
| Iter | SS | FW | Matrix Size | Iter | SS | FW | Matrix Size | Iter | SS | FW | Matrix Size |
| 2 | 21 | 2 | 128 | 2 | 24 | 2 | 128 | 2 | 21 | 2 | 128 |
| 2 | 21 | 2 | 256 | 2 | 24 | 4 | 128 | 2 | 21 | 4 | 128 |
| 2 | 21 | 4 | 128 | 2 | 24 | 6 | 128 | 2 | 21 | 6 | 128 |
| 2 | 21 | 4 | 256 | 2 | 24 | 2 | 192 | 2 | 21 | 2 | 200 |
| 2 | 21 | 6 | 128 | 2 | 24 | 4 | 192 | 2 | 21 | 4 | 200 |
| 2 | 21 | 6 | 256 | 2 | 24 | 6 | 192 | 2 | 21 | 6 | 200 |
| 3 | 21 | 2 | 128 | 2 | 24 | 2 | 256 | 2 | 21 | 2 | 256 |
| 3 | 21 | 2 | 256 | 2 | 24 | 4 | 256 | 2 | 21 | 4 | 256 |
| 3 | 21 | 4 | 128 | 2 | 24 | 6 | 256 | 2 | 21 | 6 | 256 |
| 3 | 21 | 4 | 256 | 3 | 24 | 2 | 128 | 3 | 21 | 2 | 128 |
| 3 | 21 | 6 | 128 | 3 | 24 | 4 | 128 | 3 | 21 | 4 | 128 |
| 3 | 21 | 6 | 256 | 3 | 24 | 6 | 128 | 3 | 21 | 6 | 128 |
| | | | | 3 | 24 | 2 | 192 | 3 | 21 | 2 | 200 |
| | | | | 3 | 24 | 4 | 192 | 3 | 21 | 4 | 200 |
| | | | | 3 | 24 | 6 | 192 | 3 | 21 | 6 | 200 |
| | | | | 3 | 24 | 2 | 256 | 3 | 21 | 2 | 256 |
| | | | | 3 | 24 | 4 | 256 | 3 | 21 | 4 | 256 |
| | | | | 3 | 24 | 6 | 256 | 3 | 21 | 6 | 256 |

Iter = iterations, SS = subsets, FW = filter width

The image slice going through the center of the spheres was used for analysis. Since the spheres were of different sizes, each sphere was impacted to a different degree by partial volume effects. Therefore, the SUV values within each of the spheres were not consistent. We exploited this observation and proceeded to use these spheres (all at 10:1 source to background) as a surrogate for assessing stability of heterogeneous values that are seen in patient tumors.

We first analyzed all images for each scanner (VCT, 710, mCT) that used a fixed matrix size (e.g., 128, 192, 200, or 256). This led to image groupings of 6 images of consistent matrix sizes and therefore voxel sizes. These image groups were able to be assessed in a voxel-by-voxel manner since these images able to be perfectly overlaid on another. A “range image” could then be produced where

the image showed the size of the SUV range within a particular voxel across all six analyzed images that used various iterations, subsets, and filter widths within the reconstruction. For example, a particular voxel within one of the spheres had values of 1.1, 1.4, 0.9, 2.1, 1.7, and 1.0 in the 6 images obtained from a single scanner at a particular matrix size. That voxel in the range image would have a value of 1.2 (max value: 2.1 – minimum value: 0.9). An example of a range image is shown below in Figure 5. From these range images, we could calculate what percentage of voxels within the spheres had a maximum SUV less than 1. This cutoff was chosen because in our analyses the bin size when calculating our QIFs was set to 1. Therefore, voxel changes <1 would not influence the resulting feature value more than what is observed by discretizing the image into integer bins from a floating point number (SUV) initially.

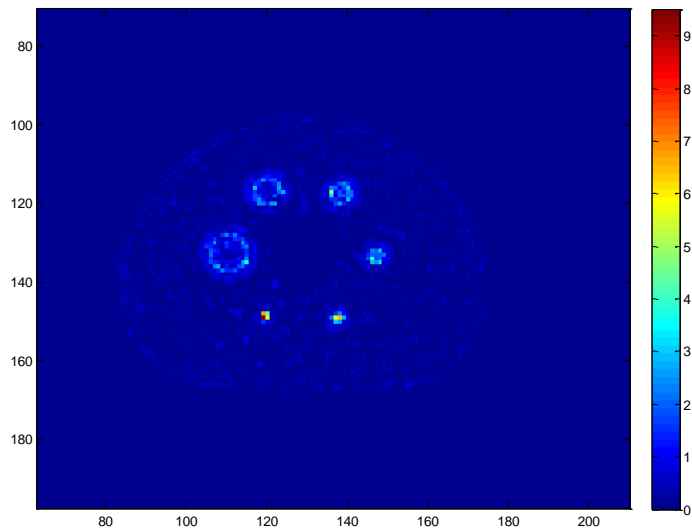


Figure 5. Example Range Image of NEMA IEC Phantom

Additionally, we calculated QIF values using values within the spheres. We used a threshold ($\text{SUV} > 1.5$) on the clinical protocol image (or image closest to what is performed clinically) in order to define the contour of the six spheres. This contour was used within each scanner as the input mask for calculating the QIFs. The purpose of this segmentation methodology was to ensure that the same

contour was being assessed across the different images in order to demonstrate that the changes in QIFs were due to the changes in values and not changes in contouring. QIF values were extracted from each of the three scanners using the various reconstruction parameters. The median, mean, and standard deviation seen within each scanner and across all scanners were calculated and compared to the values from the patient data in terms of their standard deviations (Table 13).

A similar process was performed using patient scans. We identified 5 patients (with 6 analyzed lesions) with NSCLC with primary tumors that were similar to those included in Cohort 3. In terms of volume, these 6 tumors reasonable spanned the range of volumes seen in Cohort 3 (16cc – 275cc). These patient scans were reconstructed using the same parameters as the phantom scans and described in Table 4. However, the primary was re-contoured separately for each reconstruction whereas in the phantom analysis the same contour was used for the six spheres across all reconstructions for a given matrix size.

3.7 Statistical Methods

3.7.1 Use of Cross-Validation for Assessment of Prognostic Value

All statistical analyses were conducted in R 3.0.2 with the following R packages: survival v(2.37-4), penalized (0.9-42), and survcomp v(1.10.0). The CPFs and QIFs (features calculated for primary, nodal, and total disease were defined as separate features) extracted from each patient's pretreatment CT or PET scan were entered into a penalized multivariate Cox proportional hazards model. The Cox proportional hazards model is a survival model relating a unit increase in a covariate to the hazard rate (i.e., risk of event per unit time). The hazard function for the proportional hazards model is shown below.

$$h(t|X) = h(t) \exp(X_1\beta_1 + \dots + X_p\beta_p)$$

$h(t)$ = underlying hazard, $h(t|X)$ = underlying hazard as a function of covariates

X = covariate value, β = covariate coefficient, p = number of predictors

This penalized modeling framework simultaneously carries out covariate selection alongside model development. Covariate selection is performed as L1 penalization reduces non-informative covariate coefficients to zero. This selection allows the user to input “eligible” covariates with the algorithm returning only covariates seeming to have prognostic ability via cross-validation. The penalization is directed by the L1 penalty parameter, which balances model fit and model complexity. The penalty parameter is determined by maximizing the cross-validated likelihood. The R penalized package standardizes all covariates by their unit central L2-norm prior to penalization in order to minimize the influence of covariate’s scales. The model coefficients are subsequently rescaled to reflect their original covariate’s magnitudes.

In order to adjust for the bias associated with training and testing a model on the same internal dataset, we predominantly used methodologies suggested by Simon et al. to generate cross-validated Kaplan-Meier curves.⁴⁴ This methodology allows a reasonable estimate for out-of-sample performance of our models while only using an internal dataset.⁴⁵ Cross-validated Kaplan-Meier curves are generated using model predictions for patients that are derived from models developed without the patient’s inclusion in model training. When performing leave-one-out cross validation, a patient is left out of model development and a prediction for this patient is generating using the remaining cohort. The patient who is left out is changed and this process repeated such that each observation in the sample has a prediction from when it was not involved in model development. These predictions (we utilized the linear predictor generated during each fold of cross-validation) are used to stratify patients into risk groups. The linear predictor is defined as the sum of each of the model coefficient times the corresponding covariate value of that specific patient. Therefore, the higher the linear predictor, the higher the predicted risk. We then used these predictions to generate risk groups based on a median cutoff in Cohort 1 due to the low number of patients or k-means clusters (see 3.7.5 K-Means Clustering of Predictions). In addition, we also calculated the concordance index (c-index) at multiple time points (see 3.7.3 Concordance Index at Multiple Time Points).

The outcome of overall survival was of primary interest in the analyses followed by local-regional control and freedom from distant metastases. The outcomes were measured as the time from the initiation of treatment until the corresponding event in months. Treatment initiation was defined as the first cycle of chemotherapy for patients receiving induction chemotherapy or the first day of radiation treatment for patients receiving radiotherapy upfront. Patients not experiencing an event were censored at the last known follow-up date. The MD Anderson Cancer Center Institutional Review Board approved all retrospective chart review study and waived the need for informed consent. The study complied with all Health Insurance Portability and Accountability Act (HIPAA) regulations.

To develop a model using covariates found to be predictive in cross-validation, QIFs that were included in greater than 50% of the folds were used along with the CPFs that were included in greater than 50% of folds in the preceding analysis. These nested models were compared using a likelihood ratio test to assess for impact of adding QIFs.

3.7.2 Permutation Test and Impact of Feature Reproducibility on Predictions (Cohort 1)

In Cohort 1, a permutation test was performed where the patient outcomes were randomly permuted with respect the QIFs and CPFs and the original analysis was re-run. This process was repeated 200 times in order to determine what proportion of randomly permuted data achieved a log-rank score greater than our original models (i.e., the p-value).

Test-retest scans were obtained from 10, 10, and 13 independent patients for the T_{AVG-CT} , T_{50-CT} , and $CE-CT$, respectively. The test-retest scans of the $AVG-CT$ and T_{50-CT} images were taken at MD Anderson Cancer Center and on average separated by 27 min (range: 16-47). $CE-CT$ test-retest images from patients within a close time period were not available; therefore, CE scans taken outside MD Anderson Cancer Center prior to treatment were obtained and compared to the diagnostic $CE-CT$ taken within MD Anderson Cancer Center. The average separation between these scans was 38 days (range: 17-72). The contours for the test-retest scans were performed by a single observer (DF) on separate occasions for the test and retest scans in order to incorporate intra-observer

contour variability. The classification reproducibility of our models was calculated incorporating the reproducibility seen via the test-retest scans. This was performed by utilizing the mean and standard deviation of the differences between the extracted metrics from the test and re-test scans to generate a normal distribution.

We added the values obtained by sampling this normal distribution with the associated mean and standard deviation of the differences for each feature to the original features. These values were then put through the same cross-validation process as the original feature values to determine how the reproducibility would influence the predictions and subsequent classification reproducibility.

The classification reproducibility is defined as the percent of patients categorized into the same group as the original models when incorporating the test-retest variation into the texture parameters. This was done for the models incorporating the QIFs and CPFs. The CPFs were assumed to be constant.

3.7.3 Concordance Index at Multiple Time Points

The concordance index or c-index was originally introduced by Harrell et al.⁴⁶ The purpose of this measure was to serve as an analog for area under the receiver operating characteristic curve in survival analysis. The c-index is computed by analyzing all eligible combinations of patient pairs within a cohort. In order to generate c-indices at multiple time points, restrictions were placed as to which patient pairs would be eligible to contribute to the c-index calculation. For example, the c-index at 6 months only allowed patient pairs whose outcomes differed by at least 6 months.

3.7.4 Analysis of Relationship between Quantitative Image Features, Conventional Features and Morphologic Characteristics

We used the 249 patients in Cohort 2 for this analysis. When comparing QIFs to conventional features, such as staging and volume, the QIF values for conventional feature values above and below the median were assessed using a Wilcoxon rank-sum test. When comparing QIFs to morphologic characteristics, the 249 tumors were divided into two groups (tissue type present or absent based on auto-segmentation methods previously described) and compared by their quantitative feature values.

Significant differences between the presence/absence of a particular tissue type of tumors in terms of the resulting QIFs values were assessed using a Wilcoxon rank-sum test.

Additionally, QIF values from tumors containing a particular tissue type were compared to the QIF values from the same tumors using contours that excluded one or all morphological characteristics (e.g., excluding air and necrosis, excluding air and enhancing vessels, including only tissue [i.e., excluding air, necrosis, and enhancing vessels respectively], etc.). QIF values obtained from the ROIs that excluded a single or combination of tissues were plotted versus the QIF values obtained from the entire tumor ROI in Figure 21. In addition, the plots were stratified for vessels and necrosis according to whether the volume of vessels or necrosis was greater than the average across all tumors where these tissue types were present. Differences between the values obtained from the entire tumor versus the same tumors excluding the tissue type(s) were assessed using a paired Wilcoxon signed-rank test. Differences in values of LOG_Average and uniformity between tumors with higher than average volume of vessels/necrosis versus lower than average volume of vessels/necrosis were assessed using a Wilcoxon rank-sum test. The coefficient of determination (R^2) was calculated from the linear regression for each of these plots.

The reproducibility of the values obtained for volume and percentage of tumor containing necrosis between CE-CT and FDG-PET auto-segmentation methodologies was assessed using the concordance correlation coefficient (CCC).⁴⁷

3.7.5 K-Means Clustering of Predictions

The process of using a cutoff based on optimizing the log-rank statistic, the median, or balancing patient numbers in each subgroup is quite common in many publications regarding prognostic scores or outcome prediction models. Optimization of the log-rank statistic is not a preferred method, since these results are almost always overly optimistic and can be problematic, particularly when the number of covariates being analyzed increases. Use of the median to separate patients into “high risk” and “low risk” is reasonable but not optimal. This process does not take into account the underlying distribution of predictions and frequently groups patients with similar

predictions into different risk groups. Balancing patient numbers (i.e., generating groups with the same number of patients in each category) suffers from the same issue as using a median cutoff. We proposed implementing a k-means clustering on the generated predictions in order to generate more uniform cohorts. In linear regression models, a calibration curve is often shown, plotting the predictions versus the actual values. For survival analyses this is not possible since survival curves must be generated from multiple patients and the issue of censoring makes it so not every patient has a defined “actual value.” By generating uniform cohorts, the displayed Kaplan-Meier curves are a more reflective display of calibration of the model/predictions. In our analyses, we increased the number of k-means clusters so long as each risk group contained a significant number of patients.

3.7.6 Concordance Correlation Coefficient

The concordance correlation coefficient (CCC) measures reproducibility between two covariates. Others have used this metric (or the nearly identical intra-class correlation coefficient) to assess the reproducibility of texture features under different acquisition conditions. We used the CCC when analyzing the reproducibility when resampling in Cohorts 1 and 5 along with the “pseudo” test/retest scans from Cohort 6.

3.7.7 Analysis of PET Tumor Resampling (Cohort 5)

Primary tumors in Cohort 5 were resampled to different spatial resolutions using trilinear interpolation using MIM 6.2. The resulting images of the primary tumors were re-contoured using the PETedge feature after each interpolation and grouped into similarly sized cohorts based on the number of voxels present. Voxel groupings consisted of the following approximate voxel sizes: 27 (range: 18-34), 55 (range: 43-73), 108 (range: 92-138), 226 (range: 171-322), and 488 (range: 369-641). The features calculated from the resampled ROIs were compared to the features calculated from the ROIs at native resolution using the CCC. This was performed in an effort to identify if any mathematical biases existed due to the nature of the QIF calculations.

3.7.8 Sub-cohorts Based on FDG-PET QIFs to Determine Impact of Dose Escalation

This analysis was performed on Cohort 3. Patients receiving 60-70 Gy (median: 66 Gy) were considered “low dose” while patient receiving 74 Gy were considered “high dose.” Different sub-cohorts comprised of ranges of both solidity and COM energy were examined to determine if any sub-cohorts would demonstrate a survival difference between those receiving low and high doses. Sub-cohorts were created allowing different values of solidity and COM energy at five percentile thresholds. This was done in order to observe if there existed any pattern regarding the impact of dose escalation in regards to our QIFs. A log-rank test was used to determine significance of separation between low-dose and high-dose patient Kaplan-Meier curves. We refer to the sub-cohort chosen from examining the trend of increasing COM energy and solidity values as the “high QIFs values sub-cohort” and the sub-cohort chosen from examining the trend of decreasing COM energy and solidity values as the “low QIFs values sub-cohort.” The sub-cohort chosen in each analysis was done so by balancing the number of events and sample size with the p-value being representative of all significant cells.

Chapter 4 Results

A substantial portion of the results is described in or based on following publications:

Fried DV, Mawlawi O, Zhang L, et al. Stage III non-small cell lung cancer: Prognostic value of FDG PET quantitative image features combined with clinical prognostic factors. *Radiology* doi: 10.1148/radiol.2015142920. Published online July 15, 2015. ©Radiological Society of North America.

Fried DV, Tucker SL, Zhou S, et al. Prognostic Value and Reproducibility of Pretreatment CT Texture Features in Stage III Non-Small Cell Lung Cancer. *International Journal of Radiation Oncology, Biology, and Physics* doi: 10.1016/j.ijrobp.2014.07.020. Volume 90, Issue 4, Pages 834-842. 2014. ©Elsevier.

The permissions for reuse of these materials were obtained from both the Radiological Society of North America and Elsevier B V.

4.1 Results of Specific Aim 1: Analysis of CT-based Quantitative Image Features

Specific Aim 1 examined CT-based quantitative image features for prognostic value, reproducibility, and value of each CT image type in cohort 1.

4.1.1 Results for Project 1.1: Quantify the impact of adding CT-based quantitative image features to outcome models containing only CPFs including and excluding GTV

Sixty-six QIFs were assessed in each fold of cross-validation. These features are shown in

Table 5. The predictive Kaplan-Meier curves generated from the cross-validated predictions using CPFs excluding GTV, CPFs including GTV, and CPFs including GTV and QIFs are shown. The p-values in the lower left of each figure represent the p-value of the associated log-rank test. Figure 6, Figure 7, and Figure 8 illustrate the stratification in overall survival using the aforementioned covariate combination types.

Table 5. Extracted Quantitative Image Features for Cohort 1

| Intensity Histogram (IHIST)[^] | Absolute Gradient (Grad)⁻ | Nearest Gray Tone Difference Matrix (NGTDM)⁼ | Co-Occurrence Matrix (COM)⁺ | Laplacian of Gaussian Filtration Metrics (LoG)[*] |
|---|--|--|--|---|
| Mean Variance Skewness Kurtosis Entropy Uniformity | Mean Variance Skewness Kurtosis % non-zero | Coarseness Contrast Busyness | Energy Contrast Correlation Sum of Squares Inv. Diff. Moment Sum Average Sum Variance Sum Entropy Entropy Diff. Entropy Infomc1 Infomc2 | Mean Uniformity Standard Deviation Entropy |

[^]- Histograms and gradient images were generated by first converting CT into 8-bit image (i.e. bins of 16 HU)

⁼NGTDM were computed using a neighborhood = 1 on the converted 8-bit CT

⁺COM were computed on the 8-bit CT. Features were averaged across all 3D directions.

^{*} Sigma values used for the Laplacian of Gaussian Filter of: 1.0, 1.5, 1.8, 2.0, and 2.5 for the largest axial (LA) slice and for the entire tumor with a filter size of 11 voxels.

This table has been reused with the permission of the original publisher from the following publication: Fried DV, Tucker SL, Zhou S, Liao Z, Mawlawi O, Ibbott G, Court LE. Prognostic Value and Reproducibility of Pretreatment CT Texture Features in Stage III Non-Small Cell Lung Cancer. *International Journal of Radiation Oncology, Biology, and Physics* doi: 10.1016/j.ijrobp.2014.07.020. Volume 90, Issue 4, Pages 834-842. 2014. ©Elsevier.

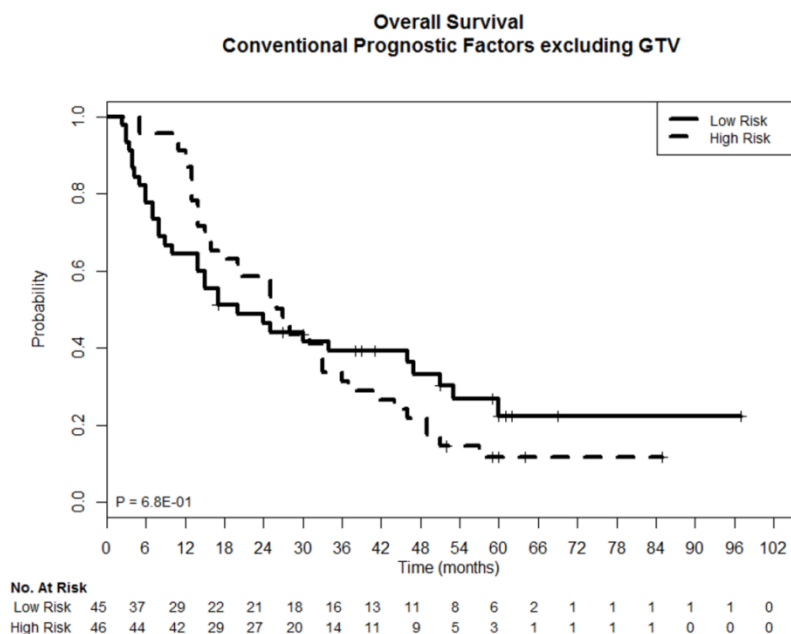


Figure 6. Overall Survival Comparing High Risk versus Low Risk Patients Using Models Incorporating CPFs Excluding GTV.

This figure has been reused with the permission of the original publisher from the following publication: Fried DV, Tucker SL, Zhou S, Liao Z, Mawlawi O, Ibbott G, Court LE. Prognostic Value and Reproducibility of Pretreatment CT Texture Features in Stage III Non-Small Cell Lung Cancer. *International Journal of Radiation Oncology, Biology, and Physics* doi: 10.1016/j.ijrobp.2014.07.020. Volume 90, Issue 4, Pages 834-842. 2014. ©Elsevier.

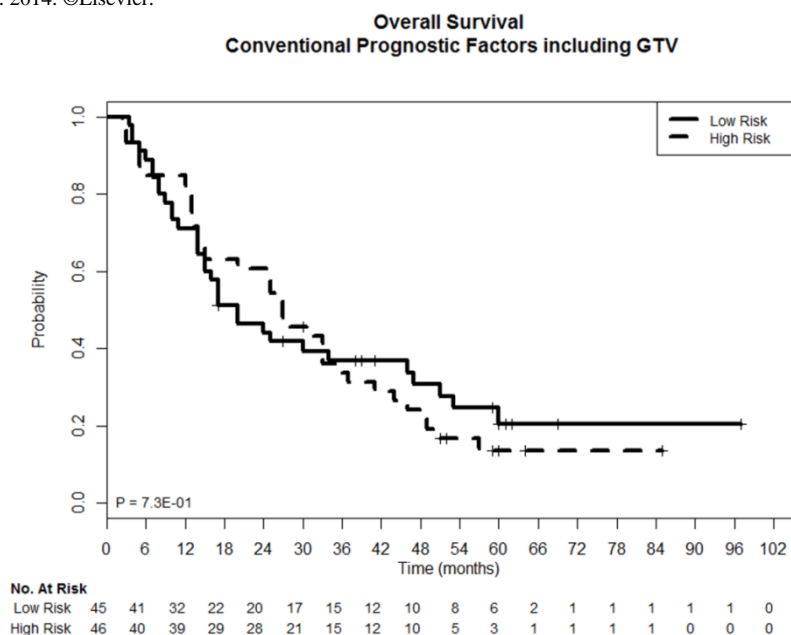


Figure 7. Overall Survival Comparing High Risk versus Low Risk Patients Using Models Incorporating CPFs Including GTV.

This figure has been reused with the permission of the original publisher from the following publication: Fried DV, Tucker SL, Zhou S, Liao Z, Mawlawi O, Ibbott G, Court LE. Prognostic Value and Reproducibility of Pretreatment CT Texture Features in Stage III Non-Small Cell Lung Cancer. *International Journal of Radiation Oncology, Biology, and Physics* doi: 10.1016/j.ijrobp.2014.07.020. Volume 90, Issue 4, Pages 834-842. 2014. ©Elsevier.

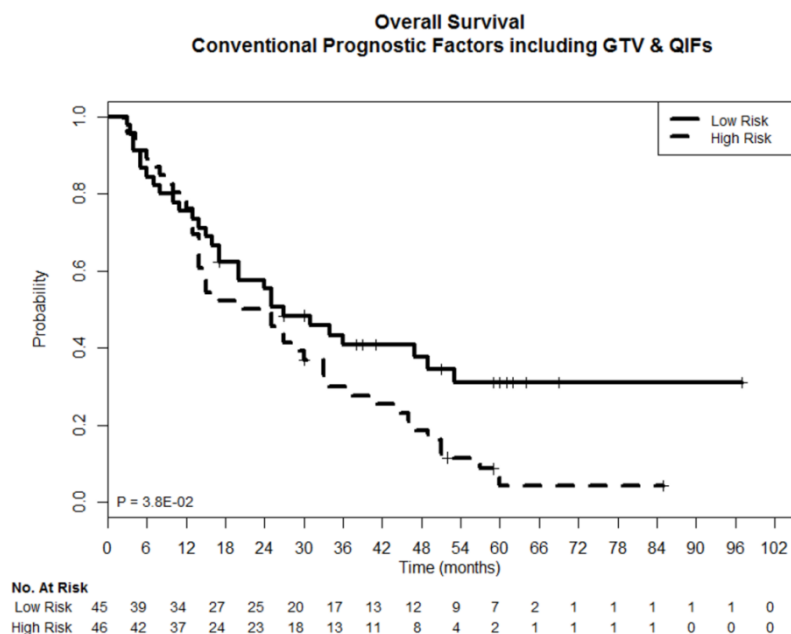


Figure 8. Overall Survival Comparing High Risk versus Low Risk Patients Using Models Incorporating CPFs Including GTV and CT-Based QIFs.

This figure has been reused with the permission of the original publisher from the following publication: Fried DV, Tucker SL, Zhou S, Liao Z, Mawlawi O, Ibbott G, Court LE. Prognostic Value and Reproducibility of Pretreatment CT Texture Features in Stage III Non-Small Cell Lung Cancer. *International Journal of Radiation Oncology, Biology, and Physics* doi: 10.1016/j.ijrobp.2014.07.020. Volume 90, Issue 4, Pages 834-842. 2014. ©Elsevier.

Figure 9, Figure 10, and Figure 11 illustrate the stratification in local-regional control.

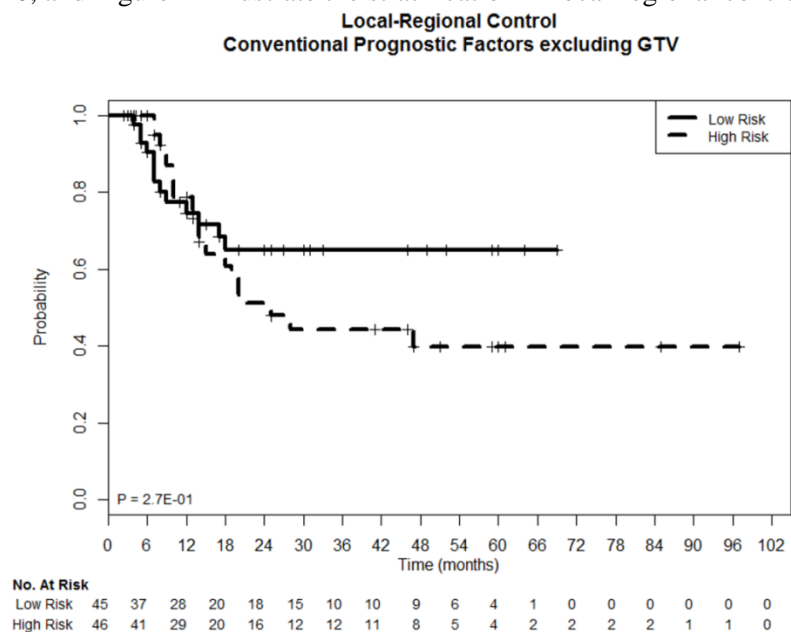


Figure 9. Local-Regional Control Comparing High Risk versus Low Risk Patients Using Models Incorporating CPFs Excluding GTV.

This figure has been reused with the permission of the original publisher from the following publication: Fried DV, Tucker SL, Zhou S, Liao Z, Mawlawi O, Ibbott G, Court LE. Prognostic Value and Reproducibility of Pretreatment CT Texture Features in Stage III Non-

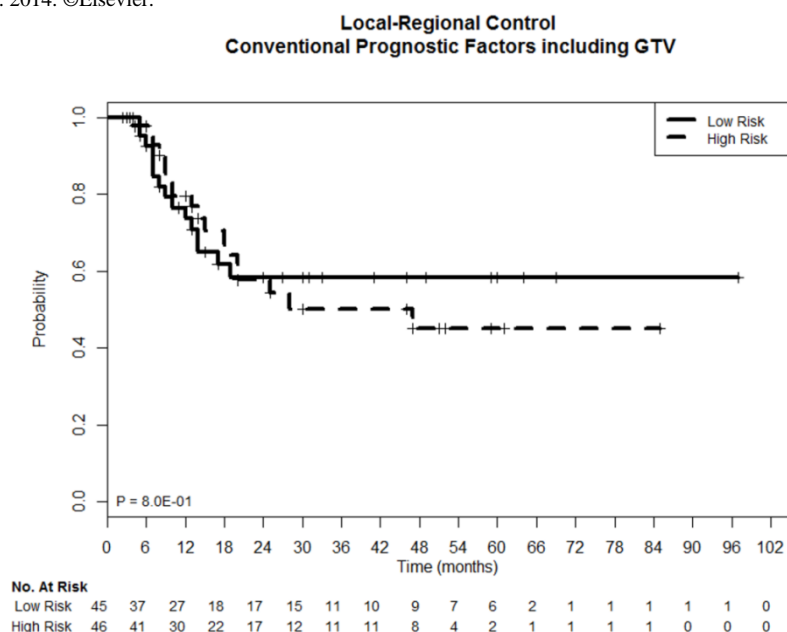


Figure 10. Local-Regional Control Comparing High Risk versus Low Risk Patients Using Models Incorporating CPFs Including GTV.

This figure has been reused with the permission of the original publisher from the following publication: Fried DV, Tucker SL, Zhou S, Liao Z, Mawlawi O, Ibbott G, Court LE. Prognostic Value and Reproducibility of Pretreatment CT Texture Features in Stage III Non-Small Cell Lung Cancer. *International Journal of Radiation Oncology, Biology, and Physics* doi: 10.1016/j.ijrobp.2014.07.020. Volume 90, Issue 4, Pages 834-842. 2014. ©Elsevier.

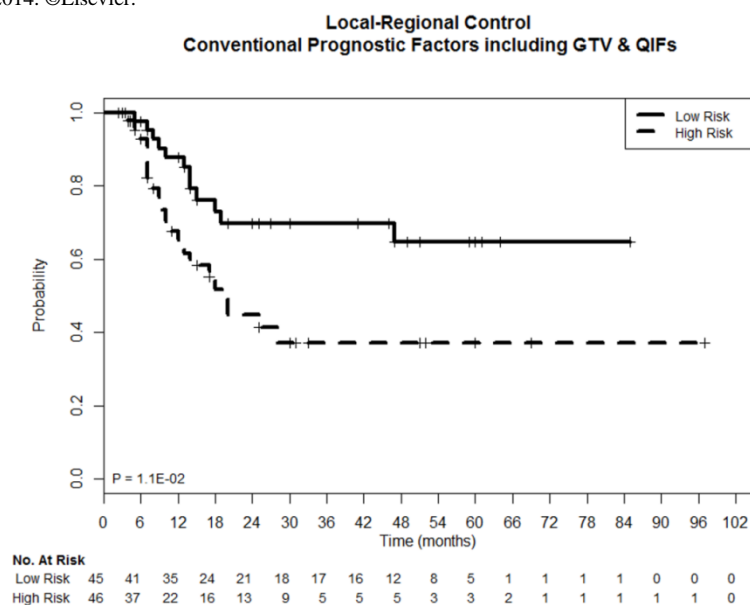


Figure 11. Local-Regional Control Comparing High Risk versus Low Risk Patients Using Models Incorporating CPFs Including GTV And CT-Based QIFs.

This figure has been reused with the permission of the original publisher from the following publication: Fried DV, Tucker SL, Zhou S, Liao Z, Mawlawi O, Ibbott G, Court LE. Prognostic Value and Reproducibility of Pretreatment CT Texture Features in Stage III Non-Small Cell Lung Cancer. *International Journal of Radiation Oncology, Biology, and Physics* doi: 10.1016/j.ijrobp.2014.07.020. Volume 90,

Figure 12, Figure 13, and Figure 14 illustrate the stratification in freedom from distant metastases.

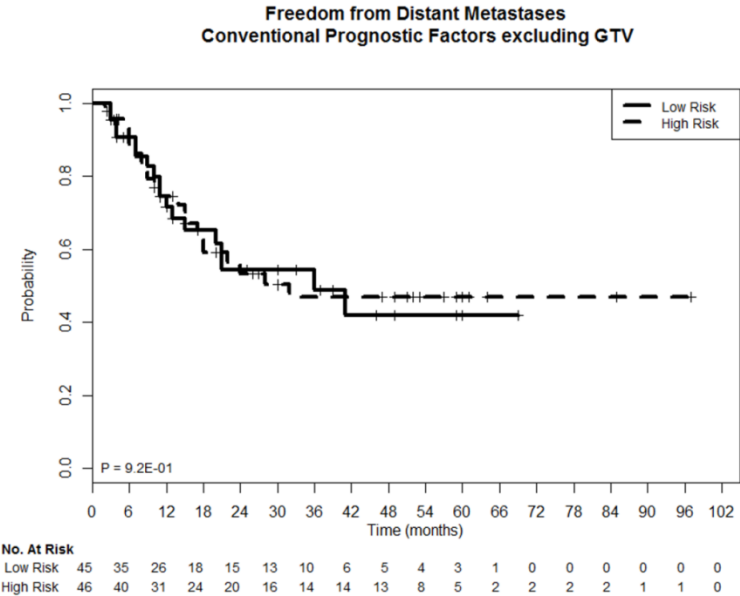


Figure 12. Freedom from Distant Metastases Comparing High Risk versus Low Risk Patients Using Models Incorporating CPFs Excluding GTV.

This figure has been reused with the permission of the original publisher from the following publication: Fried DV, Tucker SL, Zhou S, Liao Z, Mawlawi O, Ibbott G, Court LE. Prognostic Value and Reproducibility of Pretreatment CT Texture Features in Stage III Non-Small Cell Lung Cancer. *International Journal of Radiation Oncology, Biology, and Physics* doi: 10.1016/j.ijrobp.2014.07.020. Volume 90, Issue 4, Pages 834-842. 2014. ©Elsevier.

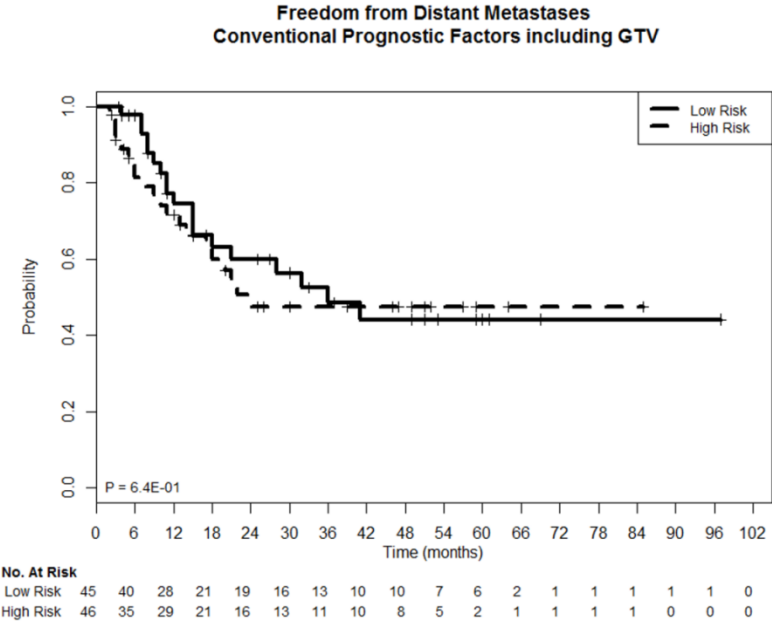


Figure 13. Freedom from Distant Metastases Comparing High Risk versus Low Risk Patients Using Models Incorporating CPFs Including GTV.

This figure has been reused with the permission of the original publisher from the following publication: Fried DV, Tucker SL, Zhou S, Liao Z, Mawlawi O, Ibbott G, Court LE. Prognostic Value and Reproducibility of Pretreatment CT Texture Features in Stage III Non-

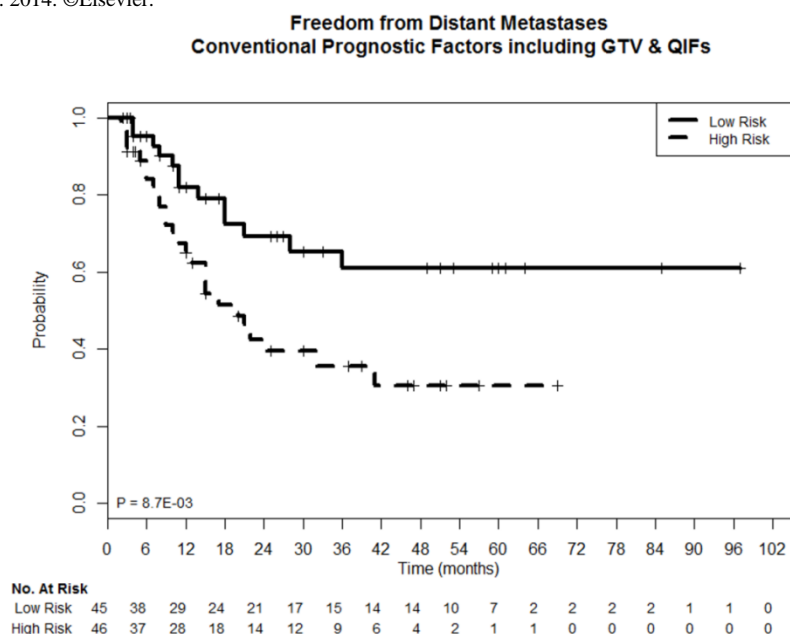


Figure 14. Freedom from Distant Metastases Comparing High Risk versus Low Risk Patients Using Models Incorporating CPFs Including GTV And CT-Based Qifs.

This figure has been reused with the permission of the original publisher from the following publication: Fried DV, Tucker SL, Zhou S, Liao Z, Mawlawi O, Ibbott G, Court LE. Prognostic Value and Reproducibility of Pretreatment CT Texture Features in Stage III Non-Small Cell Lung Cancer. *International Journal of Radiation Oncology, Biology, and Physics* doi: 10.1016/j.ijrobp.2014.07.020. Volume 90, Issue 4, Pages 834-842. 2014. ©Elsevier.

Across all outcomes (overall survival, local-regional control, and freedom from distant metastases), stratification was improved by including CT-based QIFs to models using CPFs alone or CPFs including GTV. No stratification was significant prior to adding the CT-based QIFs according to the log rank test ($p > 0.05$) and stratifications across all outcomes were significant after including CT-based QIFs ($p < 0.05$).

Overall survival stratification was much less compared to other two outcomes that were assessed. The separation between high and low risk groups did not appear until after 36 months. At this time point, the patient numbers in the high and low risk groups were low (11 and 13, respectively). While statistically different, the amplitude of difference between risk groups was small.

For both local-regional control and freedom from distant metastases, separation between risk groups appeared almost immediately and was much higher in amplitude than what was observed for

overall survival. Since models for each outcome were generated independently (i.e. patients modeled as being low risk in terms of overall survival were not necessarily the same patients modeled as low risk in terms of local-regional control or freedom from distant metastases), one can only infer that the QIFs appeared more prognostic for disease failure (locally and distantly) than patient survival.

Figure 15, Figure 16, and Figure 17 show the concordance indices for overall survival, local-regional control, and freedom from distant metastases, respectively using the multiple time point methodology described in section 3.7.3 Concordance Index at Multiple Time Points.

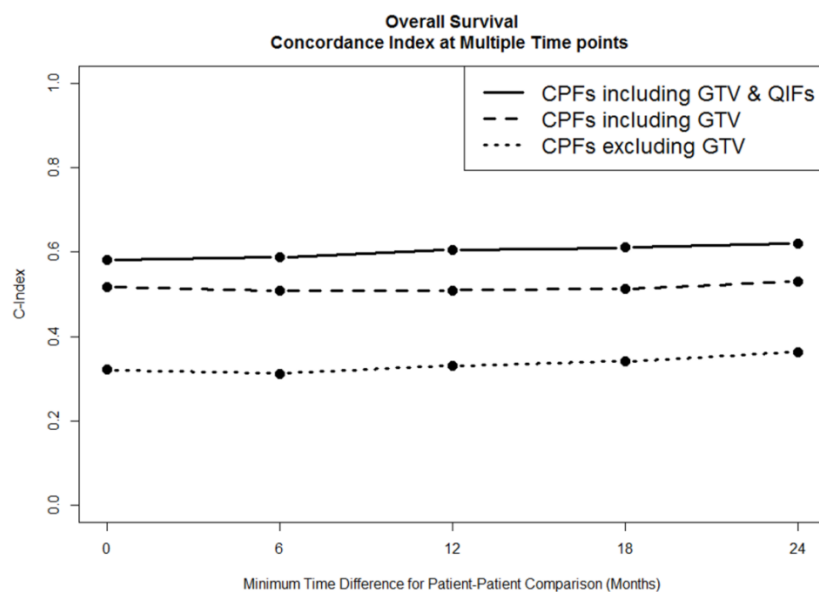


Figure 15. Concordance Indices for Overall Survival Predictions Using Minimum Outcome Differences of 6, 12, 18, and 24 Months.

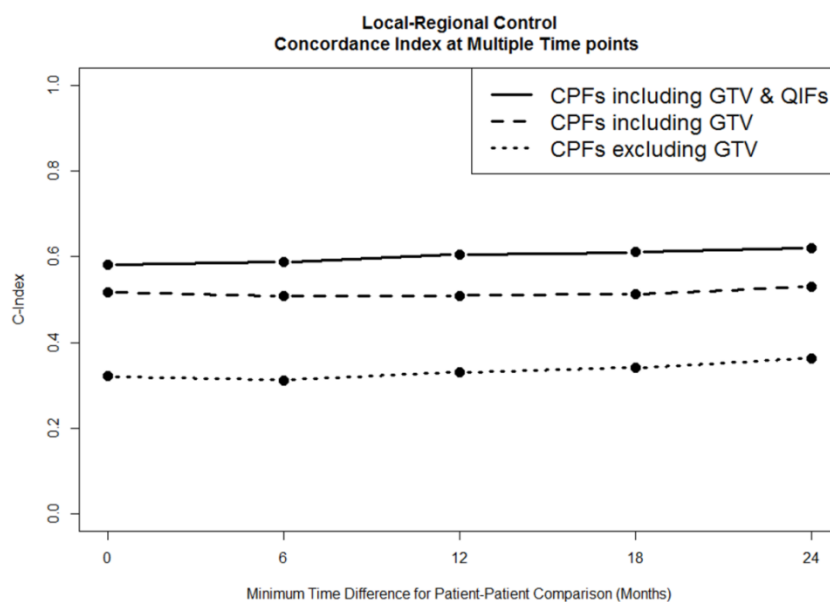


Figure 16. Concordance Indices for Local-Regional Control Predictions Using Minimum Outcome Differences Of 6, 12, 18, And 24 Months.

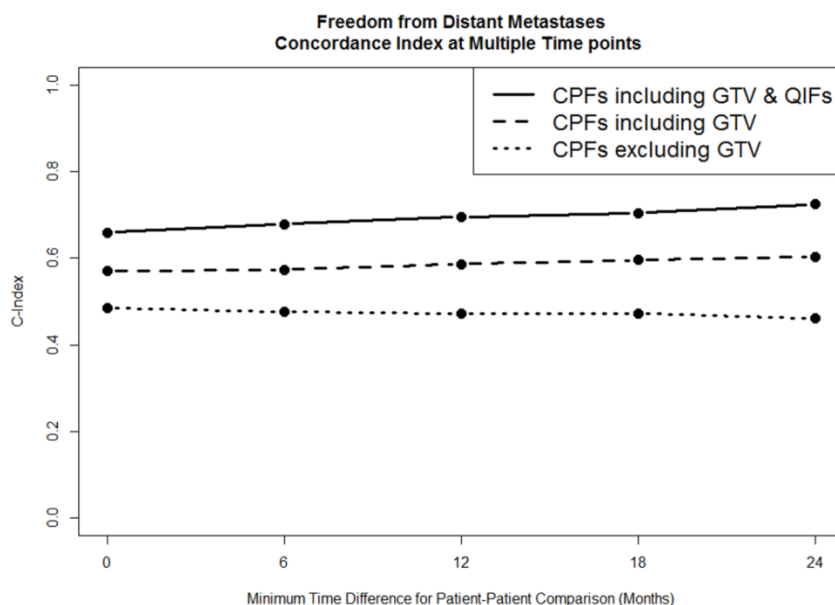


Figure 17. Concordance Indices for Freedom from Distant Metastases Predictions Using Minimum Outcome Differences of 6, 12, 18, And 24 Months.

For overall survival, local-regional control, and freedom from distant metastases predictions generated using CPFs including GTV and QIFs resulted in superior stratification of patients. Furthermore, c-indices at every time point were greater for predictions generated using CPFs including GTV and QIFs. The inclusion of GTV to the other CPFs resulted in improved c-indices in all outcomes but the inclusion of QIFs always resulted in additional improvement. Furthermore, it was observed that the c-indices increased with an increasing outcome separation between patients (i.e. for later time points). This inherently makes sense since one would hope a model would be able to predict patient outcomes more effectively when the outcomes are substantially different (e.g. patient survivals of 36 months versus 6 months) rather than close (e.g. patient survivals of 12 months versus 11 months). Larger increases in c-indices for later time points were seen in models utilizing CT-based QIFs.

Table 6 illustrates the models developed from the cross-validation methodology using CPFs including GTV and QIFs.

Table 6. Outcome Models for Covariate Combinations in Cohort 1

| Covariate | OS Model | | LRC Model | | FFDM Model | |
|---|--------------------|----------------|--------------------|----------------|--------------------|----------------|
| | <i>Coefficient</i> | <i>p-value</i> | <i>Coefficient</i> | <i>p-value</i> | <i>Coefficient</i> | <i>p-value</i> |
| CPFs: | | | | | | |
| Age (65> vs ≤65) | 0.37 | 0.19 | NI | - | NI | - |
| ECOG (0/1 vs 2) | 0.49 | 0.1 | NI | - | NI | - |
| Histology (SCC vs Other) | 0.31 | 0.31 | NI | - | NI | - |
| Gender (Male vs Female) | -0.2 | 0.46 | NI | - | -1.34 | <0.01 |
| GTV | 0.005 | <0.01 | NI | - | 0.002 | 0.35 |
| QIFs: | | | | | | |
| CE-CT | | | | | | |
| LoG_Average (LA, σ=1) | 0.45 | 0.01 | | | 0.41 | 0.03 |
| LOG_Average (σ=1) | NI | - | 0.59 | <0.01 | NI | - |
| IHIST_kurtosis | -0.05 | 0.13 | NI | - | -0.20 | 0.02 |
| NGTDM_busyness | NI | - | NI | - | 108.7 | 0.21 |
| COM_infomc1 | NI | - | NI | - | | |
| AVG-CT | | | | | | |
| LoG_SD (σ=1) | 0.04 | 0.11 | NI | - | 0.15 | <0.01 |
| LoG_SD (LA, σ=1.5) | NI | - | NI | - | 0.056 | 0.43 |
| LoG_Uniformity (LA, σ=2.5) | 1.56 | 0.04 | NI | - | NI | - |
| T50-CT | | | | | | |
| GRAD_kurtosis | NI | - | NI | - | 0.202 | <0.01 |
| LOG_Average _σ (LA, σ=1.5) | -0.29 | 0.12 | NI | - | NI | - |
| COM_sosvariance | 0.003 | 0.03 | NI | - | NI | - |
| LoG_Uniformity (LA, σ=1.5) | NI | - | -2.3 | 0.05 | NI | - |

Abbreviations: NI-not included in model, SCC-squamous cell carcinoma, GTV-gross tumor volume, SD-standard deviation, LA-largest axial slice. This table has been reused with the permission of the original publisher from the following publication: Fried DV, Tucker SL, Zhou S, Liao Z, Mawlawi O, Ibbott G, Court LE. Prognostic Value and Reproducibility of Pretreatment CT Texture Features in Stage III Non-Small Cell Lung Cancer. *International Journal of Radiation Oncology, Biology, and Physics* doi: 10.1016/j.ijrobp.2014.07.020. Volume 90, Issue 4, Pages 834-842. 2014. ©Elsevier.

Across all outcomes, features from the CE-CT (specifically the LOG_Average) appeared to be the most consistent and significant QIF. The LOG_Average was significant in both local-regional control and freedom from distant metastases ($p < 0.01$ and 0.03 , respectively). For overall survival,

the same feature but only used on the largest axial slice was significant ($p = 0.01$). The LOG_Averages calculated on the entire tumor versus the largest axial slice were highly correlated (Pearson correlation coefficient = 0.91, $p < 0.01$) and thus would be unlikely to be selected within the same model framework due to the ability of the penalized algorithm to handle covariate collinearity.

To ensure these results were not due to over fitting or the ratio of the number of features being analyzed to the number of patients, a permutation test was performed (see 3.7.2 Permutation Test and Impact of Feature Reproducibility on Predictions (Cohort 1)). The log-rank statistic derived when outcomes were permuted with respect to the QIFs and CPFs (i.e., random data) was greater than the true, non-randomized log-rank statistic in 11/200 ($p = 0.055$), 0/200 ($p < 0.005$), and 1/200 ($p = 0.005$) for OS, LRC, and FFDM, respectively.

4.1.2 Results for Project 1.2: Quantify the reproducibility of FDG-PET-based quantitative image features using “pseudo” test-retest scans

Data from test-retest scans from 10, 10, and 13 independent patients for the AVG-CT, T50-CT, and CE-CT, respectively, were used for our assessment of reproducibility. We found that 85%,(56/66), 75%,(50/66), and 23%,(15/66) of texture features had a CCC>0.9 for features generated from T50-CT, Average-CT, and CE-CT, respectively.

Incorporating reproducibility within our models yielded 80.4% (SD=3.7), 78.3 (SD=4.0), and 78.8% (SD=3.9) classification reproducibility in terms of OS, LRC, and FFDM, respectively. Figure 18 illustrates an example iteration where we compared the predicted outcome with reproducibility to the original predicted outcome in terms of FFDM and calculate the classification reproducibility.

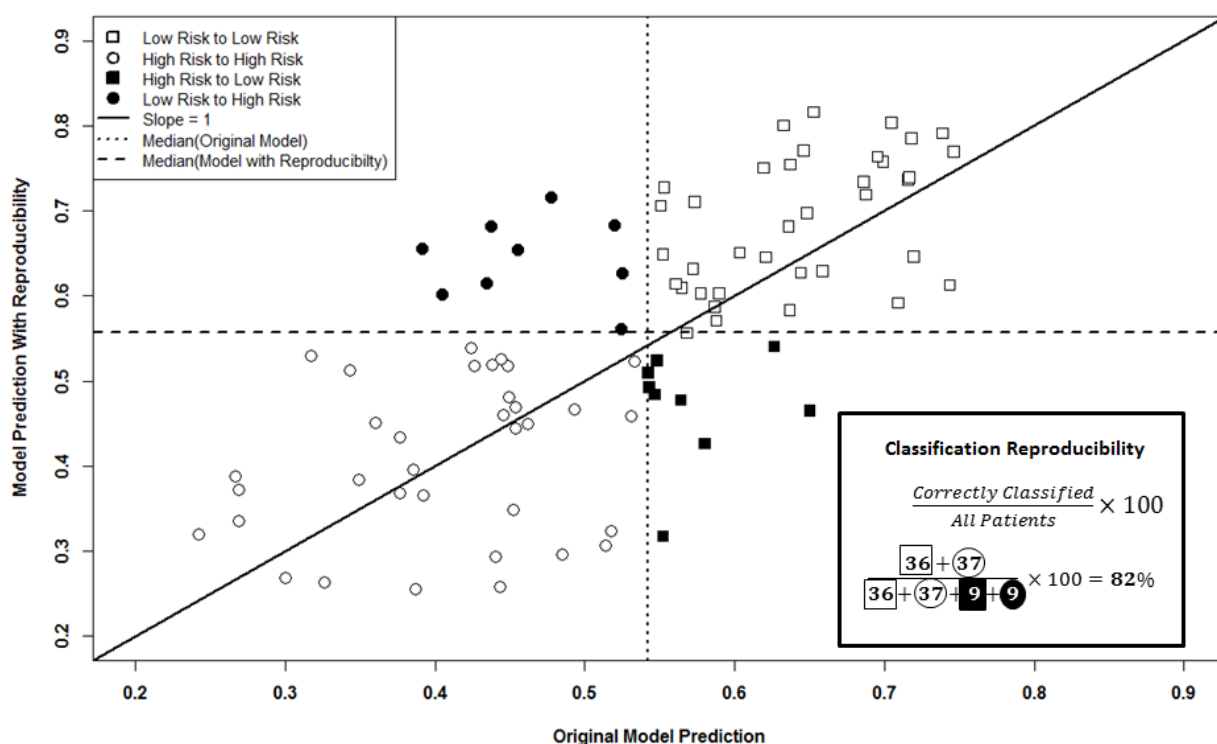


Figure 18. Example from Single Simulation of the Impact of Texture Feature Reproducibility on FFDM Estimates. Outcome Prediction from Original Model (X-Axis) Compared to Prediction Incorporation of the Variation in QIFs From Test/Retest Scans (Y-Axis).

This figure has been reused with the permission of the original publisher from the following publication: Fried DV, Tucker SL, Zhou S, Liao Z, Mawlawi O, Ibbott G, Court LE. Prognostic Value and Reproducibility of Pretreatment CT Texture Features in Stage III Non-Small Cell Lung Cancer. *International Journal of Radiation Oncology, Biology, and Physics* doi: 10.1016/j.ijrobp.2014.07.020. Volume 90, Issue 4, Pages 834-842. 2014. ©Elsevier.

4.1.3 Results for Project 1.3: Quantify the prognostic value of adding CE-CT-based quantitative image features to outcome models containing only CPFs

The 249 patients in Cohort 2 were used for analysis of prognostic value of CE-CT QIFs.

Forty three QIFs were assessed during cross-validation. These features are shown in Table 7.

Table 7. Extracted Quantitative Image Features for Cohort 2

| Intensity Histogram (IHIST)* | Nearest Gray Tone Difference Matrix (NGTDM)* | Co-Occurrence Matrix (COM)* | Laplacian of Gaussian Filtration Metrics (LoG)* | Volume/Morphologic Characteristics |
|-------------------------------------|---|------------------------------------|--|---|
| Mean | Coarseness | Energy | Mean | Volume |
| Standard Deviation | Contrast | Contrast | Standard Deviation | Surface Area |
| Skewness | Busyness | Correlation | | Air Volume |
| Kurtosis | Complexity | Sum of Squares | | Tissue Volume |
| Entropy | | Inv. Diff. Moment | | Necrosis Volume |
| Uniformity | | Sum Average | | Vessel Volume |
| | | Sum Variance | | Air Percentage |
| | | Sum Entropy | | Tissue Percentage |
| | | Entropy | | Necrosis Percentage |
| | | Diff. Entropy | | Vessel Percentage |
| | | Infomc1 | | |
| | | Infomc2 | | |

*The same parameters were used as described in Table 5 for each type of extracted QIFs

The predictive Kaplan-Meier curves generated from the cross-validated predictions using CPFs excluding GTV, CPFs including GTV, and CPFs including GTV and QIFs are shown in Figure 19, Figure 20, Figure 21, respectively. Three clusters based on k-means were used to stratify patients into low, medium, and high risk groups based for overall survival. Local-regional control and freedom from distant metastases were investigated; however CE-CT based QIFs did not appear to be prognostic for these outcomes. A comparison of the c-indices is seen in Figure 22.

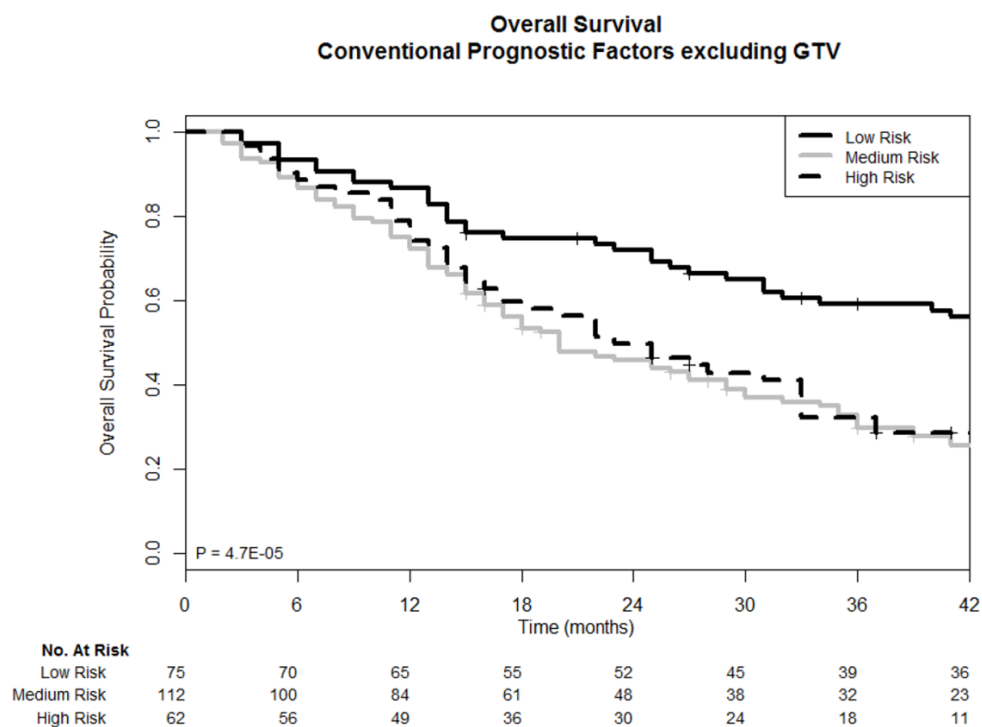


Figure 19. Overall Survival Comparing High Risk, Medium Risk, and Low Risk Patients Using Models Incorporating CPFs Excluding GTV

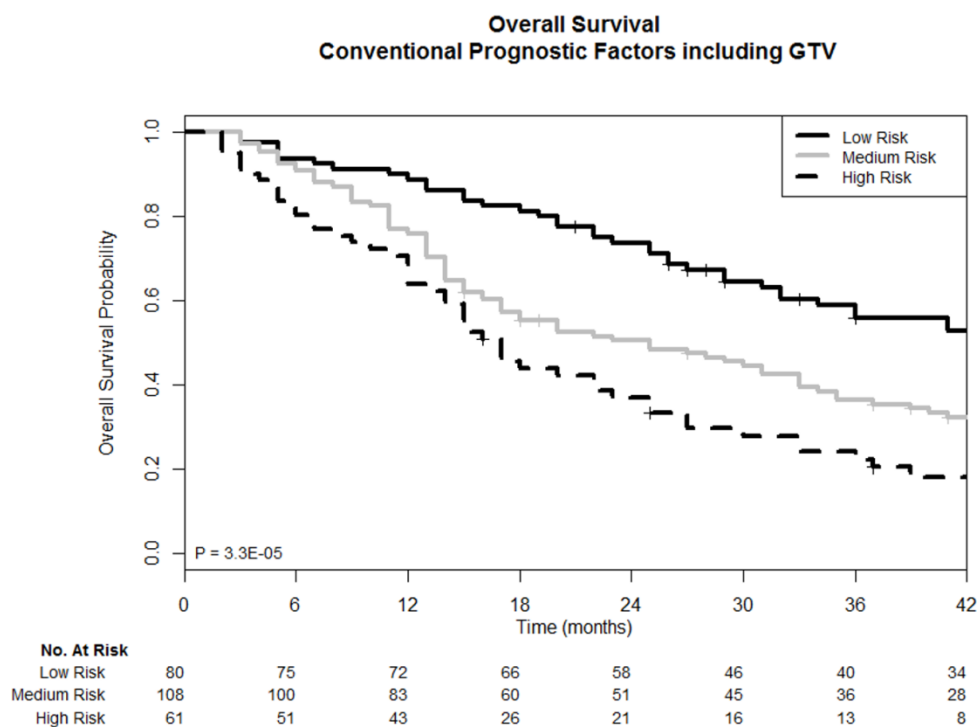


Figure 20. Overall Survival Comparing High Risk, Medium Risk, and Low Risk Patients Using Models Incorporating CPFs Including GTV

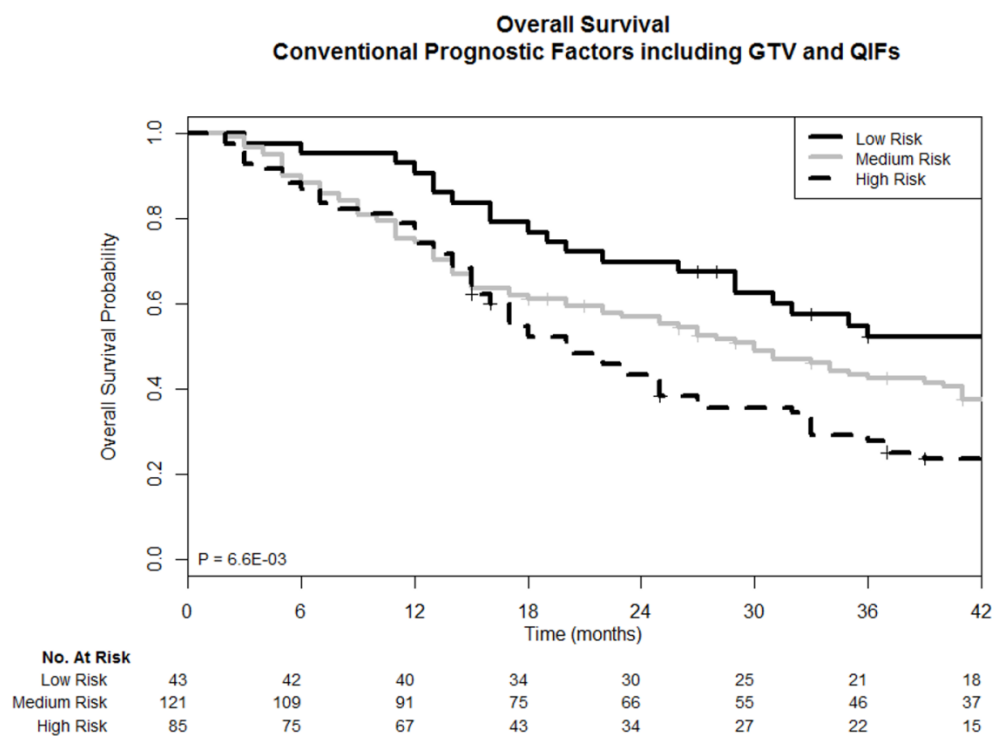


Figure 21. Overall Survival Comparing High Risk, Medium Risk, and Low Risk Patients Using Models Incorporating CPFs Including GTV And CE-CT Based QIFs

The addition of GTV to CPFs improved stratification (Figure 19 versus 20) and the c-indices at every time point (Figure 22). However, the addition of QIFs did not improve stratification nor the c-index beyond what was obtained using CPFs including GTV (Figure 21). These results are most likely due to the dominant prognostic feature being GTV. Table 8 illustrates the models developed using the covariates that were included in greater than 50% of the cross-validation folds. For the model using CPFs including GTV and QIFs, the significance of GTV was several orders of magnitude lower than any other factor. This was not observed in the analysis of Cohort 1. While significant, GTV significance was observed to be on a similar order of magnitude as other features in Cohort 1.

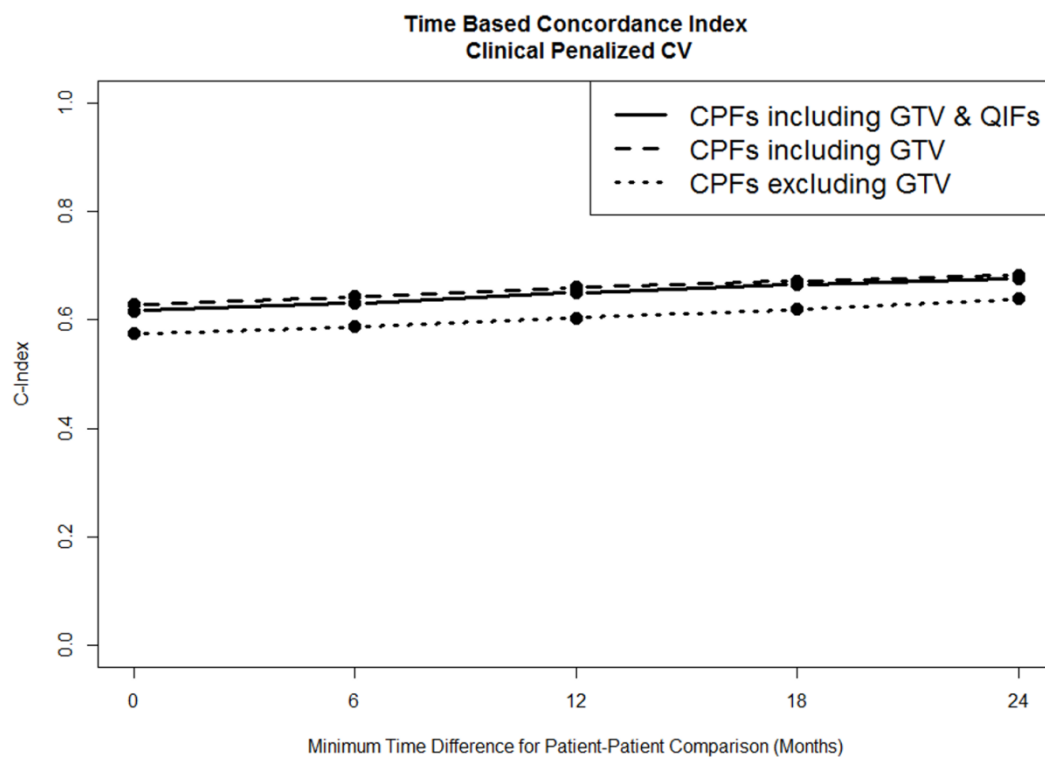


Figure 22. Concordance Indices for Overall Survival in Cohort 2 (CE-CT QIFs Only) Using Minimum Outcome Differences of 6, 12, 18, and 24 Months.

Table 8. Outcome Models for Covariate Combinations in Cohort 2

| Covariates | CPFs excluding GTV | p- value | CPFs including GTV | p-value | CPFs including GTV and QIFs | p- value |
|-------------------------------|--------------------------|-------------|--------------------------|---------|--------------------------------------|-------------|
| CPFs: | | | | | | |
| Age (continuous) | NI | - | 0.009 | 0.32 | 0.012 | 0.22 |
| GTV (log2) | NI | - | 0.296 | <0.01 | 0.340 | <0.01 |
| Gender (Male vs Female) | 0.151 | 0.33 | NI | - | NI | - |
| Histology (SCC vs Other) | 0.138 | 0.37 | NI | - | NI | - |
| Induction | -0.56 | <0.01 | -0.400 | 0.04 | -0.39 | 0.04 |
| KPS (<90 vs ≥90) | 0.354 | <0.01 | 0.282 | 0.02 | 0.25 | 0.04 |
| N Stage (N2/3 vs N0/1) | 0.628 | 0.04 | 0.826 | 0.01 | 0.84 | <0.01 |
| Overall Stage (3b vs 3a) | 0.259 | 0.09 | 0.200 | 0.20 | 0.21 | 0.17 |
| QIFs: | | | | | | |
| Global Uniformity | NI | - | NI | - | -6.11 | 0.02 |
| COM sum variance | NI | - | NI | - | -0.0002 | 0.10 |
| Percent Air | NI | - | NI | - | 1.26 | 0.53 |

Abbreviations: NI=not included in model; SCC=squamous cell carcinoma; GTV=gross tumor volume; KPS = Karnofsky performance status; SD= standard deviation; LA=largest axial slice

The addition of GTV to the Cox model containing induction, KPS, N stage, and overall stage (i.e., the CPFs excluding GTV) led to a statistically significant improvement in model fit ($p = 8 \times 10^{-6}$). Adding the QIFs from Table 7 to the model using the CPFs including GTV also led to a statically significant improvement in model fit ($p = 0.027$).

While comparisons of stratification on a Kaplan-Meier plot or c-indices are reasonable visual ways to assess for prognostic value, performing a likelihood ratio test on nested models is seen as the gold standard. Ultimately, it appears that QIFs from CE-CT add prognostic value but are not *substantially* adding information not accounted for from GTV or CPFs. Initially, it was somewhat surprising that the LOG_Average feature, which was significant in the Cohort 1 analysis, was not selected in this

analysis. However, LOG_Average and COM sumvariance were significantly correlated (Pearson correlation coefficient = 0.91, $p < 0.01$).

4.2 Results of Specific Aim 2: Analysis of FDG-PET-based Quantitative Image Features

Specific Aim 2 examined PET-based quantitative image features for prognostic value, reproducibility, and volumetric stability.

4.2.1 Results for Project 2.1: Quantify the impact of adding FDG-PET-based quantitative image features to outcome models containing only CPFs

Twenty-eight QIFs were assessed in each fold of cross-validation. These features are shown in Table 9.

Table 9. Extracted Quantitative Image Features for Cohort 3 (pretreatment FDG-PET)

| Intensity Histogram (IHIST)⁻ | Co-Occurrence Matrix (COM)⁺ | Shape/Volume |
|--|---|----------------------|
| Mean* -- SUVmean | Contrast | Volume* -- MTV |
| Maximum* -- SUVmax | Correlation | Surface Area* |
| Peak* -- SUVpeak | Energy | Convex Hull Volume |
| Entropy | Homogeneity | Solidity |
| Uniformity | | |
| Standard Deviation | | Presence of Necrosis |
| Coefficient of Variation | | |
| Cumulative Histogram | | |

*These features were calculated for the primary, nodal disease, and total disease (primary plus nodal)

[^] This features was only calculated for the nodal disease, and total disease (primary plus nodal)

⁻ Entropy, Uniformity, and cumulative histogram were generated using a bin size of 1 SUV. The other features used the raw SUV values

⁺A bin size of 1 SUV was used for COM features. COM features were averaged across all 2D directions using all axial slices of the ROI.

This table has been reused with the permission of the original publisher from the following publication: Fried DV, Mawlawi O, Zhang L, Fave X, Zhou S, Ibbott G, Liao Z, Court LE. Stage III non-small cell lung cancer: Prognostic value of FDG PET quantitative image features combined with clinical prognostic factors. *Radiology* doi: 10.1148/radiol.2015142920. Published online July 15, 2015. ©Radiological Society of North America.

The predictive Kaplan-Meier curves generated from the cross-validated predictions of overall survival using CPFs excluding GTV, CPFs including GTV, and CPFs including GTV and QIFs are shown in Figure 23, Figure 24, and Figure 25, respectively. Local-regional control and freedom from distant metastases were assessed; however FDG-PET based QIFs did not appear to be prognostic for these outcomes. The p-values in the lower left of each figure represent the p-value of the associated

log-rank test. The curves were divided into 5 risk groups via k-means clustering in order to demonstrate prediction calibration.

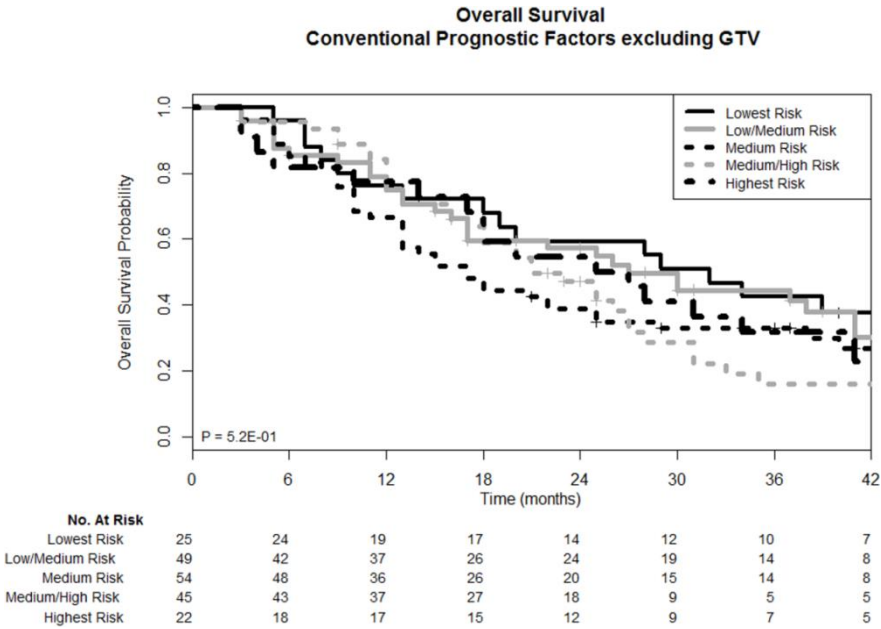


Figure 23. Overall Survival Comparing Various Risk Groups (Defined Using K-Means Clustering from Low Risk to High Risk) Using Models Incorporating CPFs Excluding GTV. This figure has been reused with the permission of the original publisher from the following publication: Fried DV, Mawlawi O, Zhang L, Fave X, Zhou S, Ibbott G, Liao Z, Court LE. Stage III non-small cell lung cancer: Prognostic value of FDG PET quantitative image features combined with clinical prognostic factors. *Radiology* doi: 10.1148/radiol.2015142920. Published online July 15, 2015. ©Radiological Society of North America.

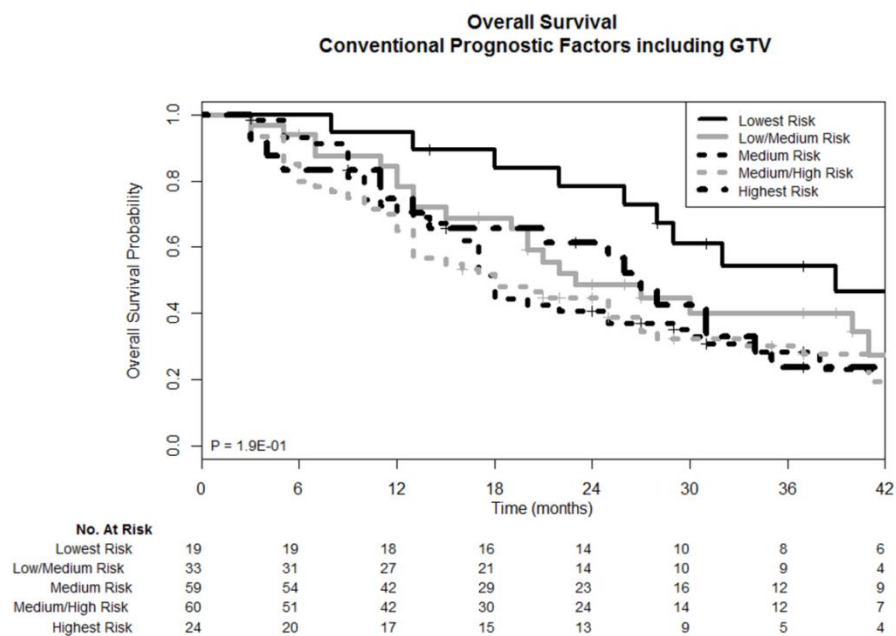


Figure 24. Overall Survival Comparing Various Risk Groups (Defined Using K-Means Clustering from Low Risk to High Risk) Using Models Incorporating Cpfs Including GTV.

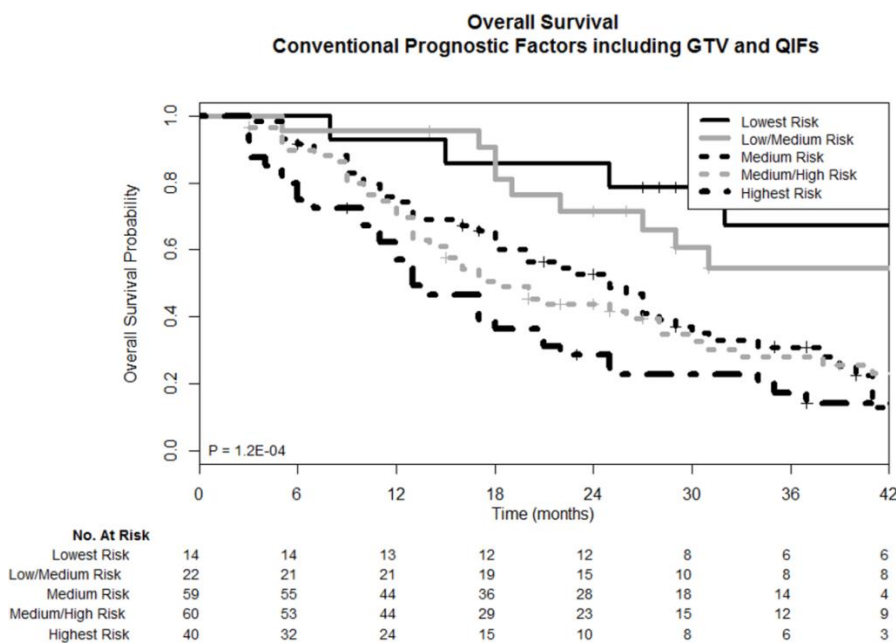


Figure 25. Overall Survival Comparing Various Risk Groups (Defined Using K-Means Clustering from Low Risk to High Risk) Using Models Incorporating Cpfs Including GTV And QIFs.

This figure has been reused with the permission of the original publisher from the following publication: Fried DV, Mawlawi O, Zhang L, Fave X, Zhou S, Ibbott G, Liao Z, Court LE. Stage III non-small cell lung cancer: Prognostic value of FDG PET quantitative image features combined with clinical prognostic factors. *Radiology* doi: 10.1148/radiol.2015142920. Published online July 15, 2015. ©Radiological Society of North America.

Figure 26 illustrates the concordance indices for overall survival.

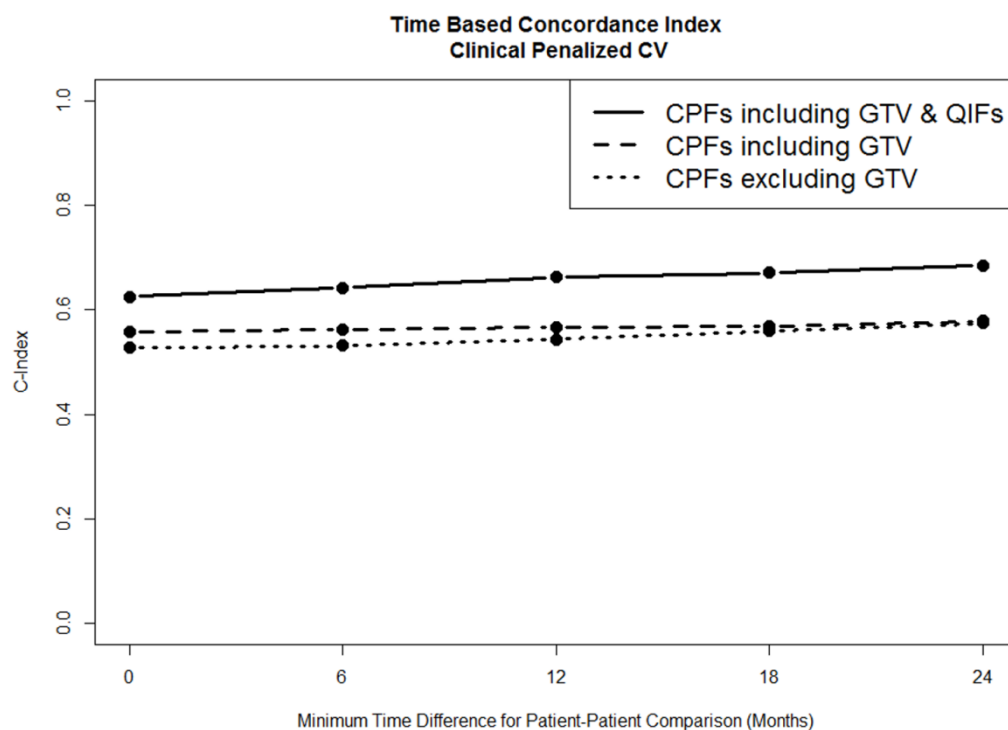


Figure 26. Concordance Indices for Overall Survival in Cohort 3 (FDG-PET Based QIFs) Using Minimum Outcome Differences of 6, 12, 18, And 24 Months.

This figure has been reused with the permission of the original publisher from the following publication: Fried DV, Mawlawi O, Zhang L, Fave X, Zhou S, Ibbott G, Liao Z, Court LE. Stage III non-small cell lung cancer: Prognostic value of FDG PET quantitative image features combined with clinical prognostic factors. *Radiology* doi: 10.1148/radiol.2015142920. Published online July 15, 2015. ©Radiological Society of North America.

For overall survival predictions generated using CPFs including GTV and QIFs resulted in superior stratification of patients. Furthermore, c-indices at every time point were greater for predictions generated using CPFs including GTV and QIFs (Figure 26). The inclusion of GTV to the other CPFs resulted in improved c-indices but the inclusion of QIFs always resulted in substantial further improvement.

Table 10 illustrates the models developed from the cross-validation methodology using CPFs including GTV and QIFs.

Table 10. Overall Survival Models for Covariate Combinations in Cohort 3

| Covariates | CPFs excluding GTV | | CPFs including GTV | | CPFs including GTV and QIFs | |
|--------------------------|--------------------|---------|--------------------|---------|-----------------------------|---------|
| | Coefficient | p-value | Coefficient | p-value | Coefficient | p-value |
| CPFs: | | | | | | |
| Age (continuous) | 0.027 | <0.01 | 0.029 | <0.01 | 0.027 | <0.01 |
| Induction | -0.226 | 0.24 | -0.130 | 0.52 | -0.138 | 0.49 |
| T Stage (T1/2 vs T3/4) | -0.200 | 0.25 | -0.286 | 0.11 | -0.198 | 0.31 |
| Gender (Male vs Female) | 0.525 | <0.01 | 0.506 | 0.01 | 0.467 | 0.02 |
| GTV (Log2) | NI | - | 0.196 | 0.02 | 0.225 | 0.01 |
| KPS (<90 vs ≥90) | 0.257 | 0.07 | 0.202 | 0.16 | 0.307 | 0.03 |
| Overall Stage (3b vs 3a) | 0.277 | 0.13 | 0.215 | 0.25 | NI | - |
| QIFs: | | | | | | |
| COM Energy | NI | - | NI | - | -7.23 | 0.05 |
| Solidity | NI | - | NI | - | -0.780 | <0.01 |

Abbreviations: NI-not included in model, GTV-gross tumor volume, COM-co-occurrence matrix, CPFs – conventional prognostic factors

This table has been reused with the permission of the original publisher from the following publication: Fried DV, Mawlawi O, Zhang L, Fave X, Zhou S, Ibbott G, Liao Z, Court LE. Stage III non-small cell lung cancer: Prognostic value of FDG PET quantitative image features combined with clinical prognostic factors. *Radiology* doi: 10.1148/radiol.2015142920. Published online July 15, 2015. ©Radiological Society of North America.

The addition of GTV to the Cox model containing induction, KPS, T stage, gender and overall stage (i.e., the CPFs excluding GTV) led to a statistically significant improvement in model fit ($p = 0.04$). Adding the QIFs from Table 9 to the model using the CPFs including GTV also led to a statically significant improvement in model fit ($p = 0.007$). Disease solidity (the volume of disease divided by the smallest convex volume that would be able to encompass all disease) along COM energy (a metric quantifying the uniformity of the SUV values within the primary) were the QIFs selected in all folds of cross-validation and were significantly (i.e., solidity; $p < 0.01$) or marginally significantly (i.e., COM energy; $p = 0.05$) associated with overall survival in the multivariate Cox model.

The presence of necrosis and percent of tumor exhibiting necrosis were examined in a separate analysis. Neither of these features was selected in cross-validation nor were they significant using univariate or multivariate Cox proportional hazards analyses ($p > 0.05$).

While comparisons of stratification on a Kaplan-Meier plot or c-indices are reasonable visual ways to assess for prognostic value, performing a likelihood ratio test on nested models is seen as the gold standard. Ultimately, it appears that QIFs from CECT add prognostic value not accounted for by CPFs or GTV.

4.2.2 Results for Project 2.2: Quantify the reproducibility of FDG-PET-based quantitative image features using “pseudo” test-retest scans

Cohort 6 was used to analyze the reproducibility of “pseudo” test-retest scans (i.e. patients having one scan taken at an outside institution followed by a scan taken at MD Anderson without any intervention between scans). The CCC was calculated for all 53 test-retest pairs, only pairs with different reconstruction dimensionality (2D vs 3D, $n = 40$), pairs with both scans having 3D reconstruction ($n = 10$), pairs with varying time differences between scans, pairs with $<25\%$ change in volume, and between scans acquired using different PET/CT models. The results from these analyses are shown in Table 11. Representative features were chosen from the various QIF types as shown in Table 9. Of note, the average CCC is more reasonable than expected at 0.78 using all scanners and imaging parameters. Furthermore, it can be seen that the QIF reproducibility improves when both scans are obtained using 3D reconstruction techniques versus a mix of 2D and 3D reconstruction (average CCC = 0.93 vs 0.72). Reproducibility also apparently worsens when scans are separated in time by more than 61 days (average CCC = 0.81 vs 0.62). Based on this data, standardization of PET acquisitions could lead to more reproducible QIFs and potentially enhance prediction models.

Table 11. CCC Values from “Pseudo” Test-Retest PET Scans

| | All | Reconstruction Type | | Time Difference (days) | | | Low Volume Change | Model Type | |
|-----------------------|--------------|---------------------|------------------|------------------------|----------------|--------------|-------------------|----------------|---------------|
| QIF | All (N = 53) | 2D - 3D (n = 40) | 3D - 3D (n = 10) | 0-30 (n = 16) | 31-60 (n = 25) | 61+ (n = 12) | <25% (n=24) | ST-RX (n = 15) | ST-STE (n=11) |
| Volume | 0.92 | 0.92 | 0.97 | 0.93 | 0.86 | 0.95 | 0.99 | 0.96 | 0.92 |
| Surface Area | 0.93 | 0.93 | 0.97 | 0.94 | 0.87 | 0.96 | 0.99 | 0.97 | 0.93 |
| Entropy | 0.82 | 0.8 | 0.95 | 0.79 | 0.87 | 0.62 | 0.84 | 0.88 | 0.75 |
| Max | 0.84 | 0.76 | 0.95 | 0.82 | 0.89 | 0.67 | 0.8 | 0.87 | 0.86 |
| Peak | 0.78 | 0.72 | 0.9 | 0.8 | 0.81 | 0.36 | 0.78 | 0.85 | 0.81 |
| Mean | 0.85 | 0.78 | 0.96 | 0.82 | 0.87 | 0.81 | 0.86 | 0.87 | 0.91 |
| Std | 0.87 | 0.8 | 0.96 | 0.83 | 0.93 | 0.68 | 0.85 | 0.89 | 0.92 |
| Uniformity | 0.77 | 0.78 | 0.9 | 0.72 | 0.82 | 0.52 | 0.74 | 0.84 | 0.54 |
| Kurtosis | 0.66 | 0.37 | 0.09 | 0.87 | 0.42 | -0.11 | 0.4 | 0.47 | 0.28 |
| Skewness | 0.57 | 0.54 | 0.26 | 0.76 | 0.46 | 0.35 | 0.58 | 0.47 | 0.43 |
| COMContrast | 0.93 | 0.62 | 0.97 | 0.88 | 0.98 | 0.5 | 0.64 | 0.87 | 0.95 |
| COMCorrelation | 0.77 | 0.7 | 0.93 | 0.71 | 0.82 | 0.74 | 0.69 | 0.88 | 0.5 |
| COMEnergy | 0.6 | 0.67 | 0.54 | 0.63 | 0.56 | 0.32 | 0.55 | 0.68 | 0.31 |
| COMHomogeneity | 0.7 | 0.69 | 0.87 | 0.53 | 0.79 | 0.61 | 0.66 | 0.8 | 0.68 |
| cumHistogram | 0.71 | 0.62 | 0.8 | 0.81 | 0.64 | 0.71 | 0.7 | 0.78 | 0.63 |
| | | | | | | | | | |
| Average (all metrics) | 0.78 | 0.72 | 0.93 | 0.81 | 0.82 | 0.62 | 0.74 | 0.87 | 0.75 |

4.2.3 Results for Project 2.3: Quantify the reproducibility of FDG-PET-based quantitative image features using retrospective reconstructions of phantom and patient data

The percent of voxels less than 1 or 2 SUV using the NEMA IEC phantom, as described in Section 3.6, is shown below in Table 12.

Table 12. Percent of Sphere Voxels with a Maximum Change in SUV <1 or < 2

| Scanner | Matrix Size | %Voxels < 1 SUV | %Voxels < 2 SUV |
|---------|-------------|-----------------|-----------------|
| VCT | 128 | 92 | 99 |
| | 256 | 75 | 94 |
| 710 | 128 | 80 | 98 |
| | 192 | 61 | 88 |
| | 256 | 62 | 90 |
| mCT | 128 | 91 | 98 |
| | 200 | 72 | 90 |
| | 256 | 69 | 88 |

QIF values were extracted from each of the three scanners using the various reconstruction parameters using fixed contours of the spheres. The median, mean, and standard deviation seen within each scanner and across all scanners were calculated and compared to the values from the patient data in terms of their standard deviations (Table 13). In addition, the same process was performed using 6 lesions from 5 different patients. The ratio of Cohort 3 patient standard deviation for each feature to the standard deviation observed in the 6 test lesions are shown in Table 14.

Table 13. Change in QIF Values due to Variation in Reconstruction Parameters and Comparison to Variation in Cohort 3 Patient QIF Values

| Scanner | Metric | Contrast | Correlation | Energy | Homogeneity | Uniformity | SUVmax | SUVmean | SD | Entropy |
|-------------------|--------------|----------|-------------|--------|-------------|------------|--------|---------|------|---------|
| All | min | 2.53 | 0.21 | 0.014 | 0.35 | 0.103 | 10.3 | 5.0 | 2.9 | 2.88 |
| All | max | 25.91 | 0.86 | 0.044 | 0.59 | 0.169 | 17.2 | 6.9 | 4.3 | 3.50 |
| All | median | 12.15 | 0.58 | 0.023 | 0.44 | 0.119 | 12.8 | 5.9 | 3.6 | 3.27 |
| All | mean | 12.91 | 0.54 | 0.024 | 0.44 | 0.123 | 13.1 | 5.9 | 3.6 | 3.27 |
| All | SD | 6.76 | 0.21 | 0.008 | 0.07 | 0.015 | 1.8 | 0.5 | 0.3 | 0.16 |
| VCT | min | 2.53 | 0.40 | 0.023 | 0.41 | 0.122 | 10.3 | 5.0 | 2.9 | 2.88 |
| VCT | max | 15.57 | 0.86 | 0.044 | 0.59 | 0.169 | 12.4 | 5.2 | 3.5 | 3.27 |
| VCT | median | 7.86 | 0.62 | 0.03 | 0.47 | 0.13 | 11.2 | 5.1 | 3.3 | 3.16 |
| VCT | mean | 8.61 | 0.63 | 0.03 | 0.49 | 0.137 | 11.1 | 5.1 | 3.3 | 3.11 |
| VCT | SD | 5.31 | 0.21 | 0.007 | 0.07 | 0.015 | 0.7 | 0.1 | 0.2 | 0.12 |
| 710 | min | 4.90 | 0.21 | 0.015 | 0.36 | 0.103 | 11.3 | 6.0 | 3.2 | 3.09 |
| 710 | max | 25.91 | 0.78 | 0.041 | 0.52 | 0.14 | 17.2 | 6.9 | 4.3 | 3.50 |
| 710 | median | 12.15 | 0.58 | 0.026 | 0.46 | 0.117 | 14.7 | 6.5 | 3.9 | 3.35 |
| 710 | mean | 14.52 | 0.51 | 0.026 | 0.44 | 0.119 | 14.2 | 6.4 | 3.8 | 3.31 |
| 710 | SD | 7.19 | 0.21 | 0.009 | 0.06 | 0.013 | 1.6 | 0.3 | 0.3 | 0.15 |
| mCT | min | 5.04 | 0.21 | 0.014 | 0.35 | 0.104 | 11.1 | 5.7 | 3.4 | 3.10 |
| mCT | max | 24.33 | 0.77 | 0.033 | 0.51 | 0.136 | 15.9 | 6.1 | 4.1 | 3.48 |
| mCT | median | 13.55 | 0.56 | 0.017 | 0.41 | 0.114 | 13.1 | 5.9 | 3.7 | 3.36 |
| mCT | mean | 14.26 | 0.51 | 0.019 | 0.42 | 0.117 | 13.2 | 5.9 | 3.7 | 3.33 |
| mCT | SD | 6.25 | 0.20 | 0.005 | 0.05 | 0.01 | 1.4 | 0.1 | 0.2 | 0.12 |
| | | | | | | | | | | |
| Cohort 3 Patients | SD | 17.1 | 0.17 | 0.031 | 0.11 | 0.07 | 7.5 | 3.7 | 1.7 | 0.68 |
| | | | | | | | | | | |
| All | SDpts/SDphan | 2.53 | 0.83 | 3.69 | 1.66 | 4.43 | 4.22 | 6.77 | 4.81 | 4.33 |
| VCT | SDpts/SDphan | 3.22 | 0.84 | 4.31 | 1.50 | 4.27 | 10.64 | 63.92 | 8.11 | 5.72 |
| 710 | SDpts/SDphan | 2.38 | 0.83 | 3.52 | 1.90 | 5.19 | 4.82 | 11.66 | 5.19 | 4.50 |
| mCT | SDpts/SDphan | 2.74 | 0.85 | 5.90 | 2.02 | 6.87 | 5.41 | 29.08 | 6.87 | 5.85 |

SD = standard deviation; SDpts = Cohort 3 patient standard deviation; SDphan; phantom standard deviation; SDpts/SDphan = ratio of standard deviation between Cohort 3 patient and phantom values; min = minimum value observed for the particular feature for the scanner used across the different parameters (18 total images); max = maximum value observed for the particular feature for the scanner used across the different parameters (18 total images); median = median value observed for the particular feature for the scanner used across the different parameters (18 total images); mean = mean value observed for the particular feature for the scanner used across the different parameters (18 total images); SD = standard deviation of values observed for the particular feature for the scanner used across the different parameters (18 total images)

Table 14. Change in QIF Values from Patient Scans due to Variation in Reconstruction Parameters and Comparison to Variation in Patient QIF Values

| | Contrast | Correlation | Energy | Homogeneity | Uniformity | SUVmax | SUVmean | SD | Entropy |
|-----------------------|----------|-------------|--------|-------------|------------|--------|---------|-------|---------|
| Patient Ratio 1 | 7.82 | 1.42 | 5.11 | 1.85 | 15.03 | 10.05 | 20.45 | 10.44 | 10.05 |
| Patient Ratio 2 | 9.45 | 1.26 | 8.81 | 2.36 | 10.49 | 12.72 | 13.89 | 13.37 | 9.92 |
| Patient Ratio 3 | 3.24 | 0.94 | 11.54 | 1.83 | 11.11 | 6.89 | 22.48 | 5.62 | 7.68 |
| Patient Ratio 4a | 9.30 | 1.43 | 5.68 | 2.19 | 6.78 | 4.36 | 26.90 | 7.27 | 6.32 |
| Patient Ratio 4b | 3.71 | 1.61 | 13.37 | 2.50 | 16.62 | 3.69 | 22.34 | 5.02 | 8.14 |
| Patient Ratio 5 | 2.71 | 1.35 | 21.97 | 2.48 | 15.57 | 4.93 | 25.52 | 5.30 | 6.97 |
| | | | | | | | | | |
| Average Patient Ratio | 6.04 | 1.33 | 11.08 | 2.20 | 12.60 | 7.11 | 21.93 | 7.84 | 8.18 |
| Phantom Ratio - 710 | 2.38 | 0.83 | 3.52 | 1.90 | 5.19 | 4.82 | 11.66 | 5.19 | 4.50 |

4.3 Results of Specific Aim 3: Assess relationships between CT-based quantitative image features, PET-based quantitative image features, conventional features, and morphologic features

Specific aim 3 examined whether relationships exist between CT-based quantitative image features, PET-based quantitative image features, conventional features, and morphologic features. Correlations were investigated using features identified in previous analyses, such as COM energy and solidity in PET and LOG_Average and Uniformity in CE-CT. Uniformity in PET was also tested as it was found to be strongly correlated with COM energy and is calculated in the same fashion only without taking into account 2D displacement associations.

4.3.1 Results for Project 3.1: Quantify correlations between prognostic FDG-PET-based and CECT-based quantitative image features

Cohort 4 was used for determining correlations between PET and CE-CT QIFs that were found to have prognostic value in sections 4.1.1 and 4.2.1. FDG-PET based uniformity was also added to this list as it was found to be significantly correlated to COM Energy and have a very similar formula for quantification. Table 15 shows the Pearson correlation coefficients and associated p-values testing whether two metrics are significantly correlated.

Table 15. Correlations between PET and CE-CT Features

| FDG-PET QIFs | CE-CT QIFs | |
|---------------------|--------------------|-------------------|
| | LOG_Average | Uniformity |
| COM Energy | -0.32 (p = 0.005) | -0.005 (p = 0.96) |
| Solidity | 0.31 (p = 0.006) | 0.26 (p = 0.02) |
| Uniformity | -0.41 (p = 0.0002) | -0.03 (p = 0.80) |

The LOG_Average from CE-CT was associated with the 3 examined PET QIFs. Uniformity from CE-CT was significantly correlated with solidity but not with COM energy or uniformity. This data found that patients having heterogeneous FDG-uptake in the primary tumor and more dispersion between primary and nodal disease were correlated with tumors found to have high intensity and/or frequent edges on CE-CT. While statistically significant in many cases, the correlation coefficients were quite low across all comparisons. Graphical representations plotting CE-CT QIFs versus FDG-PET QIFs are shown in Figure 27. This implies that relationships do exist between QIFs from different modalities; however they by no means can be used interchangeably. For example, in Figure 27 it is clear that in the bottom left figure comparing CE-CT LOG_Average to FDG-PET uniformity that higher values of FDG-PET uniformity are associated with lower values of CE-CT LOG_Average. However, one could not simply use CE-CT LOG_Average and accurately determine the FDG-PET uniformity. For instance, having an FDG-PET uniformity value of 0.1 is associated with CE-CT LOG_Average values ranging from 1 to 4.5.

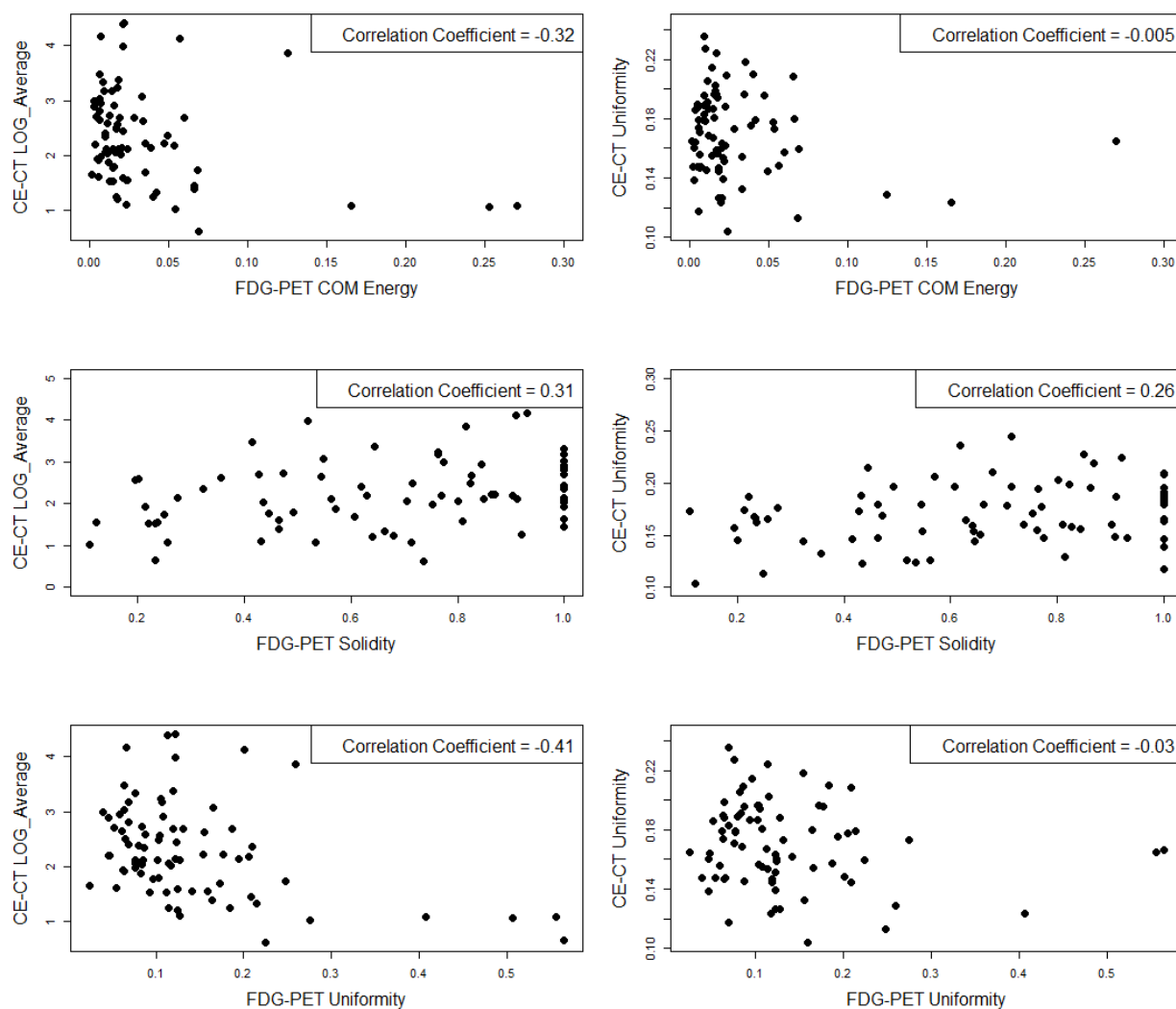


Figure 27. Assessment of Correlations between Prognostic CE-CT QIFs (LOG_Average and Uniformity) and Prognostic FDG-PET QIFs (COM Energy, Solidity, and Uniformity)

4.3.2 Results for Project 3.2: Quantify if relationships exist between CE-CT-based and FDG-PET-based quantitative image features with tumor volume and TNM staging

Cohort 2 was used to determine if relationships existed between CE-CT QIFs and tumor volume (volume = primary volume and GTV = primary volume + nodal volume) or TNM staging. Cohort 3 was used to determine if relationships existed between FDG-PET QIFs and tumor volume or TNM staging. When comparing to FDG-PET QIFs, both the primary and GTV (i.e., the metabolic tumor volume, MTV) were determined from the contours on the FDG-PET scan and not based on the CT. Box plots relating CE-CT and FDG-PET QIFs are shown below in

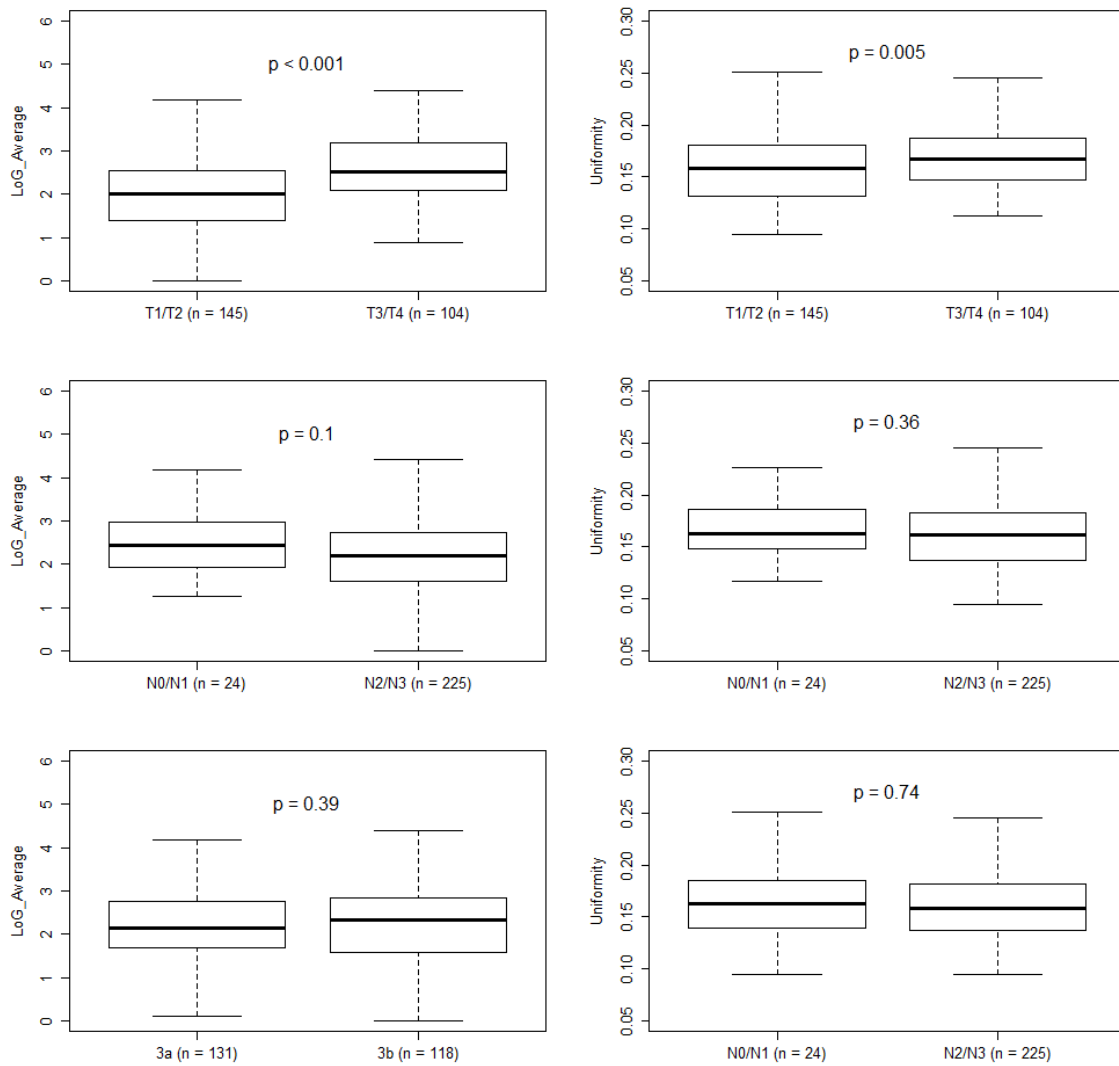
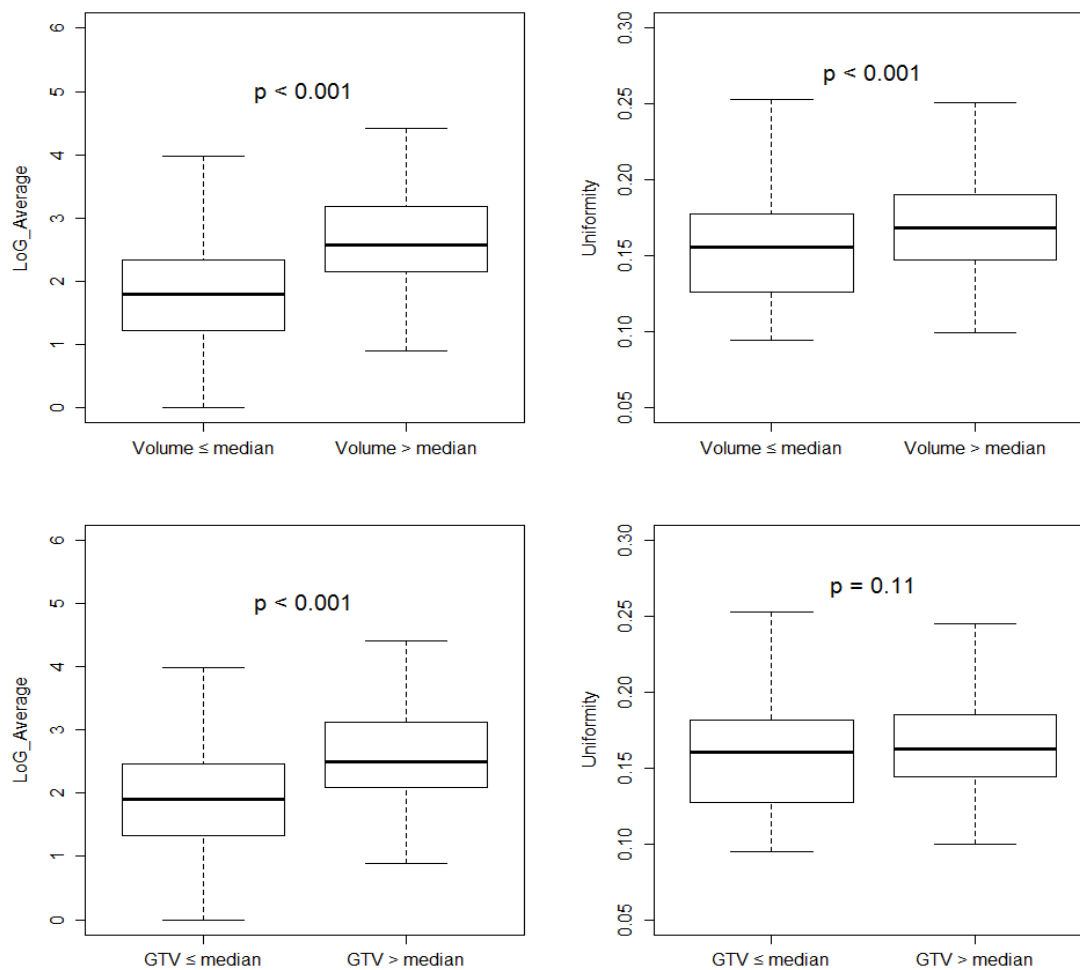


Figure 28 and Figure 29, respectively. The p-values were determined from Wilcoxon rank-sum tests.



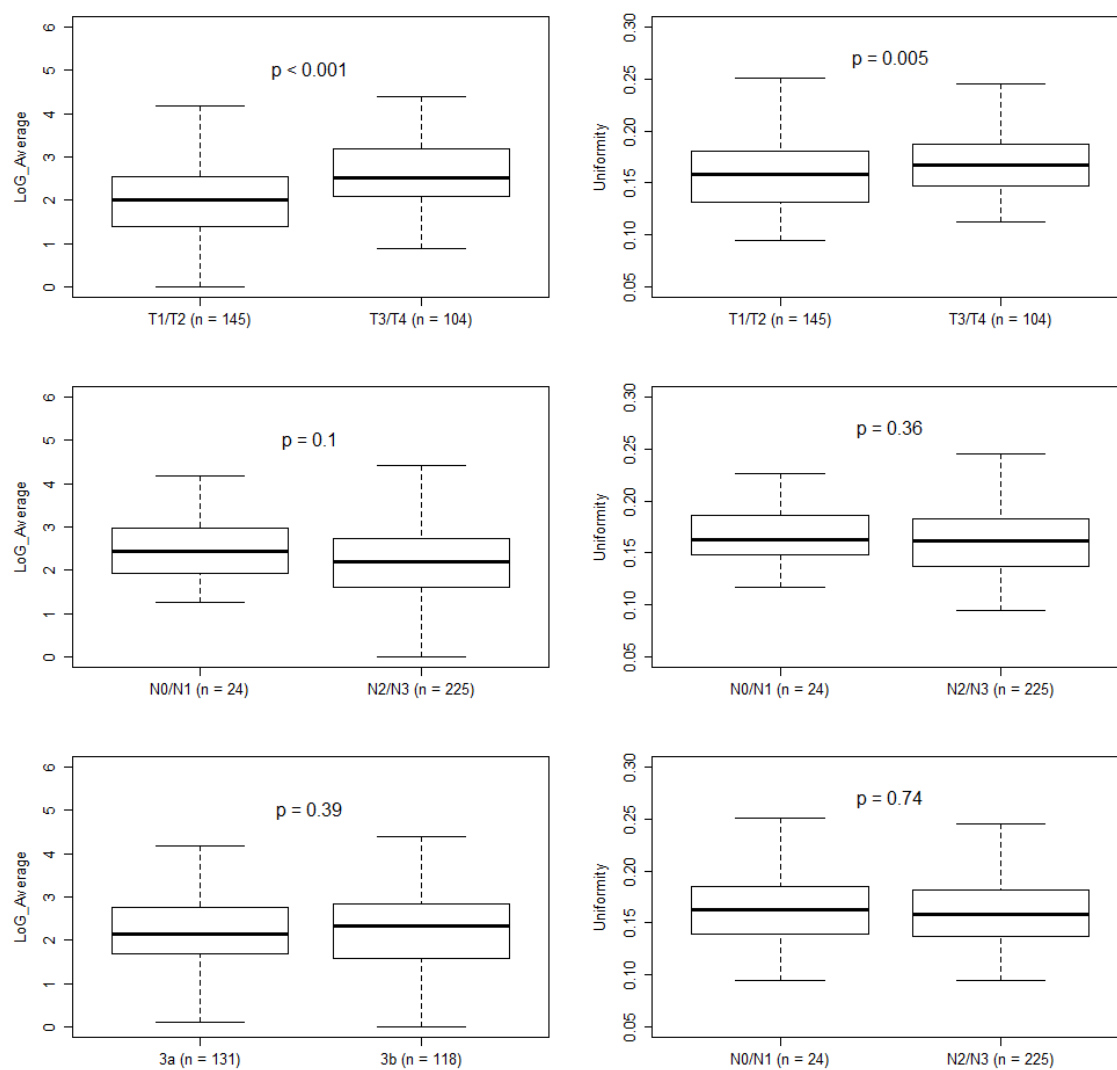
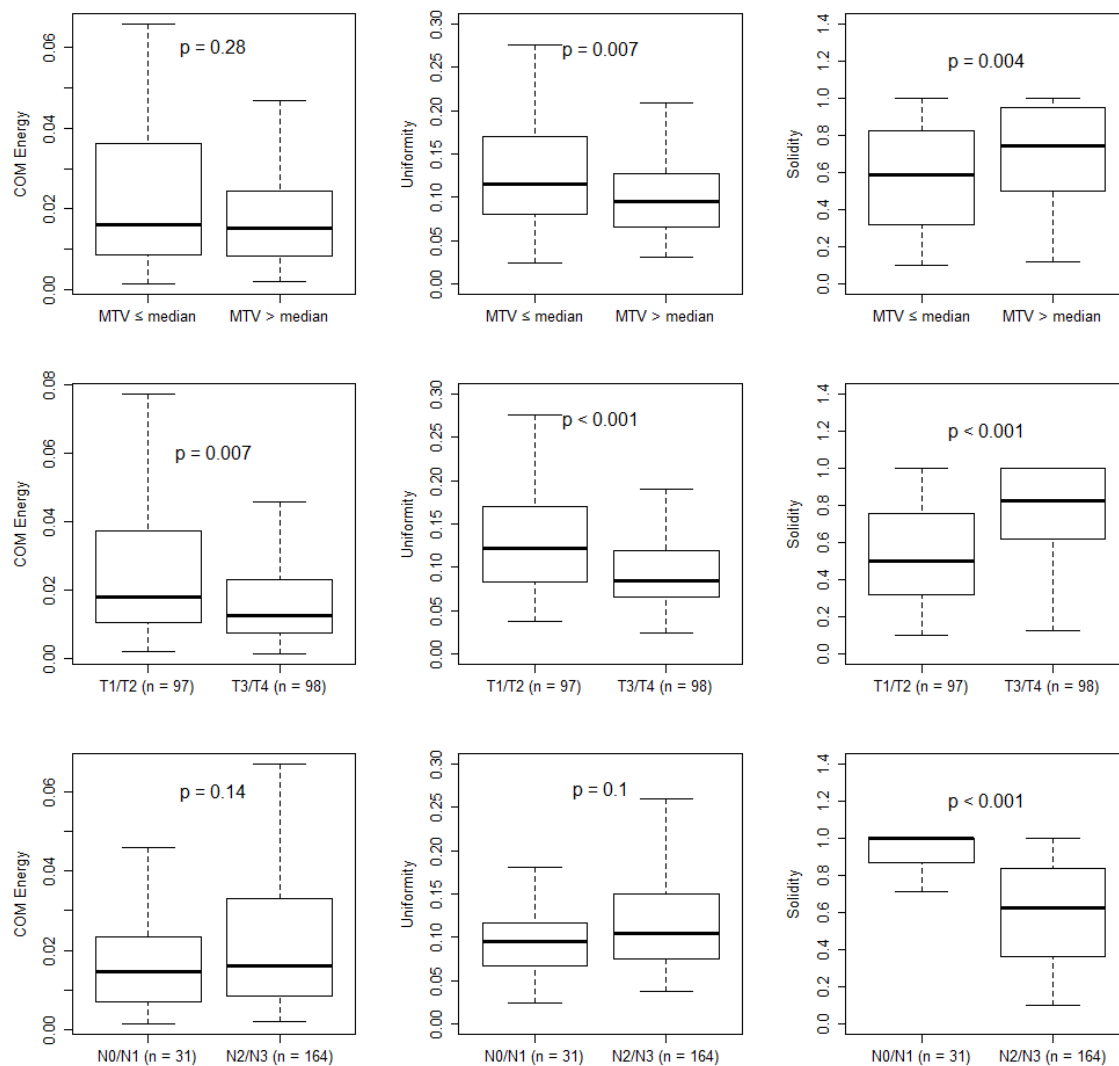


Figure 28. Comparison of LOG_Average and Uniformity from CE-CT Versus Tumor Volume and Staging



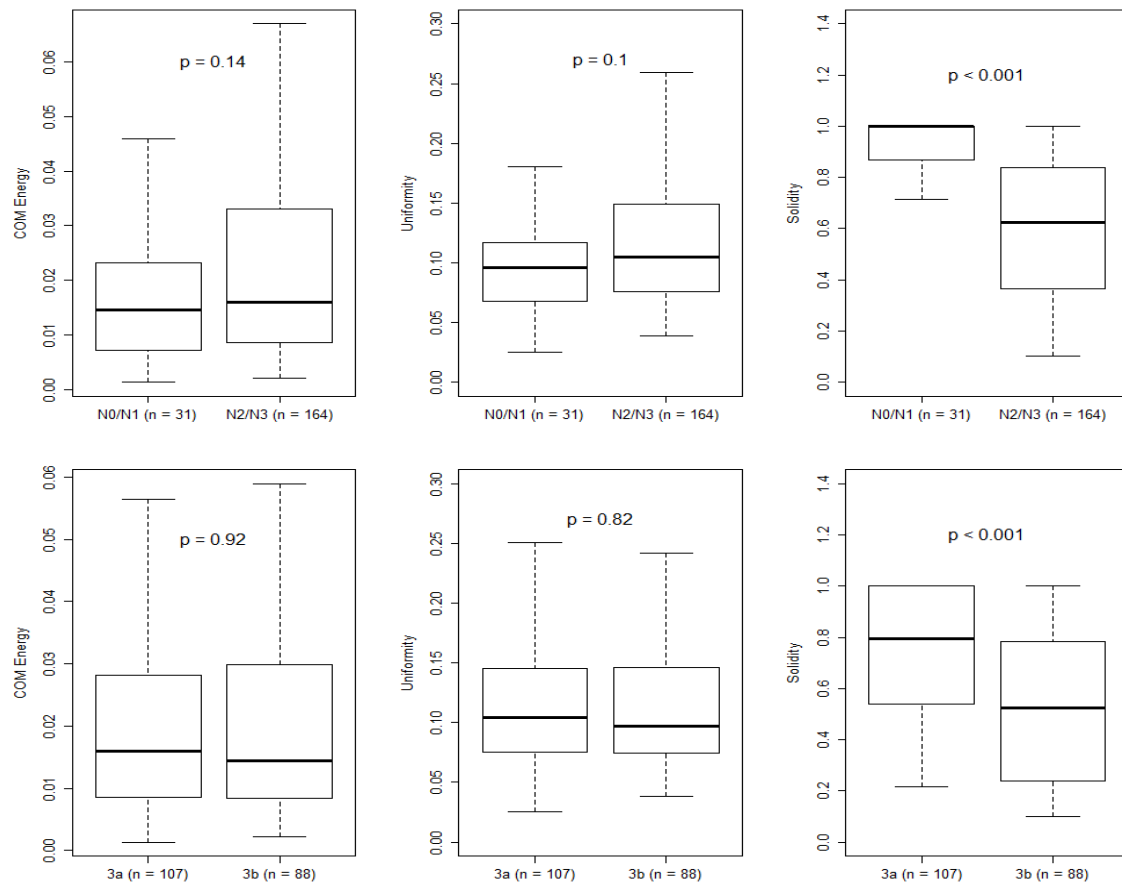


Figure 29. Comparison of COM Energy, Uniformity, and Solidity from FDG-PET versus Tumor Volume and Staging

Significant differences in CE-CT, LOG_Average, and CE-CT uniformity were observed between various CPFs. Significant differences in FDG-PET COM energy, solidity and uniformity were also observed between various CPFs. Uniformity was significantly different when stratified by primary tumor volume, GTV, and T stage. FDG-PET based COM energy was significantly different when stratified by primary volume, T stage, and N stage. FDG-PET based solidity significantly differed when stratified by all tested CPFs. FDG-based uniformity was significantly different when stratified by primary tumor volume, GTV, and T stage. Significant differences were observed between QIFs when stratified by CPFs. The results of Specific Aims 1 and 2 suggest that even though relationships exist between QIFs and CPFs, QIFs provide additional prognostic information.

4.3.3 Results for Project 3.3: *Quantify if there are correlations between FDG-PET-based quantitative image features, CECT-based quantitative image features, and morphologic characteristics (vessels, necrosis, air cavities, etc.)*

Radiologists routinely observe morphologic characteristics of tumors such as the presence of necrosis, cavitation, and heterogeneous enhancement when examining CE-CT images. The purpose of this project was to quantitatively assess how morphologic features influence the prognostic QIFs found in sections 4.1.1 and 4.2.1. Sections 3.3.5 *Contrast Enhanced CT Auto-segmentation of Morphologic Characteristics* and 3.3.6 *PET Necrosis Auto-segmentation* describe how morphologic characteristics can be extracted in a quantitative fashion. Namely, these sections describe the extraction and quantification of volume and percent of the tumor that consists of vessels, necrosis, air cavities, and tumor tissue. Features quantifying vessels, necrosis, air cavities, and tumor tissue are able to be extracted from CE-CT and features regarding necrosis are able to be extracted from PET. Figure 30 displays boxplots of dichotomous comparisons between QIFs and the presence/absence of tissue types and their associated p-values as determined by the Wilcoxon rank-sum test.

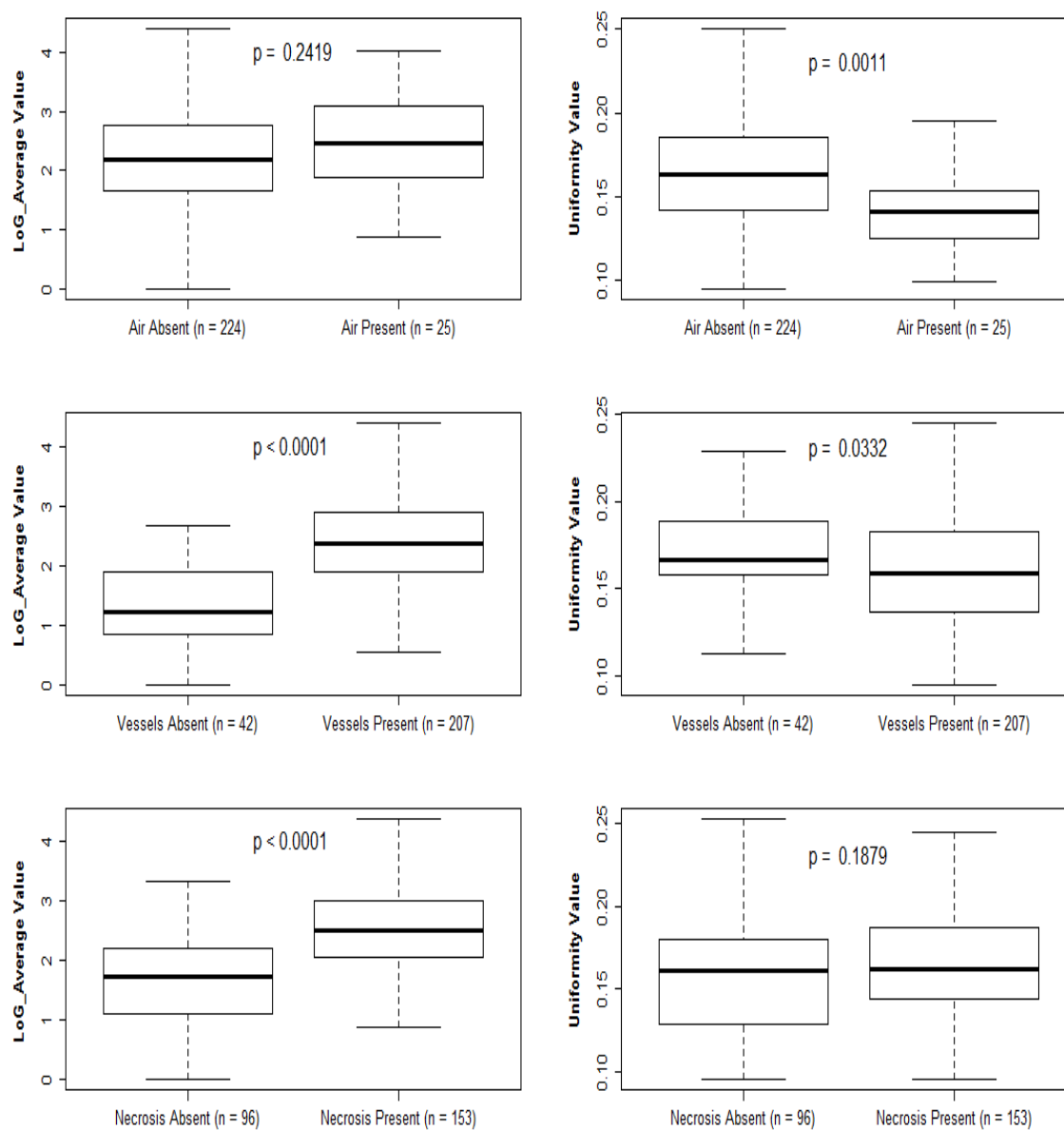


Figure 30. Comparison of LOG_Average and Uniformity on CE-CT versus The Presence/Absence of Various Tissue Types

Figure 31 and Figure 32 below compare the extracted necrosis volume and necrosis percentage of tumor from FDG-PET and CE-CT auto-segmentations, respectively.

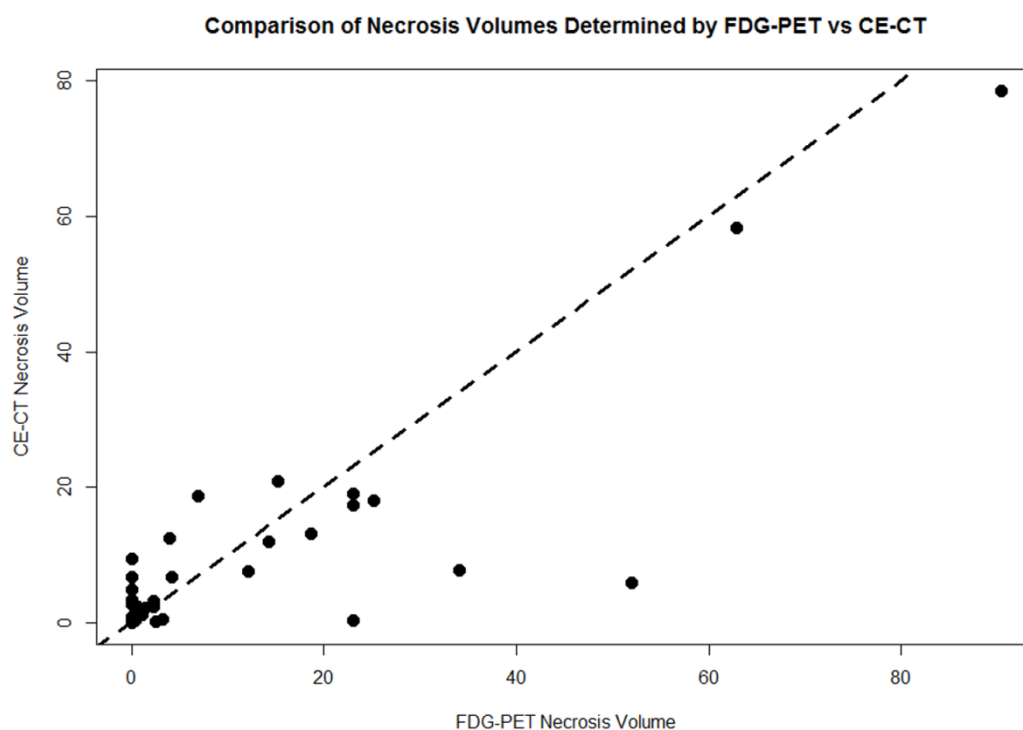


Figure 31. Comparison of Necrosis Volumes Determined By FDG-PET Vs CE-CT

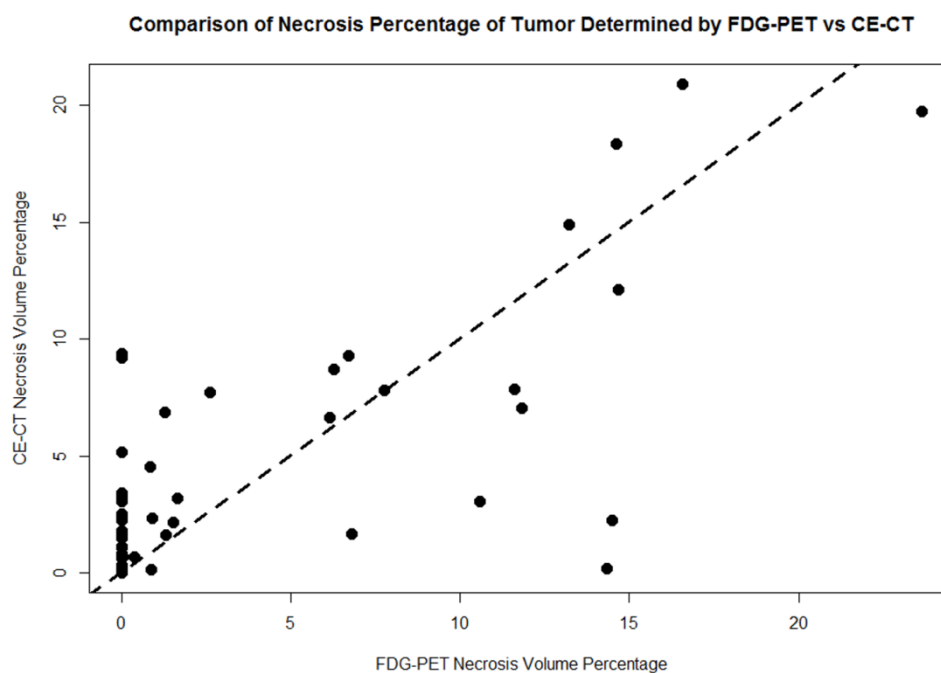


Figure 32. Comparison of Necrosis Percentage Determined By FDG-PET vs CE-CT

The CCC values from Figure 31 and Figure 32 were 0.85 and 0.76, respectively. It was observed that the auto-segmentation from CE-CT identified more small regions of necrosis compared to the auto-segmentation from FDG-PET. In general, the agreement seen between the two methodologies was reasonable based on their CCC values.

In addition to quantifying correlations between PET, CE-CT, and morphologic characteristics, it was important to determine how the presence of morphologic characteristics influences QIF metrics. To investigate this, plots were generated (Figure 33) comparing the original (i.e., total tumor) to the original tumor ROI excluding certain tissue types. The details of these analyses are described in 3.3.5 Contrast Enhanced CT Auto-segmentation of Morphologic Characteristics.

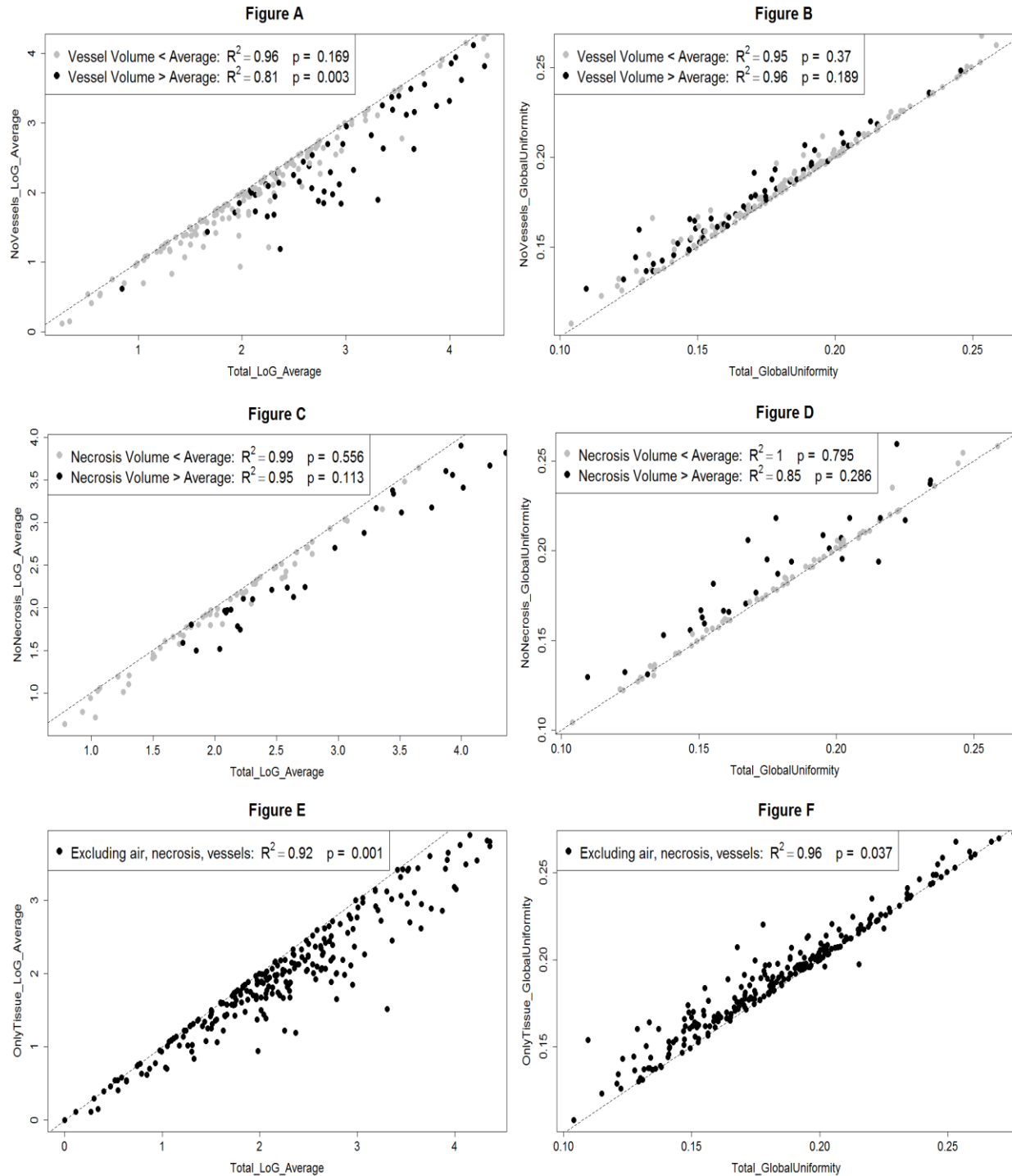


Figure 33. Comparison of CE-CT Feature Value for Entire Tumor (X-Axis) with Feature Value Excluding a Particular Tissue Type or Types (Only Tissue: Excludes Air, Necrosis, and Vessels). (A) Comparison of LOG_Average Values from Tumors with Enhancing Vessels when the Vessels Are Present (X-Axis) or Excluded (Y-Axis). (B) Comparison of Uniformity Values from Tumors with Enhancing Vessels when the Vessels are Present (X-Axis) or Excluded (Y-Axis). (C) Comparison of LOG_Average Values from Tumors with Necrosis when the Necrosis Is

Present (X-Axis) or Excluded (Y-Axis). (D) Comparison of Uniformity Values from Tumors with Necrosis when the Necrosis Is Present (X-Axis) or Excluded (Y-Axis). (E) Comparison of LOG_Average Values from Tumors with Cavitation when the Cavitation Is Present (X-Axis) or Excluded (Y-Axis). (F) Comparison of Uniformity Values from Tumors With Cavitation when the Cavitation Is Present (X-Axis) or Excluded (Y-Axis).

All R^2 values from the plots in Figure 33 were greater than 0.81. Excluding enhancing vessels and/or necrotic regions led to a decrease in measured LOG_Average values and an increase in uniformity values. As expected, a decrease in LOG_Average (metric quantifying number and intensity of “edges” within the tumor) led to an increase in uniformity. Figure 33E and Figure 33F illustrate that analyzing only tumor tissue (i.e., excluding necrosis and enhancing vessels) yielded different values of LOG_Average and intensity histogram uniformity when compared to the total tumor including all morphologic tissue types ($p = 0.001$ and $= 0.037$, respectively). While these values are statistically different, the R^2 values are generally very high (greater than 0.81). This implies that while statistically different, the values are still highly related. No differences were found when comparing uniformity values between tumors with above or below average volumes of vessels or necrosis based on our sampled cohort. However, it was found that tumors with above average volume of vessels and necrosis were found to have significantly higher values of LOG_Average ($p < 0.01$) for both the total tumor contour and ROI excluding these morphologic features.

4.4 Results of Specific Aim 4: Potential use of FDG-PET-based quantitative image features

4.4.1 Results for Project 4.1: Assess whether significant PET-based quantitative image features relate to a difference in patient survival for those treated with an escalated radiation dose

Section 4.2.1 was able to identify FDG-PET based QIFs that appeared to be prognostic for patient survival. The purpose of this project was to see if these QIFs could be used to identify patients that would benefit from receiving an escalated radiation dose. We hypothesized that patients with primary tumors with uniform FDG-uptake (i.e. high values of COM energy) may benefit from dose escalation as heterogeneous FDG-uptake has been associated with poor pathologic factors, aggression, and inferior outcome. Additionally, we also hypothesized that patients with non-dispersed local-regional disease (i.e. high values of solidity) could also potentially benefit from dose escalation as these patients are probably less likely to develop metastatic disease and would result in less dose delivered to cardiothoracic normal tissue structures (this was confirmed in an analysis shown in Appendix D: Relationship of Cardiothoracic Dosimetry with Disease Solidity). The development of metastatic disease would reduce the survival impact associated with an increase in local control and higher normal tissue doses have been shown to reduce patient survival. Therefore, we chose to investigate the impact of dose escalation considering both patient COM energy and disease solidity. Since high values of each feature were hypothesized to be beneficial with increased radiation dose, we assessed the impact of creating sub-cohorts of patient with high values of both COM energy and disease solidity.

We first stratified all patients in Cohort 3 by radiation dose (74 Gy vs 60-70 Gy). Dose escalation did not result in a difference in overall survival or progression-free survival (Figure 34).

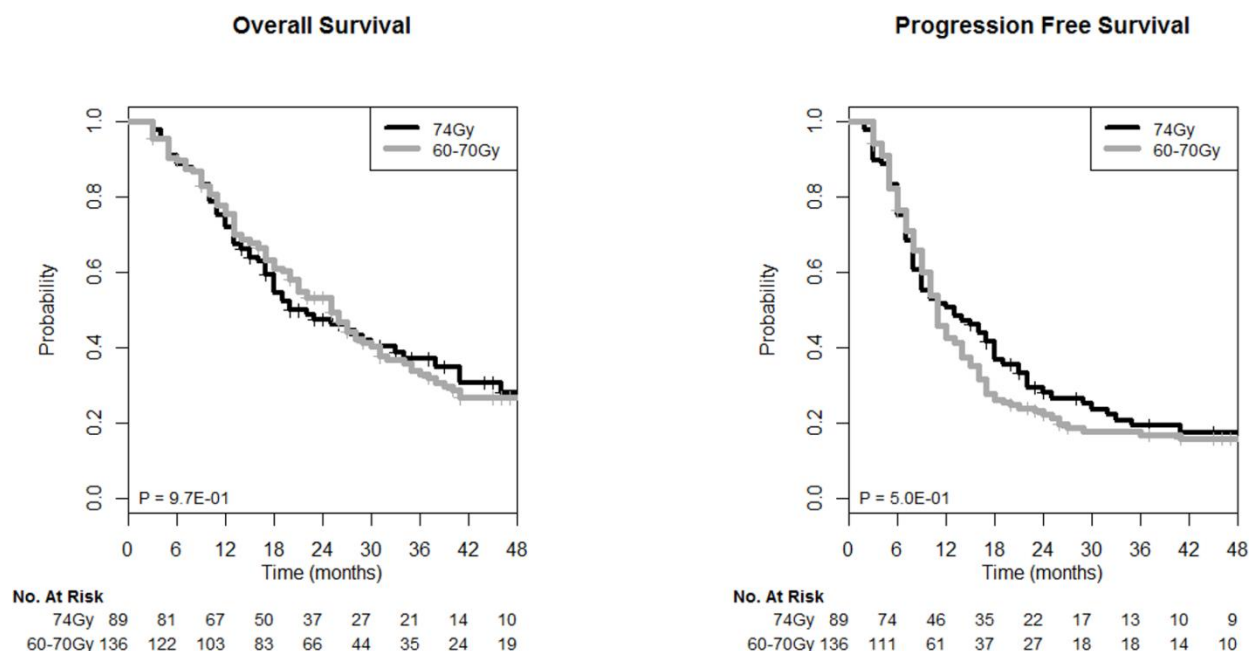


Figure 34. Stratification of Overall Survival and Progression-Free Survival by Dose Level in All Patients within Cohort 3

A grid search was performed to examine different combinations of sub-cohorts as determined by varying cutoffs of the QIFs found to be prognostic in 4.2.1 Results for Project 2.1 (i.e., solidity and COM energy). Initially, the search was performed to isolate patients with homogeneous SUVs within the primary tumor (high COM energy) and close proximity of disease (high solidity). The results of this search are seen in Figure 35. The values within the figures are the p-values from a log-rank test when stratifying the specific sub-cohort by dose level. Values at the top left have mild cutoffs in terms of the QIFs and thus include majority of all patients (180/195). Values progressively closer to the bottom right have more strict cutoffs in terms of QIFs and include patients with progressively higher values. As higher and higher cutoffs are implemented for the QIFs, the p-values from the log-rank test between dose levels become significant for both overall survival and progression-free survival.

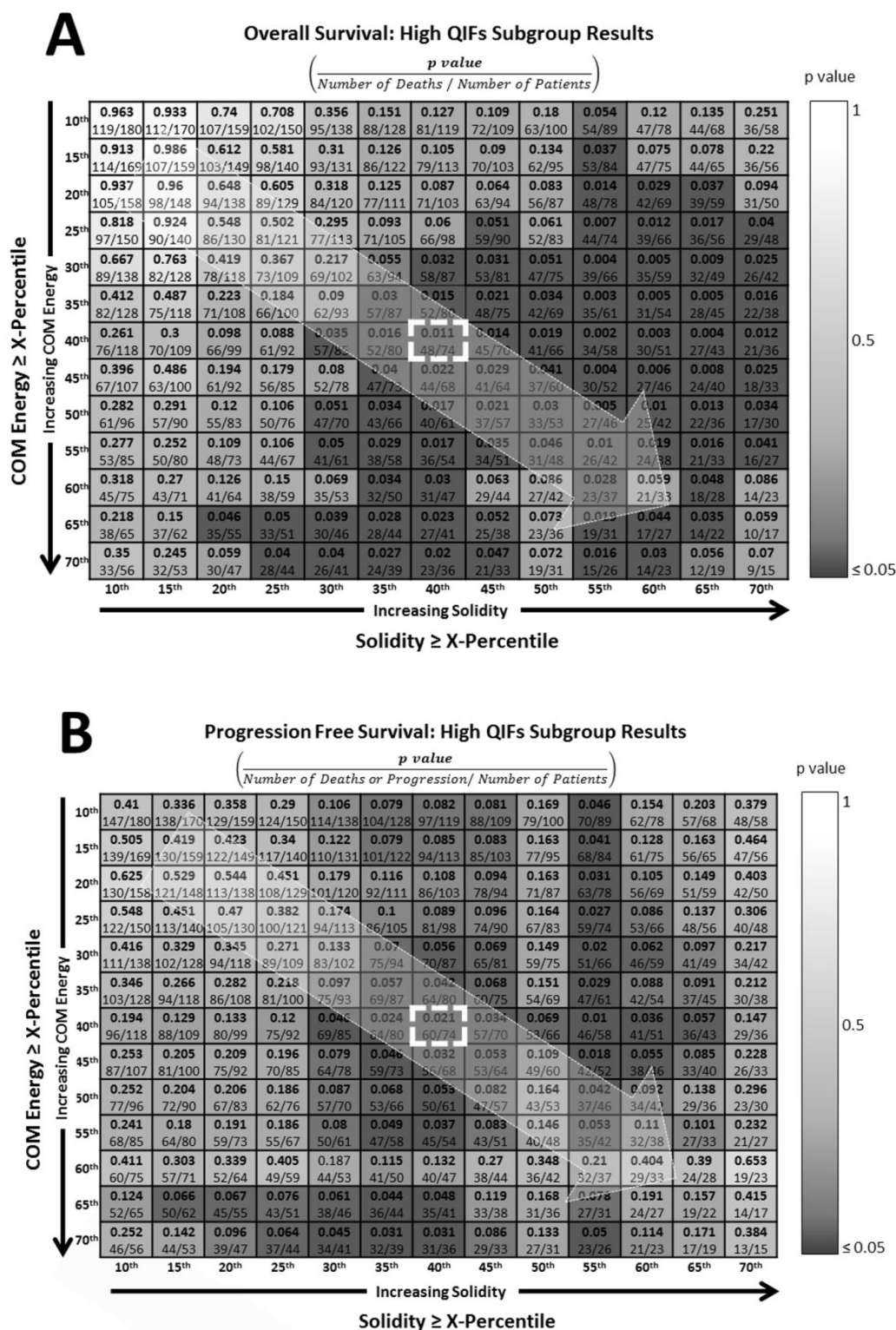


Figure 35. Log-Rank P-Values from Sub-Cohorts Based on High Values FDG-PET QIFs In Terms Of Overall Survival (A) and Progression-Free Survival (B)

Kaplan-Meier plots from the sub-cohort in the cell with the dashed white outline Figure 35 are shown in Figure 36. In this sub-cohort, patients receiving an escalated dose of 74 Gy had superior overall survival and progression-free survival compared to those receiving 60-70 Gy ($p = 0.01$ and 0.02 , respectively).

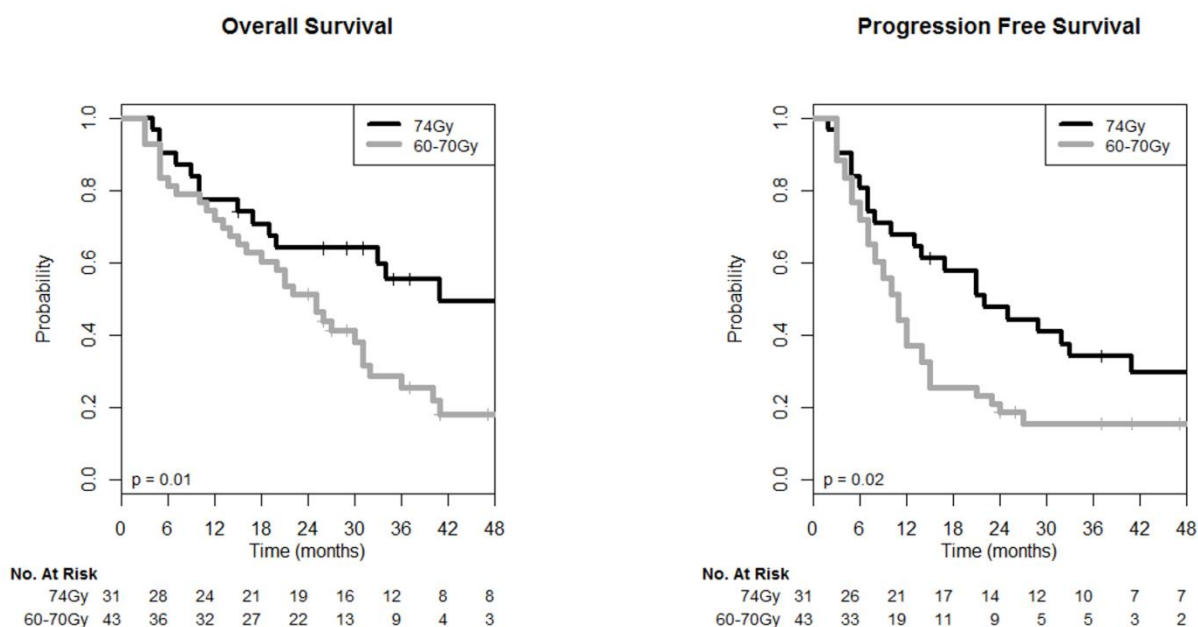


Figure 36. Kaplan-Meier Plots Stratified by Dose Level for The Sub-Cohort with High Values Of FDG-PET QIFs In Terms of Overall Survival and Progression-Free Survival

The opposite association (i.e., impact of low values of COM energy and solidity) was examined in the same manner. The results of this search are seen in Figure 37. Values at the bottom right now have mild cutoffs in terms of the QIFs and thus include the majority of all patients (179/195). Values progressively closer to the top left have more strict cutoffs in terms of QIFs and include patients with progressively lower values. It can be seen as more strict cutoffs are implemented for the QIFs, the p -values from the log-rank test between dose levels become significant for both overall survival and progression-free survival.

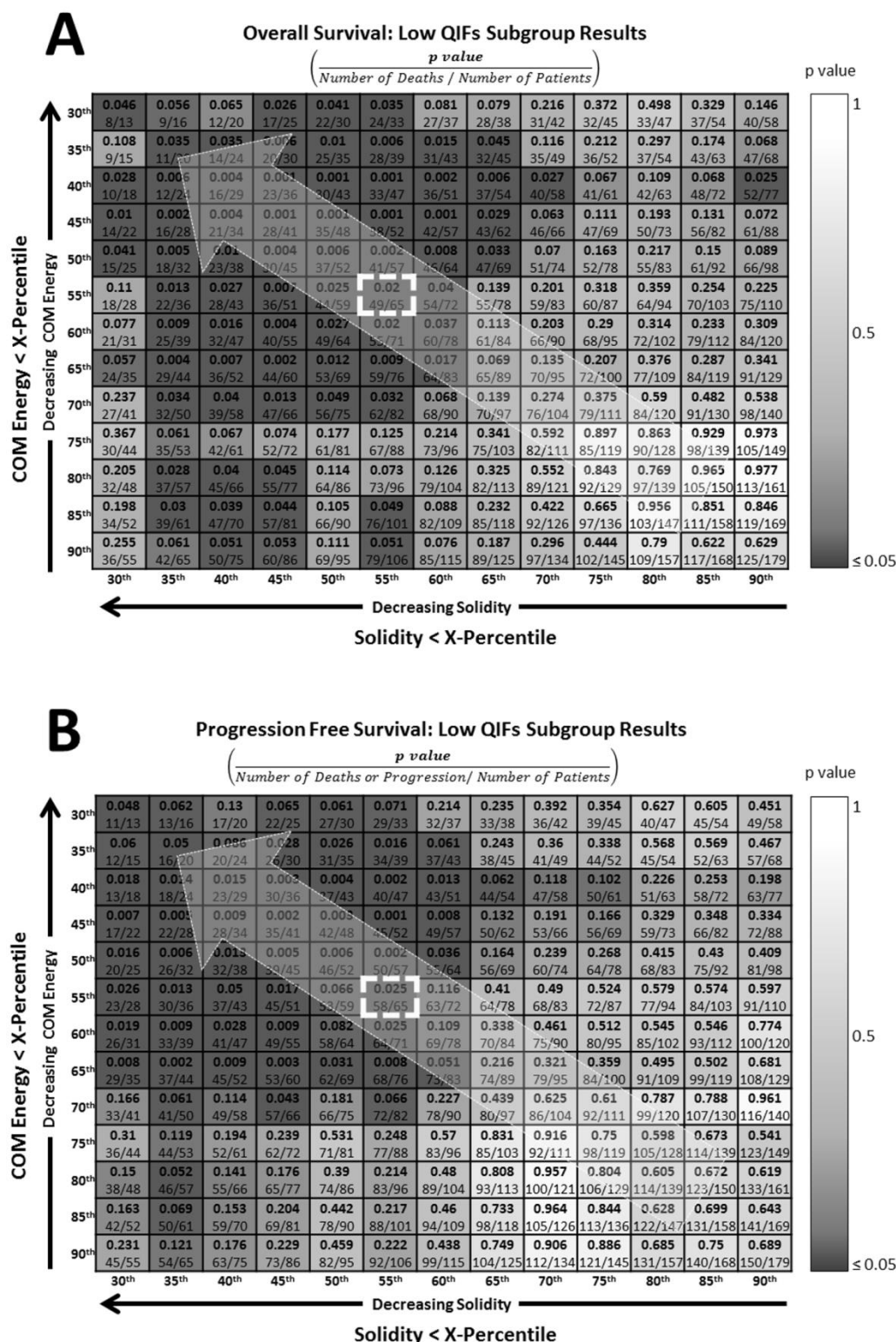


Figure 37. Log-Rank P-Values from Sub-Cohorts Based on High Values of FDG-PET QIFs In Terms of Overall Survival(A) and Progression-Free Survival(B)

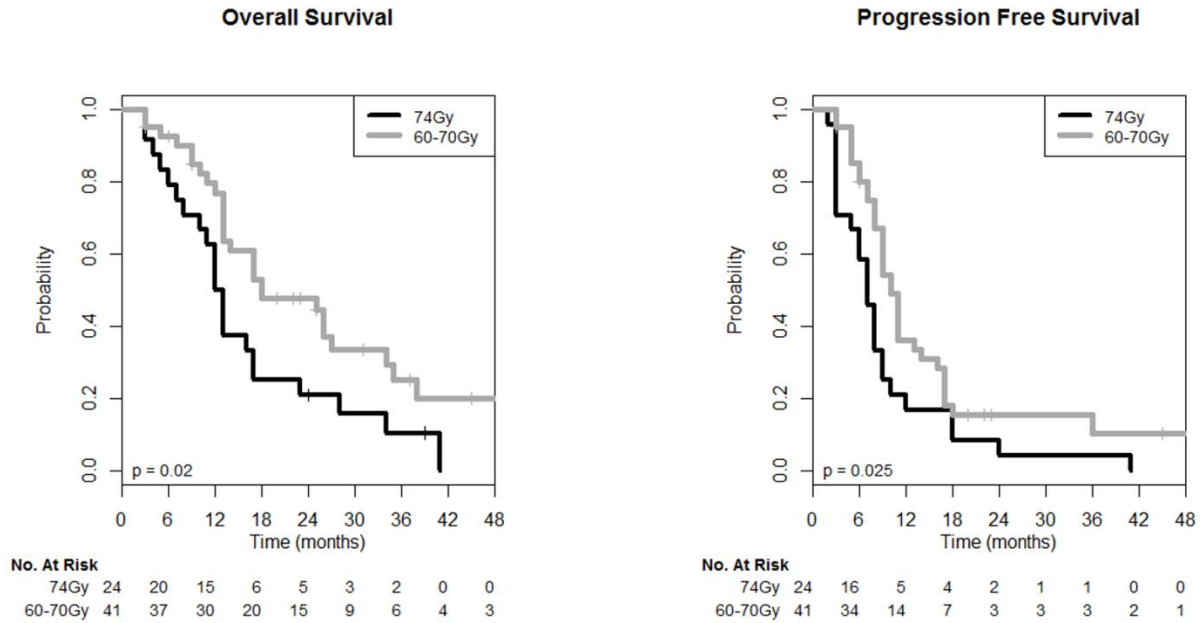


Figure 38. Kaplan-Meier Plots Stratified by Dose Level for the Sub-Cohort with Low Values of FDG-PET QIFs In Terms of Overall Survival and Progression Free Survival

Kaplan-Meier plots from the sub-cohorts in the cells with the dashed white outlines in Figure 37 are shown in Figure 38. In this sub-cohort, patients receiving an escalated dose of 74 Gy had inferior overall survival and progression-free survival compared to those receiving 60-70 Gy ($p = 0.02$ and 0.025 , respectively).

Furthermore, using a multivariate Cox model, receiving 74 Gy versus 60-70 Gy was an independent prognostic factor for both overall survival (high value QIFs: $p = 0.012$, low value QIFs: $p = 0.02$) and progression-free survival (high value QIFs: $p = 0.015$, low value QIFs: $p = 0.025$) when adjusting for overall stage, T stage, receiving induction chemotherapy, age, gender, and GTV. These CPFs were used as they were selected during cross-validation in our previous work in 4.2.1 Results for Project 2.1: *Quantify the impact of adding FDG-PET-based quantitative image features to outcome models containing only CPFs.*

Table 16. Comparison of Conventional Prognostic Factors

| | High QIFs Values Subgroup | | | Low QIFs Values Subgroup | | |
|----------------------|---------------------------|---------|---------|--------------------------|---------|---------|
| | 74Gy | 60-70Gy | p-value | 74Gy | 60-70Gy | p-value |
| Mean GTV (cc) | 200 | 219 | 0.69 | 192 | 186 | 0.87 |
| Age (mean) | 65 | 63 | 0.49 | 67 | 63 | 0.18 |
| MTV (cc) | 118 | 123 | 0.91 | 104 | 126 | 0.44 |
| Overall Stage | | | 0.31 | | | 0.79 |
| 3a | 24 | 28 | | 16 | 25 | |
| 3b | 7 | 15 | | 8 | 16 | |
| KPS | | | 0.3 | | | 0.78 |
| 60 | 0 | 3 | | 1 | 3 | |
| 70 | 3 | 2 | | 1 | 3 | |
| 80 | 18 | 29 | | 16 | 20 | |
| 90 | 10 | 8 | | 6 | 14 | |
| 100 | 0 | 1 | | 0 | 1 | |
| Induction | | | 0.29 | | | 0.42 |
| No | 21 | 34 | | 18 | 26 | |
| Yes | 10 | 9 | | 6 | 15 | |
| Concurrent | | | 0.51 | | | 0.38 |
| No | 0 | 2 | | 0 | 3 | |
| Yes | 31 | 41 | | 24 | 38 | |
| Adjuvant | | | 1 | | | 0.25 |
| No | 20 | 27 | | 20 | 28 | |
| Yes | 11 | 16 | | 4 | 13 | |
| Histology | | | 0.24 | | | 1 |
| Squamous Cell | 11 | 22 | | 10 | 16 | |
| Other | 20 | 21 | | 14 | 25 | |
| Smoking | | | 0.1 | | | 0.44 |
| Never | 0 | 4 | | 1 | 4 | |
| Former | 23 | 23 | | 19 | 26 | |
| Current | 8 | 16 | | 4 | 11 | |
| Gender | | | 0.03 | | | 1 |
| Male | 27 | 27 | | 9 | 16 | |
| Female | 4 | 16 | | 15 | 25 | |

Figure 34 through Figure 38 demonstrate that the QIFs found to be prognostic in section 4.2.1 Results for Project 2.1: *Quantify the impact of adding FDG-PET-based quantitative image features to outcome models containing only CPFs* were capable of identifying sub-cohorts of patients whose survival was influenced by dose escalation. Dose escalation did not appear to influence patient survival on the entire cohort. Figure 35 and Figure 37 show that these observations followed a definite trend and were not purely the result of selecting a significant result from a large number of tests. Dose escalation was found to be prognostic in the isolated sub-cohorts even when controlling for CPFs such as overall stage, T stage, receiving induction chemotherapy, age, gender, and GTV. The lack of imbalances in Table 16 implies that CPFs do not appear to be responsible for the observed survival differences.

Chapter 5 Discussion

Discussion Specific Aim 1

We hypothesized that the addition of CT-based quantitative image features would significantly improve outcome models compared to models using conventional prognostic factors. This hypothesis was confirmed as the addition of CT-based quantitative image features significantly improved outcome models compared to models using conventional prognostic factors. QIFs extracted from CT were able to improve model fit compared to models using CPFs excluding or including GTV. The initial analysis of Cohort 1 (91 patients with a pretreatment T_{AVG} , T_{50} , and CE-CT) found that the addition of QIFs improved the stratification of patient outcome compared to models using CPFs including or excluding GTV. In addition, incorporating the reproducibility of QIFs yielded a percent classification reproducibility of approximately 80%. QIFs from CE-CT were found to be the most significant in terms of prognostic value for patient outcome between CT types but were far less reproducible compared to features from T_{AVG} and T_{50} . The choice to analyze only CE-CT images in Cohort 2 was based on data from Cohort 1 that found the most significant source of prognostic information was from the features extracted from CE-CT. Having contrast injected facilitates greater HU differences within the tumor due to the contrast infiltrating vessels and subsequently into the tissue. We hypothesized this may be the reason the CE-CT derived features were more prognostic than those from the non-contrast 4D-CT. Therefore, we decided to develop a cohort of patients who only needed to have a CE-CT regardless of whether they received a 4D-CT scan.

The analysis of Cohort 2 (249 patients with a pretreatment CE-CT) found that QIFs significantly improved model fit but did not improve the c-indices or patient stratification. These results suggest that CE-CT-based QIFs are associated with a statistically significant improvement in outcome model fit using the most sensitive test (log-likelihood ratio). However, the results would probably not be considered a *clinically* significant improvement in predictive ability since neither the c-index nor visual stratification appeared to improve.

The results suggest that CT QIFs and more specifically CE-CT-based QIFs may be useful in patient outcome modeling. CT-based QIFs have the potential to develop clinically useful prediction models using medical images that are already obtained during routine patient staging. Therefore, implementation would require little to no added cost and would not require additional time, discomfort, or radiation dose to patients. The ability to stratify patients in ways shown to be superior to current staging methods might allow physicians to deliver more optimized, patient-specific treatment. While our initial analysis of Cohort 1 appeared more efficacious in terms of LRC and FFDM, ultimately stratifying patients in terms of overall survival would generate the most benefit to physicians and patients alike. Our analysis of Cohort 2 using only CE-CT scans did result in improved c-indices and stratification but these results may be due to the elevated prognostic ability of CPFs including GTV. The impact of CPFs and GTV seen in Cohort 2 is far superior to what was seen Cohort 1 (91 pts), Cohort 3 (195 pts), and Cohort 4 (77 pts). Since QIFs from CE-CT were still significant in improving model fit in Cohort 2, it is possible that improved stratification may be possible in alternative cohorts even though this was not observed in our analysis. There is a relatively extensive body of literature supporting the idea that QIFs are prognostic in NSCLC with a vast majority performed using non-contrasted CT scans.^{10, 14, 15, 48–50} However, work by Ravanelli et al. did find that CE-CT based QIFs (tumor uniformity * grey level) was able to predict response to first-line chemotherapy in NSCLC.⁵¹ Additionally, Al-Kadi and Watson also found that fractal based QIFs from CE-CT were able to predict malignant aggressiveness in NSCLC.¹⁶ Aerts et al. (in non-contrast CT) did find evidence that QIFs improved NSCLC patient outcome stratification in addition to tumor volume and staging.

Project 1.2 found that a majority of QIFs from CE-CT did not have a CCC value greater than 0.9. This could be due to the fact that these were not performed in a traditional test/retest fashion but were scans taken at different institutions separated by an average of 38 days. Differences in imaging within the ten patients, such as scanner type, manufacturer, imaging parameters, contrast timing, etc. (see 17), along with growth/underlying change in the tumor could easily be responsible for the low feature reproducibility.

Table 17. CE-CT Test/Retest Scan Information

| Field | Outside Scan (n = 10) | MD Anderson Scan (n = 10) |
|--------------------------------|------------------------------|----------------------------------|
| Pixel Dimension (range) | 0.70 – 0.86 | 0.78 – 0.86 |
| Manufacturer/Model | | |
| GE/Lightspeed 16 | 0 | 7 |
| GE/Lightspeed VCT | 2 | 2 |
| GE/Lightspeed Plus | 1 | 1 |
| GE/Lightspeed Ultra | 2 | 0 |
| GE/Lightspeed Pro | 1 | 0 |
| Phillips/Brilliance64 | 1 | 0 |
| Toshiba/Aquilon | 2 | 0 |
| Siemens/Sensation 16 | 1 | 0 |
| Reconstruction Kernel | | |
| Standard | 8 | 10 |
| FC03 | 1 | 0 |
| FC13 | 1 | 0 |

The definition of a CCC value of 0.9 being “reproducible” is also somewhat arbitrary. While this value may be justified for assessing reproducibility of some measurements (e.g., measuring the dimensions of an object), this cut-off value may not be ideal in our context. Models including covariates with a CCC lower than 0.9 could feasibly still generate valuable information and be independently validated. For example, a patient’s performance status has been shown to be vital in predicting survival. Yet, assigning performance status is quite subjective between physicians and may not be reproducible in the sense of quantitatively having a CCC index greater than 0.9. The summary of the results for Specific Aim 1 projects (1.1-1.3) are shown in Table 18.

Table 18. Summary of Results for Specific Aim 1 Projects

| Hypothesis | Result(s) |
|--|---|
| Project 1.1 Quantify the impact of adding CT-based quantitative image features to outcome models containing only CPFs including and excluding GTV | OS stratification: $p=0.046$ LRC stratification: $p=0.01$ FFDM stratification: $p=0.005$ |
| Project 1.2 Quantify the reproducibility of CT-based quantitative image features and its impact on outcome models | <u>% QIFs where $CCC > 0.9$</u> T _{avg} (85%) T ₅₀ (75%) CE-CT(23%) Reproducibility: 78-80% |
| Project 1.3 Quantify the prognostic value of adding CE-CT-based quantitative image features to outcome models containing only CPFs | Likelihood Ratio Test: $p = 0.027$ Stratification not improved |

While the addition of CT-based QIFs into survival models has shown significant potential, various downsides do exist. Data is still only available from preliminary studies, which require external validation and appropriate assessment of predictive power/accuracy. Careful consideration, of values for parameters involved with each methodology as well as whatever preprocessing steps are applied, needs be taken in deciding which QIFs/analysis methods are appropriate for particular tasks. Differing quantification methods and their associated parameters have the potential to greatly impact study results. Quantitative analysis is also not applicable to all patients. Those with a small primary tumor or severe imaging artifacts are not appropriate to undergo analysis. Advances in robust, auto-segmentation methods would also be exceedingly useful in this field in order to standardize tumor contouring. Physician-generated contours are most commonly used in these types of analyses but are far from perfect. Thresholding is a useful strategy to enhance contour reproducibility, particularly in lung tumors. Other factors that would influence reproducibility but not included in our analysis are the stability of the CPFs between institutions/physicians, such as staging, performance status, tumor volume, etc. Additionally, in

this study the imaging protocols were well controlled. The impact of changing image parameters (e.g., tube voltage, reconstruction algorithm, pixel size, manufacturer, etc.) as well as consistency of QIFs should be considered when evaluating data from multiple institutions. Multiple groups have investigated the impact of variation in imaging on feature reproducibility.^{41, 48, 52} Balagurunathan et al. found that only 30% of features examined were reproducible using test/retest scans with a criteria of having a CCC greater than or equal to 0.9 and approximately one third of these features were redundant.⁵² Hunter et al. identified that QIFs are not only dependent on the scanner type but also image phase on 4D-CT.⁴¹ Care should always be taken when determining the appropriateness of feature calculations used for each application.

Our work builds upon the preliminary evidence shown in recent publications supporting the use of CT-based QIFs in NSCLC.^{10, 14, 15, 48–50} Our chosen patients are different in that our cohorts are comprised only of patients deemed stage III rather than multiple stages. Furthermore, we examined QIFs in relation to a more substantial list of CPFs whereas most of the literature does not perform multivariate comparisons or adjust for one or two CPFs such as volume and/or staging. Our analysis of cohort 1 found that the LOG_Average feature was significant across multiple outcomes. This supports the findings of Ganeshan et al. that have multiple reports of this feature type being significant not only in NSCLC but in other disease sites such as liver, breast, and esophageal cancers.^{10, 14, 15, 17, 19, 27, 28} Our analysis of cohort 2 found that histogram uniformity was significant in multivariate analysis relating to patient survival. Aerts et al. used this feature (called total energy in their work [not to be confused with our COM energy feature]) as part of a radiomics signature that was found to be prognostic in both NSCLC and head and neck cancer cohorts in addition to relating to tumor gene-expression.⁸ Furthermore, their energy feature alone (i.e. not in a radiomics signature) was able to stratify patient survival in both NSCLC and head and neck cancer cohorts and these results were independently validated. However, this work was done using non-contrasted CT scans.

Large, prospective studies are required in order to fully understand the potential impact that CT-based QIFs could have on outcome prediction models. This work represents a good foundation from

which prospective studies could be based on due to the use of cross-validation and preliminary analysis of reproducibility. Further work needs to investigate QIFs' role as a potential source of prognostic information as well as ways to ensure/correct for variation in features due to differences in scanners, reconstruction, phase (in 4D-CT), etc. Taking into account feature robustness alongside prognostic potential is necessary. Data needs to be collected to assess whether images and/or QIFs can be "normalized" and/or "corrected" to a particular baseline in terms of scanner, reconstruction, etc. to facilitate large scale investigations. Issues relating to the use of contrast material also need additional exploration. Very little is known regarding the impact of injection timing or the use of contrast versus non-contrasted scans. A cohort of NSCLC patients receiving sequential CT scans in the area of their primary tumor after contrast injection would be valuable in determining the impact of scan timing post injection on QIFs. Furthermore, a cohort of patients receiving both non-contrasted and CE-CT and comparing the resulting QIFs should be performed to assess the impact of contrast on feature values.

Our work provides some insight as to the influence of contrasted vessels on prognostic CT-based QIFs, but additional work validating our findings/methods increase our understanding of what these features may or may not be ultimately measuring. In general, more uniformity across all aspects of feature analysis (such as homogenizing features extracted, feature nomenclature, feature formulas, feature parameters used, modeling techniques, image acquisition, etc.) needs to be implemented for the field to advance. This process of homogenizing workflow should be the focus of future research in order to properly vet the ability of QIFs to provide prognostic information in addition to what is already known from CPFs.

Discussion Specific Aim 2

We hypothesized that the addition of FDG-PET-based quantitative image features will significantly improve outcome models compared to models using conventional prognostic factors. This hypothesis was confirmed as the addition of FDG-PET-based quantitative image features significantly improved outcome models compared to models using conventional prognostic factors. Incorporating pretreatment PET QIFs alongside CPFs in survival models enabled improved model fit and better

stratification of patients in terms of overall survival compared to models using CPFs alone. The use of cross-validation allows the use of all data in both training and testing and thus is more efficient than splitting data into independent test and validation sets. The results from cross-validation should more aptly reflect how the model would perform in an independent cohort comprised of similar patients.

Recent data has suggested that quantification of intratumoral heterogeneity may yield prognostic information that could improve prediction or response and/or prognosis in patients with NSCLC.^{12, 21, 23} It is hypothesized that tumor heterogeneity in FDG-PET tracer uptake may reflect underlying tumor biology such as hypoxia, angiogenesis, and necrosis.¹⁴ Therefore, these methods could be used to identify tumors that are predisposed to aggressive behavior. In NSCLC specifically, preliminary data suggests a relationship between QIFs and patient outcome.^{12, 21} However, these studies do not sufficiently adjust for CPFs when assessing significance of new QIFs. This work is unique in that significant effort was made to generate multivariate models that implement both QIFs and CPFs to assess the added benefit of QIFs to models using CPFs. Furthermore, prediction models frequently utilize cohorts comprised of patients of varying stages whereas our cohort is comprised solely of stage III NSCLC. Models capable of stratifying patients that are homogeneously staged may be more clinically useful as different stages of disease frequently dictate different treatment courses. Furthermore, solidity and COM energy were consistently selected during cross-validation and conventional PET metrics, such as SUVmean, SUVmax, MTV, etc. , were not selected with nearly the same frequency. This observation suggests that perhaps QIFs examining spatial heterogeneity of uptake may be more predictive than conventional PET metrics when adjusting for CPFs. Solidity quantifies how dispersed the primary and nodal disease are in a local region context (all stage III patients). COM energy quantifies the uniformity of the SUV values within the primary tumor while taking into account the spatial orientation of the voxels. The COM energy metric is calculated by determining the probabilities for different voxel-adjacent voxel-pairs within the tumor, squaring these values, and summing them together. Therefore, a completely uniform tumor would have a COM energy of 1 while a heterogeneous tumor where few adjacent voxels have the same SUV value would have a COM energy value that is very small.

Use of QIFs from routinely obtained images has the potential to provide value to clinical practice without any added expense or radiation exposure. Pretreatment risk stratification could enable clinicians to deliver more patient-specific treatment tailored to individual risk. Particularly in advanced NSCLC patients, accurate predictions might aid in determining the appropriate level of treatment aggressiveness and maintaining as much of a patient's quality of life as possible. Additionally, more accurate prediction models could ensure more balanced and/or appropriate treatment arms in prospective trials.

Prediction models that include QIFs have been shown to have significant potential; however, a few limitations should be noted. First, most of the evidence for the prognostic ability of QIFs (including this study) comes from retrospective reviews and not from prospective assessment. Additionally, in order to generate sizable cohorts, several studies have used patient data acquired on a variety of scanners implementing various/outdated reconstruction parameters (e.g., differing voxel sizes or use of 2D reconstruction). Recently, literature has emerged that found these differences have a significant impact on the reproducibility of the extracted QIFs.^{53–56} Leijenaar et al. have analyzed the reproducibility of QIFs using test-retest scans and specifically found that the COM energy feature had an intra-class correlation coefficient (this is the same as the CCC used in this work) of 0.96.

However, our work found using “pseudo” test/retest scans that COM energy had a much lower CCC (0.56). This result suggests that this feature may be “portable” enough to have broad implementation but normalizing acquisition/reconstruction parameters may be beneficial. Specific Aim 2.2 found that a majority of features are reasonably reproducible even when using scans separated by time, reconstruction, scanner type, etc. Numerous publications suggest that standard quantitative FDG-PET features, such as SUV_{mean} , SUV_{max} , etc., are variable across scanners/institutions.^{57–59} Therefore, the observation that QIFs were reasonably reproducible in our “pseudo” test/re-test cohort was initially surprising considering the nonhomogeneous nature of the protocols/scanners used. The reason for this observation was found to be differences in the metric used to define reproducibility. For instance, the CCC index for SUV_{mean} for all test/re-test pairs was 0.85, which is viewed as reproducible. However, the average absolute percent difference for SUV_{mean} was 21%, which is similar to observations in the literature.⁵⁷

Tumor delineation on PET images is also less than straightforward. Many delineation methods exist such as manual contouring, value thresholding, percentage thresholding, and a variety of other semi-automated techniques. This work used a semi-automated gradient technique because a review by Werner-Wasik et al. found this method to be the most robust in terms of accuracy and consistency for NSCLC tumors.⁴² This study found PETedge to be superior to manual contour and thresholding due to its lower percent error in segmented volume and very low systematic bias. Variations in tumor delineation could easily influence the extracted QIFs.

This investigation generated retrospective reconstruction datasets from a NEMA IEC phantom as well as patient scans. The purpose of this data was to determine feature reproducibility in a more controlled manner than our “pseudo” test/re-test dataset. The phantom data did a reasonable job of replicating the reproducibility seen within the patient reconstructions. A majority of features were found to have a high ratio of the standard deviation of clinical patient QIFs to standard deviation of QIFs from phantom/patient reconstructions. COM correlation and homogeneity performed the poorest, having standard deviation ratios of approximately two or less. These features may not be sufficiently robust to quantify features from a variety of sources due to their sensitivity to changes in reconstruction. While the ratio of standard deviations may seem high, it is important to realize that the potential for substantial changes still exists. Our investigation only evaluated reconstruction parameters in routine clinical use. Images generated using reconstruction parameters outside the limits of those investigated may still generate QIF values that vary substantially from images generated using more routine reconstructions. One should also consider the implications of using a ratio of standard deviations between patient data and reconstructed data. Substantial percent changes in a particular feature can exist (e.g., SUV max change of 30%) and still yield a reasonably high ratio of standard deviations. The standard deviation ratio takes into account the variance seen within patients and therefore a 30% change may not be substantial when taking into account the range of values seen clinically. While changes in most QIFs due to the different reconstruction techniques investigated appear to be minor, future studies should strive to collect imaging data with as limited variation in reconstructions/parameters as possible. The balance between

homogenizing imaging and patient numbers is something that should be considered on a project-to-project basis. Future work determining if images can be retrospectively normalized to a particular baseline may prove beneficial.

One important observation that should also be noted is that COM energy and uniformity of FDG-PET are very closely correlated. Since uniformity was seen to be more reproducible, it may be wise to analyze both COM energy and uniformity in future studies. However, uniformity was found to be more susceptible to changes in tumor volume than COM energy. The summary of the results for Specific Aim 1 Hypotheses (2.1-2.3) are shown in Table 19.

Table 19. Summary of Results for Specific Aim 2 Hypotheses

| Hypothesis | Result(s) |
|--|--|
| Project 2.1 Quantify the impact of adding FDG-PET-based quantitative image features to outcome models containing only CPFs | Likelihood Ratio Test: OS: $p=0.007$ |
| Project 2.2 Quantify the reproducibility of FDG-PET-based quantitative image features using “pseudo” test-retest scans | 12/15 (80%) of QIFs did not have a $CCC > 0.9$ CCC of 2D-3D reconstruction (average = 0.72) vs 3D-3D (average = 0.93) |
| Project 2.3 Quantify the reproducibility of FDG-PET-based quantitative image features using retrospective reconstructions of phantom and patient data | 3/9 (33%) of QIFs in NEMA phantom and 2/9 (22%) QIFs in patients did not have ratio of standard deviations greater than 3 |

This promising work has several limitations. First, retrospective data derived only from a single institution cohort is hypothesis generating. Proper validation using a sizeable independent cohort of patients is needed. Second, we originally considered 26 distinct QIFs and did not solely perform our analysis using the two QIFs found to be predictive. However, the use of cross-validation for simultaneous

multivariate selection of these features should provide a better assessment of predictive model fit than re-substitution statistics.

The reproducibility of most FDG-PET QIFs was found to be reasonable considering the variation seen in features from Cohort 3. The “pseudo” test/retest analysis found reproducibility values that were lower than what has been previously reported in the literature. This is likely due to additional sources of variability such as scanner manufacturer/reconstruction methodology, scan timing, tumor growth, change in underlying FDG-uptake, etc., that are not present in other publications examining this issue. Feature variability within Cohort 3 is likely somewhere in between what was observed in the “pseudo” test/retest cohort and patient data using retrospective reconstructions.

Our work provides additional data supporting the use of FDG-PET QIFs in NSCLC to the growing body of literature currently available.^{12, 21, 23, 35} Similar to our work performed in CT, the analyzed FDG-PET cohort from section 4.2.1 was comprised entirely of stage III patients. Kang et al. also investigated FDG-PET based QIFs in a cohort comprised solely of stage III NSCLC patients.²¹ They found the area under the curve of the cumulative SUV histogram (AUC-CSH) was an independent prognostic factor for progression-free survival, locoregional-recurrence free survival, and distant metastases-free survival ($p < 0.05$). We examined the AUC-CSH metric in section 4.2.1 but it was not selected using our methodology in terms of its association with patient survival. AUC-CSH was also not found to be associated with overall survival, local-regional control, or freedom from distant metastases on univariate analysis. One reason why the prognostic ability of this metric was not able to be validated could be that Kang et al. extensively used optimal cut-offs for AUC-CSH as well as other conventional prognostic factors. This type of methodology has been associated with an increase in Type I error.⁶⁰ In addition, Tixier et al. also did not find AUC-CSH to be significant predictor of overall survival or recurrence free survival.²³

Other work has been performed examining the prognostic ability of FDG-PET based QIFs in NSCLC in patients with multiple stages.^{12, 23} Cook et al. investigated 4 nearest gray tone difference matrix features relationship to tumor response along with overall survival, progression-free survival, and local

progression-free survival. Tumor coarseness, contrast, and complexity were found to be significantly different between responders and non-responders ($p < 0.05$). Tumor contrast had the largest area under the curve for predicting tumor response ($AUC = 0.82$). Tumor coarseness, contrast, and complexity were all found to be significantly predictive of one or more of the outcomes measured. However, the relationship between these QIFs and patient outcomes were tested using optimal-cutoffs in a univariate manner. These methods could potentially be problematic and have led to overly optimistic results.

Accurate knowledge of a patient's prognosis is a valuable tool in medicine and particularly in oncology. We demonstrated that QIFs extracted from pretreatment PET images enhance stratification of patients based on overall survival compared to CPFs. Appropriate use of these models could greatly aid the treating clinicians and the patients themselves. More studies need to be conducted to validate PET derived QIFs and determine whether these techniques could one day be implemented clinically.

Moving forward, additional research should be conducted to standardize all aspects of feature analysis (such as homogenizing features extracted, feature nomenclature, feature formulas, feature parameters used, modeling techniques, image acquisition, etc.). This is very similar to the standardization needed in CT and is perhaps even more difficult. FDG-PET has scanner-based variation just like CT but since FDG-PET measures a biological process it has inherent biologic variation from the patient as well as variation stemming from the use of a radiotracer (variation injection-scan time interval, injected dosage, etc.). A large cohort of patients with imaging from multiple scanners, injection timings, reconstruction parameters, etc. would be useful in determining thFeature/methodology standardization analyses need to be conducted in order for large scale retrospective/prospective studies to reach their full potential in terms of advancing the field.

Discussion Specific Aim 3

We hypothesized that there would be significant relationships between some quantitative image features between modalities and with tumor volume, staging, and morphologic characteristics. This hypothesis was confirmed as significant relationships were found between some quantitative image features between modalities and with tumor volume, staging, and morphologic characteristics. A theme

throughout this work is the importance of not only determining the prognostic value of QIFs but to quantify their significance *in addition* to CPFs. Furthermore, it is important to gain an understanding of *why* these features seem to be prognostic and if features from different modalities are *related*. FDG-PET based QIFs identified in Specific Aim 2 were significantly correlated with the LOG_Average extracted from CE-CT ($p < 0.05$). Solidity as measured on FDG-PET was also significantly correlated with CE-CT uniformity ($p < 0.05$). A variety of significant associations were also seen between QIFs (from FDG-PET and CE-CT) and CPFs such as volume and staging and morphologic characteristics, such as visualized necrosis and vascularity.

The QIFs extracted in Specific Aims 1 and 2 seem to be related not only to one another in some capacity but also to CPFs and morphologic characteristics of tumors. The fact that QIFs have associations with CPFs is not a surprise nor does it invalidate our findings since QIFs were shown to have prognostic potential beyond CPFs. However, it is important to be cognizant of the fact that these relationships do exist in order to not overstate possible conclusions.

One of the more interesting results from Specific Aim 3 is the identification of a relationship between FDG-PET-based QIFs and the LOG_Average and Uniformity QIFs extracted from CE-CT. This relationship links features from a predominantly anatomical imaging modality (CT) to a functional imaging modality (FDG-PET). The linkage from anatomical to functional supports a hypothesis that morphologic characteristics/phenotypes of tumors can relate to underlying functional/biologic phenotypes. However, while some associations were statistically significant the correlation coefficients were quite low. To our knowledge, there is only one publication in the literature that suggests that two different imaging modalities may be related in terms of their extracted QIFs.¹⁰ However, this was done in non-contrasted CT and was not done using factors found to be significant in any sort of outcome analysis.

We were able to demonstrate that morphologic characteristics of tumors such as enhancing vasculature, necrosis, and air within tumor cavitation can influence the two examined QIF measurements (LOG_Average and histogram based uniformity). LOG_Average was found to be influenced by these morphologic characteristics of tumors to a greater extent than histogram uniformity. Tumors with necrosis

and enhancing vasculature demonstrated an increase in LOG_Average ($p < 0.0001$ and $p < 0.0001$, respectively) but did not have the same magnitude of impact on histogram uniformity ($p=0.19$ and $p = 0.03$, respectively). The presence of air within cavitated tumors led to a decrease in uniformity ($p = 0.0001$). Tumor morphologic characteristics, such as presence of enhancing vasculature and necrosis, may contribute to the underlying reason why QIFs appear to be prognostic in a variety of tumors. Our results demonstrated a significant difference in QIF values between tumors with and without air within cavitated tumors, necrosis, and enhancing vasculature.

Excluding morphologic characteristics was found to alter the QIFs measured from tumors; however, this was only significant when examining LoG Average and excluding tumors with a large volume of vessels. Excluding air, necrosis, and/or vessels from the analyzed contour did not impact the resulting QIF values substantially compared to examining only “tumor tissue”. The tumor tissue alone appeared to explain a vast majority of the resulting QIF values for LOG_Average and histogram heterogeneity compared to analyzing the entire tumor ($R\text{-squared} = 0.92$ and 0.96 , respectively).

At first glance this result may seem inconsistent; however, this could be due to the interaction between these morphologic characteristics and the tumor tissue itself. For example, the presence of enhancing vessels or necrosis within the tumor may lead to heterogeneity in contrast uptake within the tumor tissue causing more “edges” being measured by LOG_Average. This can be observed in Figure 1. Tumors with higher than average volume of vessels (black points in Figure 33) have higher values of LOG_Average than tumors with lower than the average volume of vessels (gray points in Figure 33) ($p < 0.05$). The same pattern can also be observed in Figure 33 for necrosis ($p < 0.05$). These types of processes may contribute to the values obtained from QIFs.

Literature attempting to determine what is fundamentally being measured from QIFs from CT in NSCLC or what tissue-related factors may influence QIF measurement is scant. Ganeshan et al. and van Gomez et al. both examined the relationship between CT-based QIFs and SUV features from FDG-PET.^{10, 35} Ganeshan et al. found that coarse texture features correlated with mean tumor SUV and van Gomez et al. found correlations between a variety of CT-based QIFs and FDG-PET features. A separate

publication from Ganeshan et al. established relationships between CT-based QIFs and hypoxia markers from FDG-PET and pathology.¹⁴ Additionally, Aerts et al. observed that certain QIF clusters are associated with primary tumor stage, overall stage, and histology. Aerts et al. also found that the QIFs incorporated into their “radiomics signature” were found to have high normalized enrichment scores in their analysis of gene expression.⁸ However, a majority of these publications were performed in non-CE-CT whereas our study was performed using contrast enhanced scans.

The correlated nature between CE-CT and FDG-PET QIFs suggests that analysis of both modalities in tandem has the potential to provide additional prognostic information.

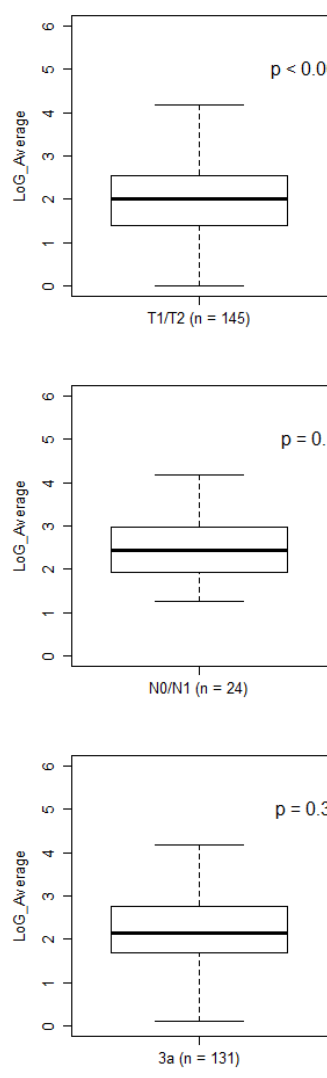
When designing the autosegmentation algorithms, we found that the most difficult portion was differentiating necrosis from tumor tissue. The Hounsfield unit values of air and enhancing vessels are substantially different from tissue and therefore amenable to a simple thresholding approach. However, tumor necrosis contains Hounsfield unit values that are only slightly lower than the rest of the tumor tissue. Tumor necrosis is more easily identified on FDG-PET than on CT (see Figure 3). Therefore, establishing that both autosegmentation methods on both image types resulted in similar values of necrosis volume and percentage of necrosis within the tumor is encouraging and supports the accuracy of segmentation on CE-CT (CCC = 0.85 and 0.76, respectively). The decrease in CCC value between necrosis volume and percentage of tumor containing necrosis could be due to variability in delineating the tumor itself on CT versus PET and not the autosegmentation. Furthermore, necrotic regions with low FDG-avidity may not always appear on CE-CT as having lower Hounsfield unit values which would cause discrepancies in quantifying necrosis volume. Overall, the agreement seen between the two methodologies was reasonable based on their CCC values and supports the accuracy of CE-CT autosegmentation.

Quantifying and relationships and/or correlations between QIFs from different modalities, CPFs, and morphologic characteristics are needed in order to increase our understanding of how and why QIFs appear to provide prognostic information. Furthermore, these types of studies are of paramount

importance if QIFs are ever to be optimized in terms of their preprocessing and method of quantification.

The summary of the results for Specific Aim 3 Hypotheses are shown in Table 20.

Table 20. Summary of Results for Specific Aim 3 Hypotheses

| Hypothesis | Result(s) |
|---|--|
| Project 3.1. Quantify correlations between prognostic FDG-PET-based and CECT-based quantitative image features | See Table 15 |
| Project 3.2. Quantify if relationships exist between CE-CT-based and FDG-PET-based quantitative image features with tumor volume and TNM staging | <p>See</p>  <p>Figure 28,</p> |

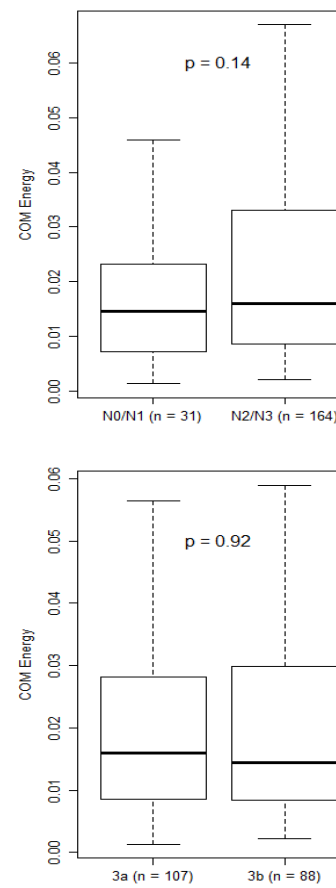


Figure 29

Project 3.3. Quantify if there are correlations between FDG-PET-based quantitative image features, CECT-based quantitative image features, and morphologic characteristics (vessels, necrosis, air cavities, etc.)

See Figure 30, Figure 31

While this work is a good initial step, there are several limitations. This was done retrospectively using only 78 patients with both a CE-CT and FDG-PET taken at a single institution. This does not address whether significant associations/correlations would be observed across a wider array of patients, scanners, etc. Statistical significance in terms of correlation or separation of groups also does not necessarily translate to clinical significance. For instance, while LOG_Average and COM Energy were found to be significantly correlated, this does not imply that one could use these features interchangeably or that the prognostic value of one feature is comparable to another.

While we believe our segmentation methodology to be sufficiently robust due to the correlation of necrosis volumes between CT and FDG-PET segmentations, further validation of its accuracy would be useful. Possible avenues to validate this segmentation would be to consult a radiologist for their opinion on segmentations across a range of patient tumors or examine pathology from excised lesions. In addition, enhancing the autosegmentation to be more robust to image artifacts, alternative reconstruction filters (i.e. standard versus lung versus bone), and non-contrasted scans would increase its utility. As the reconstructions rely heavily on HU cutoffs, the ability to adapt these cutoffs on an image-to-image basis may also improve its accuracy.

Discussion Specific Aim 4

We hypothesized that significant FDG-PET-based based quantitative image features found in Specific Aim 2 would allow for identification of sub-cohorts that will demonstrate a significant stratification of patients based on radiation dose. This hypothesis was confirmed as the significant FDG-PET-based based quantitative image features (COM energy and solidity) allowed for identification of sub-cohorts that demonstrated a significant survival stratification of patients based on radiation dose. Quantitative image features from CT and PET have been shown to be prognostic in a variety of solid tumors including NSCLC.^{12, 21, 23, 25} To our knowledge, this work is the first to examine the possible influence that these factors could have on modifying treatment. This study found that there was no difference in overall and progression free survival between patients being treated with 74Gy vs 60-70Gy

when examining the entire cohort. When examining subgroups of our cohort based on values of QIFs, we found that dose escalation benefits those with high values of COM energy and solidity (i.e. those who have a higher predicted survival based on our previous work) and is detrimental to those with low values of COM energy and solidity. We found that receiving 74Gy versus 60-70Gy was an independent prognostic factor both overall survival (high value QIFs: $p = 0.012$, low value QIFs: $p = 0.02$) and progression free survival (high value QIFs: $p = 0.015$, low value QIFs: $p = 0.025$) in a multivariate Cox proportional hazards model.

The literature is conflicting whether or not escalating dose yields an improved survival in patients with NSCLC. It has been suggested that an increase in the delivered radiation dose may improve patient survival in NSCLC.⁶¹ RTOG 0617 examined whether dose escalation improved survival in stage III NSCLC patients treated with chemoradiation.⁶² This study found that treating to 74Gy versus 60Gy led to an increased incidence of grade 3 pneumonitis was detrimental to patient survival at one year (70.4% versus 81%, respectively). The reasons for these surprising results are the subject of intense debate.⁶³ There has been the suggestion that patient heart dose could be responsible for the reduced survival. Speirs et al found that cardiac dosimetric parameters such as mean/max heart dose were significant using univariate analysis for overall survival but this association was not seen when using multivariate techniques adjusting for factors accounting for tumor volume.⁶⁴ However, a study by Liao et al found that the use of IMRT led to an improvement in overall survival compared to 3D conformal radiation therapy (3D-CRT) ($p = 0.039$).⁶⁵ One could hypothesize that this may have been due to the dosimetric improvement allowed by IMRT compared to 3D-CRT in the cardiopulmonary structures. In our study comprised almost exclusively of IMRT patients, we did not observe a survival difference between those treated to at least 74Gy versus 60-70Gy. It is quite apparent that patient survival is contingent upon a multifactorial process and not purely radiation dose. In order to optimize treatment for patients, it would be advantageous to know if certain populations of stage III NSCLC are more likely to benefit from escalating radiation treatment dose.

The present work was able to demonstrate that subgroups with certain values of disease solidity and primary COM energy can yield significantly different survival rates based on the use of dose escalation. Figure 35 and Figure 37 illustrate that these observations were not obtained by merely selecting a subgroup in which a survival difference was seen but that an overall trend is evident regarding the impact of dose escalation with high/low values of these two QIFs. It should also be noted that since we chose different percentile cutoffs for the dashed outline cells in Figure 35 and Figure 37, there are patients (4) that are in both figures. This was done merely to display the concept with a reasonable sample size in each plot. We also could have displayed subgroups where the cutoffs were the same and therefore there would be no patient overlap. For example, using the values of 45% COM energy and 55% solidity in both Figure 35 and Figure 37, the separation would be significant ($p < 0.05$) for high/low values of QIFs in terms of overall and progression free survival however this would only yield sample sizes of 42 and 52 patients, respectively. The summary of the results for Specific Aim 4 Hypothesis are shown in Table 21.

Table 21. Summary of Results for Specific Aim 4 Hypotheses

| Hypothesis | Result(s) |
|--|--|
| Project 4.1. Assess whether significant PET-based quantitative image features relate to a difference in patient survival for those treated with an escalated radiation dose | See Figure 35, Figure 36, Figure 37, Figure 38 |

While the results of our work are interesting and have the potential to allow physicians to better select patients for dose escalation, there are several limitations. First and foremost this study is retrospective in nature and therefore has all the limitations that are associated with retrospective studies such as possible selection biases, attrition bias, methodological changes, etc. Our cohort also had a mixture of patients treated with both proton and photon therapies whereas most studies examining dose escalation are comprised of patients treated with photon therapy. Due to the retrospective nature and variations in patient treatment within this study it should be seen exclusively as hypothesis generating

work. In the future, we hope to extract dosimetric and toxicity data from our patients to determine if any trends exists relating to radiation dose and whether these metrics could also be useful in better selecting patients for treatment.

The use of pretreatment QIFs to better select patients for different treatments would be a substantial advance in how radiation oncologists treat patients. We demonstrated that the use of QIFs from pretreatment FDG-PET scans in stage III NSCLC patient's scans may allow clinicians to better predict who would benefit from a higher radiation dose. Further work is needed in order to validate these findings and better understand what factors influence which patients should be treated to a standard or escalated dose.

There are numerous limitations of this work that need to be considered. More variation in feature values could be obtained if images were expanded to include various manufacturers, reconstruction parameters, etc. A large, independent cohort would be needed to fully validate the findings from this work and further exploration is needed into the applicability of the metrics across institutions.

As echoed in the previous discussion sections, feature/methodology standardization is needed in order to properly put these results into context and potentially validate their results. A future study confirming these results as well as identifying a specific cutoff of both solidity and COM energy should be performed before any attempt of prospective assessment.

Discussion Overall

Almost all previous studies examining the relationship of QIFs to patient outcome are performed in patients staged differently according to AJCC staging. Stratifying patients of the same stage is fundamentally more difficult than stratifying patients of different stages for a couple of reasons: 1. There will be inherently more variation in terms of patient outcome for patients of various stages and 2. Patients of different stages are frequently accompanied by differences in non-cancer related prognostic factors such as performance status and comorbidities. Models capable of stratifying patients who are homogeneously staged may also be more clinically useful as different stages of disease frequently dictate different treatment courses. Every effort was made to try and quantify the *improvement* that QIFs can

make to outcome models when adjusting for an extensive list of CPFs. Furthermore, proper validation, whether using cross-validation or independent cohorts, was implemented whenever possible. A majority of studies in the literature do not comprehensively investigate the added benefit of QIFs but merely examine their impact alone or adjusting for 1-2 CPFs such as primary tumor volume or TNM staging. Most publications also only focus on features relating to the primary tumor and disregard information from nodal disease, which has been shown to relate to patient survival.^{4, 66} Numerous issues exist when using re-substitution statistics or optimal cut-offs and these issues permeate through the existing literature.

Some may argue that, for those with a poor predicted prognosis, this knowledge is usually not beneficial since patients are already receiving the maximum tolerable treatment. I would argue that treatment optimization does not necessarily equate to treatment escalation. In some instances, perhaps de-escalation of treatment and initiation of early palliative care may provide the best care for the patient. An accurate prognosis would also be beneficial to patients and their caregivers. Those identified as having a poor prognosis may be better equipped to make decisions regarding palliative versus definitive treatment and the role of hospice care. Section 4.4.1 found that QIFs may be potentially useful in modifying treatment for identified patients. This modification of treatment was shown to include both potential escalation and de-escalation in terms of ultimately providing patient benefit.

The major message from this work is that many factors can and do contribute to patient outcome/survival. This work explored not only novel QIFs but also CPFs and these factors are only the tip of the iceberg. Other indicators, such as genetics, social factors, blood-based markers, etc., all may potentially play a role in determining outcome. Furthermore, the cohort of patients being analyzed also plays a significant role in the ultimate findings. In our work, for example, the role of GTV in the CE-CT (Cohort 2) versus FDG-PET (Cohort 3) was found to be drastically different even when these patients were extremely similar in terms of CPFs. These prognostic discrepancies in basic CPFs really highlight the need of performing these types of analyses on extremely large patient cohorts.

There is a desperate need for the development of large-scale cohorts that contain not only patient clinical data but patient imaging, pathology, and genetics. A big data approach to the field of quantitative image analysis (a.k.a. radiomics) would be a tremendous advance. The use of natural language processing for efficient extraction of patient clinical data, pathology, and genetics would increase the ease of database construction. Automated workflows for extracting images from radiation treatment planning systems and/or PACs followed by robust and validated disease segmentation algorithms should be employed in the future to facilitate high throughput data collection. Ideally, the use of a single quantitative analysis software would be ideal but simply standardizing feature nomenclature and calculation across platforms would be a good first step. Databases exist that contain a portion of the required information such as the Surveillance, Epidemiology, and End Results (SEER) program. SEER could be used as a model in the generation of future databases with more comprehensive information needed to truly develop and investigate aspects personalized patient care.

Ultimately, outcome model development is an extremely difficult process that requires knowledge in a variety of areas. Researchers need to be familiar with both the medical and quantitative/analysis sides of the problem if they hope to conduct work that may one day translate into clinical benefit. The aforementioned work found encouraging results in terms of analyzing the prognostic value of QIFs from pretreatment CT and FDG-PET scans in stage III NSCLC patients undergoing definitive radiation therapy but has many potential caveats and drawbacks as discussed previously. This work has the potential to one day improve the quality of care for these patients and the processes described are translatable to other stages and/or tumor types. Additional investigation and/or optimization regarding standardization of image acquisition(s), quantification methods, and statistical analysis/validation are necessary and encouraged in order to realize the maximum potential of using image-based features for improving personalized cancer care.

References

- ¹ R. Siegel, J. Ma, Z. Zou, and A. Jemal, "Cancer statistics, 2014.," *CA. Cancer J. Clin.* **64**(1), 9–29.
- ² J.R. Egner, *AJCC Cancer Staging Manual*, JAMA J. Am. Med. Assoc. **304**, 1726 (2010).
- ³ L.A.G. Ries, J.L. Young Jr, G.E. Keel, M.P. Eisner, Y.D. Lin, and M.-J.D. Horner, "Cancer survival among adults: US SEER program, 1988-2001," *Patient tumor Charact. SEER Surviv. Monogr. Publ.* 7–6215 (2007).
- ⁴ T. Berghmans, M. Paesmans, and J.-P. Sculier, "Prognostic factors in stage III non-small cell lung cancer: a review of conventional, metabolic and new biological variables.," *Ther. Adv. Med. Oncol.* **3**, 127–138 (2011).
- ⁵ C. Oberije, G. Nalbantov, A. Dekker, L. Boersma, J. Borger, B. Reymen, A. van Baardwijk, R. Wanders, D. De Ruyscher, E. Steyerberg, A.-M. Dingemans, and P. Lambin, "A prospective study comparing the predictions of doctors versus models for treatment outcome of lung cancer patients: a step toward individualized care and shared decision making.," *Radiother. Oncol.* **112**(1), 37–43 (2014).
- ⁶ Y. Zhang, Y. Feng, Y. Zhang, X. Ming, J. Yu, D.J. Carlson, J. Kim, and J. Deng, "Is It the Time for Personalized Imaging Protocols in Cancer Radiation Therapy?," *Int. J. Radiat. Oncol.* **91**(3), 659–660 (2015).
- ⁷ R.L. Schilsky, "Implementing personalized cancer care," *Nat. Rev. Clin. Oncol.* **11**(7), 432–438 (2014).
- ⁸ H.J.W.L. Aerts, E.R. Velazquez, R.T.H. Leijenaar, C. Parmar, P. Grossmann, S. Cavalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebors, M.M. Rietbergen, C.R. Leemans, A. Dekker, J. Quackenbush, R.J. Gillies, and P. Lambin, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nat. Commun.* **5**, (2014).
- ⁹ R.M. Haralick, K. Shanmugam, and I. Dinstein, "Textural Features for Image Classification," *IEEE Trans. Syst. Man. Cybern.* **3**, (1973).
- ¹⁰ B. Ganeshan, S. Abaleke, R.C.D. 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 (2010).
- ¹¹ Amadasum, "Textural features corresponding to textural properties - Systems, Man and Cybernetics, *IEEE Transactions on*," **19**(5), (1989).
- ¹² G.J.R. Cook, C. Yip, M. Siddique, V. Goh, S. Chicklore, A. Roy, P. Marsden, S. Ahmad, and D. Landau, "Are pretreatment 18F-FDG PET tumor textural features in non-small cell lung cancer associated with response and survival after chemoradiotherapy?," *J. Nucl. Med.* **54**(1), 19–26 (2013).
- ¹³ S. Tan, S. Kligerman, W. Chen, M. Lu, G. Kim, S. Feigenberg, W.D. D'Souza, M. Suntharalingam, and W. Lu, "Spatial-temporal [¹⁸F]FDG-PET features for predicting pathologic response of esophageal cancer to neoadjuvant chemoradiation therapy.," *Int. J. Radiat. Oncol. Biol. Phys.* **85**(5), 1375–82 (2013).

- 14 B. Ganeshan, V. Goh, H.C. Mandeville, P.J. Hoskin, and K.A. Miles, "Non – Small Cell Lung Cancer : Histopathologic Correlates for Texture," **266**(1), (2013).
- 15 B. Ganeshan, E. Panayiotou, K. Burnand, S. Dizdarevic, and K. Miles, "Tumour heterogeneity in non-small cell lung carcinoma assessed by CT texture analysis: a potential marker of survival.," *Eur. Radiol.* **22**(4), 796–802 (2012).
- 16 O.S. Al-Kadi and D. Watson, "Texture analysis of aggressive and nonaggressive lung tumor CE CT images.," *IEEE Trans. Biomed. Eng.* **55**(7), 1822–30 (2008).
- 17 T. Win, K. a Miles, S.M. Janes, B. Ganeshan, M. Shastry, R. Endozo, M. Meagher, R.I. Shortman, S. Wan, I. Kayani, P.J. Ell, and A.M. Groves, "Tumor heterogeneity and permeability as measured on the CT component of PET/CT predict survival in patients with non-small cell lung cancer.," *Clin. Cancer Res.* **19**(13), 3591–9 (2013).
- 18 F. Tixier, C.C. Le Rest, M. Hatt, N. Albarghach, O. Pradier, J.-P. Metges, L. Corcos, and D. Visvikis, "Intratumor heterogeneity characterized by textural features on baseline 18F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer.," *J. Nucl. Med.* **52**(3), 369–78 (2011).
- 19 B. Ganeshan, K. Skogen, I. Pressney, D. Coutroubis, and K. Miles, "Tumour heterogeneity in oesophageal cancer assessed by CT texture analysis: preliminary evidence of an association with tumour metabolism, stage, and survival.," *Clin. Radiol.* **67**(2), 157–64 (2012).
- 20 K.A. Miles, M.R. Griffiths, R.C.D. Young, and C.R. Chatwin, "Colorectal Cancer : Texture Analysis of Portal Phase Hepatic CT Images as a Potential Marker of Survival 1 Purpose : Methods : Results : Conclusion :," **250**(2), (2009).
- 21 S.-R. Kang, H.-C. Song, B.H. Byun, J.-R. Oh, H.-S. Kim, S.-P. Hong, S.Y. Kwon, A. Chong, J. Kim, S.-G. Cho, H.J. Park, Y.-C. Kim, S.-J. Ahn, J.-J. Min, and H.-S. Bom, "Intratumoral Metabolic Heterogeneity for Prediction of Disease Progression After Concurrent Chemoradiotherapy in Patients with Inoperable Stage III Non-Small-Cell Lung Cancer.," *Nucl. Med. Mol. Imaging* (2010). **48**(1), 16–25 (2014).
- 22 N.-M. Cheng, Y.-H.D. Fang, J.T.-C. Chang, C.-G. Huang, D.-L. Tsan, S.-H. Ng, H.-M. Wang, C.-Y. Lin, C.-T. Liao, and T.-C. Yen, "Textural features of pretreatment 18F-FDG PET/CT images: prognostic significance in patients with advanced T-stage oropharyngeal squamous cell carcinoma.," *J. Nucl. Med.* **54**(10), 1703–9 (2013).
- 23 F. Tixier, M. Hatt, C. Valla, V. Fleury, C. Lamour, S. Ezzouhri, P. Ingrand, R. Perdrisot, D. Visvikis, and C. Cheze Le Rest, "Visual Versus Quantitative Assessment of Intratumor 18F-FDG PET Uptake Heterogeneity: Prognostic Value in Non-Small Cell Lung Cancer.," *J. Nucl. Med.* (2014).
- 24 F. Tixier, M. Hatt, C. Valla, V. Fleury, C. Lamour, S. Ezzouhri, P. Ingrand, R. Perdrisot, D. Visvikis, and C.C. Le Rest, "Visual Versus Quantitative Assessment of Intratumor 18F-FDG PET Uptake Heterogeneity: Prognostic Value in Non-Small Cell Lung Cancer.," *J. Nucl. Med.* **55**(8), 1235–1241 (2014).

- 25 D. V. Fried, S.L. Tucker, S. Zhou, Z. Liao, O. Mawlawi, G. Ibbott, and L.E. Court, "Prognostic Value and Reproducibility of Pretreatment CT Texture Features in Stage III Non-Small Cell Lung Cancer.," *Int. J. Radiat. Oncol. Biol. Phys.* **90**(4), 834–42 (2014).
- 26 F. Davnall, C.S.P. 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**(6), 573–89 (2012).
- 27 B. Ganeshan, K. a Miles, R.C.D. Young, and C.R. Chatwin, "Texture analysis in non-contrast enhanced CT: impact of malignancy on texture in apparently disease-free areas of the liver.," *Eur. J. Radiol.* **70**(1), 101–10 (2009).
- 28 B. Ganeshan, 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 (2011).
- 29 V. Goh, P. Nathan, J.K. Juttla, and K.A. Miles, "Assessment of Response to Tyrosine Kinase Inhibitors in Metastatic Renal Cell Cancer : CT Texture as a Predictive Biomarker 1 Methods : Results : Conclusion :," **261**(1), 165–171 (2011).
- 30 H. Wang, X.-H. Guo, Z.-W. Jia, H.-K. Li, Z.-G. Liang, K.-C. Li, and Q. He, "Multilevel binomial logistic prediction model for malignant pulmonary nodules based on texture features of CT image.," *Eur. J. Radiol.* **74**(1), 124–9 (2010).
- 31 H.-J. Im, K. Pak, G.J. Cheon, K.W. Kang, S.-J. Kim, I.-J. Kim, J.-K. Chung, E.E. Kim, and D.S. Lee, "Prognostic value of volumetric parameters of (18)F-FDG PET in non-small-cell lung cancer: a meta-analysis.," *Eur. J. Nucl. Med. Mol. Imaging* (2014).
- 32 L.-F. de Geus-Oei, H.F.M. van der Heijden, F.H.M. Corstens, and W.J.G. Oyen, "Predictive and prognostic value of FDG-PET in nonsmall-cell lung cancer: a systematic review.," *Cancer* **110**, 1654–1664 (2007).
- 33 T. Berghmans, M. Dusart, M. Paesmans, C. Hossein-Foucher, I. Buvat, C. Castaigne, A. Scherpereel, C. Mascaux, M. Moreau, M. Roelandts, S. Alard, A.-P. Meert, E. Patz, J.-J. Lafitte, and J.-P. Sculier, "Primary Tumor Standardized Uptake Value (SUVmax) Measured on Fluorodeoxyglucose Positron Emission Tomography (FDG-PET) is of Prognostic Value for Survival in Non-small Cell Lung Cancer (NSCLC).," *J. Thorac. Oncol.* **3**, 6–12 (2008).
- 34 R.L. Wahl, H. Jacene, Y. Kasamon, and M.A. Lodge, "From RECIST to PERCIST: Evolving Considerations for PET response criteria in solid tumors.," *J. Nucl. Med.* **50 Suppl 1**(Suppl_1), 122S–50S (2009).
- 35 O. van Gómez López, A.M. García Vicente, A.F. Honguero Martínez, A.M. Soriano Castrejón, G.A. Jiménez Londoño, J.M. Udías, and P. León Atance, "Heterogeneity in [18F]Fluorodeoxyglucose Positron Emission Tomography/Computed Tomography of Non-Small Cell Lung Carcinoma and Its Relationship to Metabolic Parameters and Pathologic Staging.," *Mol. Imaging* **13**, 1–12 (2014).

- 36 S. Chicklore, V. Goh, M. Siddique, A. Roy, P.K. Marsden, and G.J.R. Cook, *Quantifying tumour heterogeneity in 18F-FDG PET/CT imaging by texture analysis*, Eur. J. Nucl. Med. Mol. Imaging **40**, 133–140 (2013).
- 37 S.B. Edge, D.R. Byrd, C.C. Compton, a G. Fritz, F.L. Greene, and a Trotti, *AJCC cancer staging manual*, 7th ed. (Springer-Verlag New York, 2010).
- 38 R.T.H. Leijenaar, G. Nalbantov, S. Carvalho, W.J.C. van Elmpt, E.G.C. Troost, R. Boellaard, H.J.W.L. Aerts, R.J. Gillies, and P. Lambin, “The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis,” Sci. Rep. **5**, 11075 (2015).
- 39 D. V Fried, O. Mawlawi, L. Zhang, X. Fave, S. Zhou, G. Ibbott, Z. Liao, and L.E. Court, “Stage III Non-Small Cell Lung Cancer: Prognostic Value of FDG PET Quantitative Imaging Features Combined with Clinical Prognostic Factors,” Radiology 142920 (2015).
- 40 L. Zhang, D. V. Fried, X.J. Fave, L.A. Hunter, J. Yang, and L.E. Court, “ibex: An open infrastructure software platform to facilitate collaborative work in radiomics,” Med. Phys. **42**(3), 1341–1353 (2015).
- 41 L.A. Hunter, S. Krafft, F. Stingo, H. Choi, M.K. Martel, S.F. Kry, and L.E. Court, “High quality machine-robust image features: identification in nonsmall cell lung cancer computed tomography images,” Med. Phys. **40**(12), 121916 (2013).
- 42 M. Werner-Wasik, A.D. Nelson, W. Choi, Y. Arai, P.F. Faulhaber, P. Kang, F.D. Almeida, Y. Xiao, N. Ohri, K.D. Brockway, J.W. Piper, and A.S. Nelson, “What is the best way to contour lung tumors on PET scans? Multiobserver validation of a gradient-based method using a NSCLC digital PET phantom,” Int. J. Radiat. Oncol. Biol. Phys. **82**, 1164–1171 (2012).
- 43 J.J. Sunderland and P.E. Christian, “Quantitative PET/CT scanner performance characterization based upon the society of nuclear medicine and molecular imaging clinical trials network oncology clinical simulator phantom,” J. Nucl. Med. **56**(1), 145–52 (2015).
- 44 R.M. Simon, J. Subramanian, M.-C. Li, and S. Menezes, “Using cross-validation to evaluate predictive accuracy of survival risk classifiers based on high-dimensional data,” Brief. Bioinform. **12**(3), 203–14 (2011).
- 45 T. Fushiki, “Estimation of prediction error by using K-fold cross-validation,” Stat. Comput. **21**, 137–146 (2009).
- 46 F.E. Harrell, K.L. Lee, and D.B. Mark, “Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors,” Stat. Med. **15**(4), 361–87 (1996).
- 47 L.I. Lin, “A concordance correlation coefficient to evaluate reproducibility,” Biometrics **45**(1), 255–268 (1989).
- 48 Y. Balagurunathan, Y. Gu, H. Wang, V. Kumar, O. Grove, S. Hawkins, J. Kim, D.B. Goldgof, L.O. Hall, R.A. Gatenby, and R.J. Gillies, “Reproducibility and Prognosis of Quantitative Features Extracted from CT Images,” Transl. Oncol. **7**(1), 72–87 (2014).

- 49 H.J.W.L. Aerts, J. Bussink, W.J.G. Oyen, W. van Elmpt, A.M. Folgering, D. Emans, M. Velders, P. Lambin, and D. De Ruyscher, "Identification of residual metabolic-active areas within NSCLC tumours using a pre-radiotherapy FDG-PET-CT scan: a prospective validation.," *Lung Cancer* **75**(1), 73–6 (2012).
- 50 S. Basu, L.O. Hall, D.B. Goldgof, Y. Gu, V. Kumar, J. Choi, R.J. Gillies, R.A. Gatenby, and A.D. Sets, "Developing a Classifier Model for Lung Tumors in CT-scan Images," 1306–1312 (2011).
- 51 M. Ravanelli, D. Farina, M. Morassi, E. Roca, G. Cavalleri, G. Tassi, and R. Maroldi, "Texture analysis of advanced non-small cell lung cancer (NSCLC) on contrast-enhanced computed tomography: prediction of the response to the first-line chemotherapy.," *Eur. Radiol.* **23**(12), 3450–5 (2013).
- 52 Y. Balagurunathan, V. Kumar, Y. Gu, J. Kim, H. Wang, Y. Liu, D.B. Goldgof, L.O. Hall, R. Korn, B. Zhao, L.H. Schwartz, S. Basu, S. Eschrich, R.A. Gatenby, and R.J. Gillies, "Test-Retest Reproducibility Analysis of Lung CT Image Features.," *J. Digit. Imaging* (2014).
- 53 F.J. Brooks and P.W. Grigsby, "The effect of small tumor volumes on studies of intratumoral heterogeneity of tracer uptake.," *J. Nucl. Med.* **55**(1), 37–42 (2014).
- 54 P.E. Galavis, C. Hollensen, N. Jallow, B. Paliwal, and R. Jeraj, "Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters.," *Acta Oncol.* **49**(7), 1012–6 (2010).
- 55 R.T.H. Leijenaar, S. Carvalho, E.R. Velazquez, W.J.C. van Elmpt, C. Parmar, O.S. Hoekstra, C.J. Hoekstra, R. Boellaard, A.L.A.J. Dekker, R.J. Gillies, H.J.W.L. Aerts, and P. Lambin, "Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter-observer variability.," *Acta Oncol.* **52**(7), 1391–7 (2013).
- 56 F. Tixier, M. Hatt, C.C. Le Rest, A. Le Pogam, L. Corcos, and D. Visvikis, "Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18F-FDG PET.," *J. Nucl. Med.* **53**(5), 693–700 (2012).
- 57 J.S. Scheuermann, J.R. Saffer, J.S. Karp, A.M. Levering, and B.A. Siegel, "Qualification of PET scanners for use in multicenter cancer clinical trials: the American College of Radiology Imaging Network experience.," *J. Nucl. Med.* **50**(7), 1187–93 (2009).
- 58 R. Boellaard, "Standards for PET image acquisition and quantitative data analysis.," *J. Nucl. Med.* **50 Suppl 1**(Suppl 1), 11S–20S (2009).
- 59 R. Boellaard, "Need for standardization of 18F-FDG PET/CT for treatment response assessments.," *J. Nucl. Med.* **52 Suppl 2**(Supplement_2), 93S–100S (2011).
- 60 D.G. Altman, B. Lausen, W. Sauerbrei, and M. Schumacher, "Dangers of Using 'Optimal' Cutpoints in the Evaluation of Prognostic Factors," *JNCI J. Natl. Cancer Inst.* **86**(11), 829–835 (1994).
- 61 M. Machtay, K. Bae, B. Movsas, R. Paulus, E.M. Gore, R. Komaki, K. Albain, W.T. Sause, and W.J. Curran, "Higher biologically effective dose of radiotherapy is associated with improved outcomes for locally advanced non-small cell lung carcinoma treated with chemoradiation: an

- analysis of the Radiation Therapy Oncology Group.,” *Int. J. Radiat. Oncol. Biol. Phys.* **82**(1), 425–34 (2012).
- ⁶² J.D. Bradley, R. Paulus, R. Komaki, G.A. Masters, K. Forster, S.E. Schild, J. Bogart, Y.I. Garces, S. Narayan, V. Kavadi, L.A. Nedzi, J.M. Michalski, D. Johnson, R.M. MacRae, W.J. Curran, H. Choy, and Radiation Therapy Oncology Group, “A randomized phase III comparison of standard-dose (60 Gy) versus high-dose (74 Gy) conformal chemoradiotherapy with or without cetuximab for stage III non-small cell lung cancer: Results on radiation dose in RTOG 0617.,” *ASCO Meet. Abstr.* **31**(15_suppl), 7501 (2013).
- ⁶³ J.D. Cox, “Are the results of RTOG 0617 mysterious?,” *Int. J. Radiat. Oncol. Biol. Phys.* **82**(3), 1042–4 (2012).
- ⁶⁴ C.K. Speirs, S. Rehman, A. Molotievschi, M.A. Velez, T.A. DeWees, D. Mullen, S. Fergus, J.D. Bradley, and C.G. Robinson, “Comprehensive Analysis of Dosimetric Predictors of Overall Survival for Stage III Non-Small Cell Lung Cancer (NSCLC) Treated With Definitive Radiation Therapy,” *Int. J. Radiat. Oncol.* **90**(1), S667 (2014).
- ⁶⁵ Z.X. Liao, R.R. Komaki, H.D. Thames, H.H. Liu, S.L. Tucker, R. Mohan, M.K. Martel, X. Wei, K. Yang, E.S. Kim, G. Blumenschein, W.K. Hong, and J.D. Cox, “Influence of technologic advances on outcomes in patients with unresectable, locally advanced non-small-cell lung cancer receiving concomitant chemoradiotherapy.,” *Int. J. Radiat. Oncol. Biol. Phys.* **76**(3), 775–81 (2010).
- ⁶⁶ S. Markovina, F. Duan, B.S. Snyder, B.A. Siegel, M. Machtay, and J.D. Bradley, “Regional Lymph Node Uptake of [18F]Fluorodeoxyglucose After Definitive Chemoradiation Therapy Predicts Local-Regional Failure of Locally Advanced Non-Small Cell Lung Cancer: Results of ACRIN 6668/RTOG 0235,” *Int. J. Radiat. Oncol.* **93**(3), 597–605 (2015).
- ⁶⁷ A.-S. Dewalle-Vignion, N. Yeni, G. Petyt, L. Verscheure, D. Huglo, A. Béron, S. Adib, G. Lion, and M. Vermandel, “Evaluation of PET volume segmentation methods: comparisons with expert manual delineations.,” *Nucl. Med. Commun.* **33**(1), 34–42 (2012).
- ⁶⁸ U. Nestle, S. Kremp, A. Schaefer-Schuler, C. Sebastian-Welsch, D. Hellwig, C. Rube, and C.-M. Kirsch, “Comparison of Different Methods for Delineation of 18F-FDG PET-Positive Tissue for Target Volume Definition in Radiotherapy of Patients with Non-Small Cell Lung Cancer,” *J. Nucl. Med.* **46**(8), 1342–1348 (2005).
- ⁶⁹ M. Wanet, J.A. Lee, B. Weynand, M. De Bast, A. Poncelet, V. Lacroix, E. Coche, V. Grégoire, and X. Geets, “Gradient-based delineation of the primary GTV on FDG-PET in non-small cell lung cancer: a comparison with threshold-based approaches, CT and surgical specimens.,” *Radiother. Oncol.* **98**(1), 117–25 (2011).
- ⁷⁰ M. Hatt, C. Cheze-le Rest, A. van Baardwijk, P. Lambin, O. Pradier, and D. Visvikis, *Impact of Tumor Size and Tracer Uptake Heterogeneity in 18F-FDG PET and CT Non-Small Cell Lung Cancer Tumor Delineation*, *J. Nucl. Med.* **52**, 1690–1697 (2011).

Appendix A: Matlab Code for CE-CT Autosegmentation

This function was used in the work performed in section 4.3.3 Results for Project 3.3: Quantify if there are correlations between FDG-PET-based quantitative image features, CECT-based quantitative image features, and morphologic characteristics (vessels, necrosis, air cavities, etc.) in order to segment the morphologic characteristics.

```
function [Air_out,Necrosis_out,Tissue_out,Vessel_out]=TissueSeg3(CDataSetInfo)
disp_img = 0;
Air_bounds=[0 950];%%% white
Tissue_bounds = [1020 1120];%%% blue
Vessel_bounds = [1120 2000];%%% green

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%Initiation of Outputs%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Tissue_out =
zeros(size(CDataSetInfo.ROIImageInfo.MaskData,1),size(CDataSetInfo.ROIImageInfo.MaskData,2),size(CDataSet
Info.ROIImageInfo.MaskData,3));
Vessel_out =
zeros(size(CDataSetInfo.ROIImageInfo.MaskData,1),size(CDataSetInfo.ROIImageInfo.MaskData,2),size(CDataSet
Info.ROIImageInfo.MaskData,3));
Air_out =
zeros(size(CDataSetInfo.ROIImageInfo.MaskData,1),size(CDataSetInfo.ROIImageInfo.MaskData,2),size(CDataSet
Info.ROIImageInfo.MaskData,3));

disp('1: Identifying necrotic regions')
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%Identify Necrosis in 3D%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Necrosis_use = necrosis3d(CDataSetInfo); %Separate Code (see Appendix B)
disp('2: Finished necrosis region growing')
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%Identify Air, Vessels, Tissue on each 2D slice%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
for slice = 1:size(CDataSetInfo.ROIInfo.MaskData,3);
    IMG=CDataSetInfo.ROIImageInfo.MaskData(:,:,slice);
    Mask = CDataSetInfo.ROIInfo.MaskData(:,:,slice);
    filt = fspecial('gaussian',[3 3],0.7);
    TumorTest1 = roifilt2(filt,IMG,Mask);
    if isempty(TumorTest1)
        TumorTest1 = zeros(size(Mask));
    end
    Mask(IMG<875) = 0;
    Maskfill=imfill(Mask,'holes');

    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
    %%%%Identify Air and Vessels%%%
    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
    TumorTest1 = double(TumorTest1).*double(Maskfill);
```

```

TumorTestinit = double(IMG).*double(Maskfill);
Air_use=TumorTestinit>Air_bounds(1) & TumorTestinit<=Air_bounds(2);
Air_use = imerode(Air_use,ones(2));Air_use = imerode(Air_use,ones(2));
Air_use = imdilate(Air_use,ones(2));Air_use = imdilate(Air_use,ones(2));
Air_use=imfill(Air_use,'holes');
Vessel1=TumorTest1>Vessel_bounds(1) & TumorTest1<=Vessel_bounds(2);
Vessel_use=double(bwareaopen(Vessel1,2));
%%%%%%
%%%%%Identify Tissue%%%%%
%%%%%%%%%
Tissue1=TumorTest1>Tissue_bounds(1) & TumorTest1<=Tissue_bounds(2);
Tissue_use=double(bwareaopen(Tissue1,5));
Tissue_use=abs(bwareaopen(abs(Tissue_use-1),20)-1);
Tissue_use = Tissue_use - (Tissue_use&(Air_use|Vessel_use|Necrosis_use(:,slice)));
%%%%%%%%%
%%%%% Define the outputs%%%%%
%%%%%%%%%
Vessel_out(:,slice) = Vessel_use;
Necrosis_out = Necrosis_use;
Air_out(:,slice) = Air_use;
Tissue_out(:,slice) = Tissue_use;
end
disp('3: Identified Air and Vessels')
%%%%%%%%%
%%%%%Check that Necrosis is Bounded%%%%%
%%%%%%%%%
for i = 1:size(CDataSetInfo.ROIInfo.MaskData,3);
    IMG=CDataSetInfo.ROIInfo.MaskData(:,i);
    Mask = CDataSetInfo.ROIInfo.MaskData(:,i);
    Mask(IMG<875) = 0;
    Maskfill=imfill(Mask,'holes');

    Regions=bwlabel(Necrosis_use(:,i));
    Seg_out = 3*Vessel_out(:,i)+Necrosis_out(:,i)+2*Air_out(:,i)+4*Tissue_out(:,i);
    for r = 1:max(Regions(:))
        region = Regions==r;
        SE = strel('disk', 4);
        Rdilated = imdilate(region,SE);
        perim = bwperim(Rdilated);
        [num,~]=hist(Seg_out(perim==1),[0,1,2,3,4]);
        if(sum(num(2:5))/sum(num(:))>0.75)
            disp(sum(num(2:5))/sum(num(:)))
        else
            Necrosis_use(:,i)=Necrosis_use(:,i)-region;
            Tissue_out(:,i) = Tissue_out(:,i)+region;
        end
    end
    Tissue_out(:,i) = imfill(Tissue_out(:,i));
end
disp('4: Checked that necrosis met constraints')
Necrosis_out = Necrosis_use;
Tissue_out = Tissue_out - (Tissue_out&(Air_out|Vessel_out|Necrosis_out));

%%%%%%%%%

```

```

%% %% % Initialize Colors for Display %% %%
%% %% % %% %% %% %% %% %% %% %% %%
Red = zeros(size(CDataSetInfo.ROIImageInfo.MaskData,1),size(CDataSetInfo.ROIImageInfo.MaskData,2),3);
Green = Red;
Blue = Red;
White = Red;
Red(:, :, 1) = ones(size(CDataSetInfo.ROIImageInfo.MaskData,1),size(CDataSetInfo.ROIImageInfo.MaskData,2));
Green(:, :, 2) = ones(size(CDataSetInfo.ROIImageInfo.MaskData,1),size(CDataSetInfo.ROIImageInfo.MaskData,2));
Blue(:, :, 3) = ones(size(CDataSetInfo.ROIImageInfo.MaskData,1),size(CDataSetInfo.ROIImageInfo.MaskData,2));
White(:, :, 1) = ones(size(CDataSetInfo.ROIImageInfo.MaskData,1),size(CDataSetInfo.ROIImageInfo.MaskData,2));
White(:, :, 2) = ones(size(CDataSetInfo.ROIImageInfo.MaskData,1),size(CDataSetInfo.ROIImageInfo.MaskData,2));
White(:, :, 3) = ones(size(CDataSetInfo.ROIImageInfo.MaskData,1),size(CDataSetInfo.ROIImageInfo.MaskData,2));

%% %% % %% %% %% %% %% %% %% %% %%
%% %% % Code for Displaying Results %% %%
%% %% % %% %% %% %% %% %% %% %% %%
if (disp_img == 1)
    for slice = 1:size(CDataSetInfo.ROIInfo.MaskData,3);
        image = slice;
        IMG = CDataSetInfo.ROIImageInfo.MaskData(:, :, image);
        Mask = CDataSetInfo.ROIInfo.MaskData(:, :, image);
        I = double(IMG).*double(Mask);
        figure(1); subplot(1,2,1);
        figure(1); imagesc(I,[900 1300]); colormap(gray); hold on;
        figure(1); type1 = imagesc(Red);
        set(type1, 'AlphaData', Necrosis_out(:, :, slice)*0.2);
        figure(1); type2 = imagesc(Green);
        set(type2, 'AlphaData', Tissue_out(:, :, slice)*0.2);
        figure(1); type3 = imagesc(Blue);
        set(type3, 'AlphaData', Vessel_out(:, :, slice)*0.2);
        figure(1); type4 = imagesc(White);
        set(type4, 'AlphaData', Air_out(:, :, slice)*0.5); hold off;
        subplot(1,2,2);
        imagesc(I,[900 1300]);
        pause(1)

    end
else
    disp('display off')
end

```


Appendix B: Matlab Code for CE-CT Necrosis Identification

This is a sub-function used in Appendix A: Matlab Code for CE-CT Autosegmentation.

```
function Necrosis_use=necrosis3d(CDataSetInfo)
disp_img =0;%set to zero in IBEX

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%Identify Initial Image and Mask Data%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
IMG=CDataSetInfo.ROIImageInfo.MaskData;
IMGinit = IMG;
Maskfill = CDataSetInfo.ROI BWInfo.MaskData;
Mask_noair = Maskfill;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%removal of air not encapsulated by tumor%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Maskfill(IMGinit<975) = 0;
Mask_noair(IMGinit<950)=0;
for i = 1:size(CDataSetInfo.ROIImageInfo.MaskData,3);
Maskfill(:,i) = imfill(Maskfill(:,i),'holes');
end
IMGinit(Maskfill==1 & IMGinit<975)=1005;%Fill encapsulated air with 1005

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%Establish guess for necrosis based off threshold and filtered image%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
Necr_bounds = [975 1020];
for i = 1:size(CDataSetInfo.ROIImageInfo.MaskData,3);
    filt = fspecial('gaussian',[5 5],1.5);
    if max(Maskfill(:,i))>0
        IMGinit(:,i) = roifilt2(filt,IMGinit(:,i),Maskfill(:,i));
    else
    end
end
use=IMGinit.*uint16(Mask_noair);
N_guess = (use>Necr_bounds(1)).*(use<=Necr_bounds(2));
for i = 1:size(CDataSetInfo.ROIImageInfo.MaskData,3);
    N_guess(:,i) = imfill(logical(N_guess(:,i)),'holes');
end
SE = strel('disk', 2);
N_guess = imdilate(imerode(N_guess,SE),SE);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%Determine if multiple regions within the guess exists, analyze largest%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
if(max(N_guess(:))>0)
    [L,num] = bwlabeln(N_guess);
    [value,~]=hist(nonzeros(L),1:num);
    biggest_region = find(max(value)==value);
    Necr_cent = L==biggest_region(1);
    N=BWcentroid(Necr_cent);
    disp('Into region growing')
    [~,Necrosis_use] = regionGrowing(use, [N(2) N(1) N(3)],35,[],false);
    disp('Out of region growing')
    for i = 1:size(CDataSetInfo.ROIImageInfo.MaskData,3);
```

```

    Necrosis_use(:,:,i)=imerode(Necrosis_use(:,:,i),ones(3));
    Necrosis_use(:,:,i) = bwareaopen(Necrosis_use(:,:,i),50);
    Necrosis_use(:,:,i) = imdilate(Necrosis_use(:,:,i),ones(3));
    Necrosis_use(:,:,i) = bwareaopen(Necrosis_use(:,:,i),50);
end
else
    Necrosis_use = zeros(size(N_guess,1),size(N_guess,2),size(N_guess,3));
end

%%%%%Necrosis_use is the output of this function%%%%%

%%%%%%%%%%%%%
%%%%%Display Results%%%%%
%%%%%%%%%%%%%

R = zeros(size(IMG,1),size(IMG,2),3);
R(:,:,1) = 1;
if(dispatch==1)%This is set to 0 within IBEX and therefore not executed
    for i = 1:size(CDataSetInfo.ROIImageInfo.MaskData,3);
        figure(1);
        subplot(1,3,1);imagesc(use(:,:,i),[900 1300]);colormap(gray);
        subplot(1,3,2);imagesc(IMG(:,:,i),[900 1300]);colormap(gray);
        subplot(1,3,3);imagesc(IMG(:,:,i),[900 1300]);colormap(gray); hold on
        subplot(1,3,3);red=imagesc(R);
        set(red, 'AlphaData', Necrosis_use(:,:,i)*0.2);
        pause(1); hold off
    end
end
end

```

Appendix C: R Code for Cross-Validation Technique

```

####Used for both CT and PET analyses###
####Import data###
datapath <- "Y:/NSCLC_Texture/November2014/"
data <- read.csv(paste(datapath,"PETall_HDFU.csv", sep = ""), header = T)

####Load Required Packages###
library("penalized")
library("survival")
source("http://bioconductor.org/biocLite.R")
biocLite("survcomp")
library("survcomp")
require(survival)
require(survcomp)
require(MASS)
require(penalized)
require(pensim)
require(Hmisc)

####Custom functions which may or may not be used###
source('Y:/NSCLC_Texture/Jul2014/make_formula.R', echo=FALSE)

#####set parameters#####
out = "surv" #outcome of interest (e.g. surv,dm,loco)
out_time <- eval(parse(text = paste("data$",out,"rtstartmos",sep = "")))
out_stat <- eval(parse(text = paste("data$",out,"stat",sep = "")))

####Initialize/reformat various covariates####
lorig<-dim(data)[2]
Solidity<-data$MVVolume/data$MConvexHullVolume3D
data[,lorig+1]<-Solidity
names(data)[lorig+1]<- "Solidity"

data[,lorig+2]<-data$Total.Dose/100*(1+(as.numeric(as.character(data$D.fx))/100)/10)
names(data)[lorig+2]<- "BED"
data$BED[which(data$D.fx=="split")]=92.1

data[,lorig+3]<-data$PGlobalStd/data$PGlobalMean
names(data)[lorig+3]<- "COV"

data[,lorig+4]<-data$PNecrVolume>0.5
names(data)[lorig+4]<- "NEC"

data[,which(names(data)=="MVVolume")]<-log2(data$MVVolume)
data[,which(names(data)=="MConvexHullVolume3D")]<-log2(data$MConvexHullVolume3D)
data[,which(names(data)=="GTV")]<-log2(data$GTV)
data[,which(names(data)=="PVVolume")]<-log2(data$PVVolume)
data[,which(names(data)=="PSurfaceArea")]<-log2(data$PSurfaceArea)
data[,which(names(data)=="MSurfaceArea")]<-log2(data$MSurfaceArea)

data$ECOG<-as.ordered(as.numeric(data$ECOG>0))
data$KPS<-as.factor(as.numeric(data$KPS<90))
data$Smoking<-as.ordered(data$Smoking)

```

```

data$PY<-as.numeric(data$PY)
data$Hist<-as.factor(data$Hist)
data$Tstage<-as.factor(as.numeric(data$Tstage>2))
data$Nstage<-as.factor(as.numeric(data$Nstage>1))
data$Overall<-as.factor(as.numeric(data$Overall))
data$Induction<-as.factor(data$Induction)
data$PGlobalStd<-as.numeric(data$PGlobalStd)
data$Gender<-as.numeric(data$Gender)
data$Solidity[data$Solidity>1]<-1

###Conditions for Inclusion###
rule1<-data$PVOLUME>log2(5)
data<-data[which(rule1),]
out_time <- out_time[which(rule1)]
out_stat <- out_stat[which(rule1)]

###select clinical and QIFs###
set.seed(5126)
vars<-names(data)[c(3:13,17,67:114,lorig+1,lorig+2,lorig+3,lorig+4)]
evaluate_clin<-select.list(vars,multiple=TRUE,graphics=TRUE,preselect=vars[c(1:3,5:12,62)])
evaluate_all<-select.list(vars,multiple=TRUE,graphics=TRUE,preselect=vars[c(1:3,5:64)])

###Initialize Cross-Validated scores###
predPENsurv_all <- rep(0,length(out_stat))
predPENsurv_clin <- rep(0,length(out_stat))
predPENscore_all <- rep(0,length(out_stat))
predPENscore_clin <- rep(0,length(out_stat))
termsOPTouter_clin = NULL
termsOPTouter_all = NULL

pb<-winProgressBar("SVM-Progress Bar","Initializing",0,100,0)

###Start of Cross-Validation###
for(i in 1:length(out_stat)){
  data_minus_fold = data[-i,]
  out_time_temp<-out_time[-i]
  out_stat_temp<-out_stat[-i]
  ###L1 penalization###
  opt_clin<-optL1(Surv(out_time_temp,out_stat_temp), penalized =
data.matrix(data_minus_fold[evaluate_clin]),data=data_minus_fold,standardize = TRUE,trace=TRUE,fold = 10)
  opt_all<-optL1(Surv(out_time_temp,out_stat_temp), penalized =
data.matrix(data_minus_fold[evaluate_all]),data=data_minus_fold,standardize = TRUE,trace=TRUE,fold = 10)
  ###predictions###
  predPENsurv_clin[i]<-survival(predict(opt_clin$fullfit,penalized=data.matrix(data[i,evaluate_clin])),36)
  predPENsurv_all[i]<-survival(predict(opt_all$fullfit,penalized=data.matrix(data[i,evaluate_all])),36)
  ###determination of coefficients used in each fold###
  coefs_clin<-names(coefficients(opt_clin$fullfit))
  coefs_all<-names(coefficients(opt_all$fullfit))
  termsOPTouter_clin<-c(termsOPTouter_clin,coefs_clin)
  termsOPTouter_all<-c(termsOPTouter_all,coefs_all)
  clin<-coefficients(opt_clin$fullfit)
  tex_clin<-coefficients(opt_all$fullfit)
  ###generation of linear predictors###
  if(length(clin)!=0){
    predPENscore_clin[i]<-rowSums(t(matrix(rep(clin,1),nrow =
length(clin)))*as.numeric(data.matrix(data[i,names(clin)])))

```

```

}
if(length(tex_clin)!=0){
  predPENscore_all[i]<-rowSums(t(matrix(rep(tex_clin,1),nrow =
length(tex_clin)))*as.numeric(data.matrix(data[i,names(tex_clin)])))
}

setWinProgressBar(pb,round(i/length(out_stat)*100),label = paste("Working on Fold",i,"of",length(out_stat),sep=
'))

}

###Calculate C-indices###
cindex_clin<-CTindex(out_time,out_stat,-predPENscore_clin)
cindex_all<-CTindex(out_time,out_stat,-predPENscore_all)

###k-means for stratification###
numclusters<-5
set.seed(1)
clin_cluster<-rep(0,length(predPENscore_clin))
all_cluster<-rep(0,length(predPENscore_clin))
clin<-kmeans(predPENscore_clin,numclusters)
all<-kmeans(predPENscore_all,numclusters)
for(i in 1:length(clin$centers)){
  clin_cluster[clin$cluster==order(clin$centers)[i]]<-i
}
for(i in 1:length(all$centers)){
  all_cluster[all$cluster==order(all$centers)[i]]<-i
}

###Plot of KM curves for generated clusters from CV linear predictors###
layout(matrix(c(1,2,3,3),2,2,byrow=TRUE),heights=c(2,1))
km.coxph.plot(formula.s=Surv(out_time,out_stat) ~ clin_cluster, data.s=data,x.label="Time (months)",
y.label="Overall Survival Probability", main.title="Overall Survival\n Conventional Prognostic Factors including
GTV", leg.pos="topright", leg.text=paste(c("Lowest Risk ", "Low/Medium Risk ", "Medium Risk
", "Medium/High Risk ", "Highest Risk ")),leg.inset=0, .col=c("black","darkgray"),.lty=c(1,1,3,3,4),.lwd =
c(4,5,5,5,6), show.n.risk=TRUE, n.risk.step=6, xlim = c(0,42),verbose=TRUE)
km.coxph.plot(formula.s=Surv(out_time,out_stat) ~ all_cluster, data.s=data,x.label="Time (months)",
y.label="Overall Survival Probability", main.title="Overall Survival\n Conventional Prognostic Factors including
GTV and QIFs", leg.pos="topright", leg.text=paste(c("Lowest Risk ", "Low/Medium Risk ", "Medium Risk
", "Medium/High Risk ", "Highest Risk ")),leg.inset=0, .col=c("black","darkgray"),.lty=c(1,1,3,3,4),.lwd =
c(4,5,5,5,6), show.n.risk=TRUE, n.risk.step=6, xlim = c(0,42),verbose=TRUE)

###Plot of generated c-indices###
plot(cindex_clin[1:5,1],cindex_clin[1:5,2],lty = 2,lwd = 3,type="b",pch = 16, cex = 1.5, main = "Time Based
Concordance Index\n Clinical Penalized CV",xlab="Minimum Time Difference for Patient-Patient Comparison
(Months)",ylab="C-Index",xaxt="n",ylim = c(0,1))
lines(cindex_all[1:5,1],cindex_all[1:5,2],type="b",lty = 1,lwd = 3,pch = 16,cex = 1.5, main = "Time Based
Concordance Index\n Clinical + QIF Penalized CV",xlab="Minimum Time Difference for Patient-Patient
Comparison (Months)",ylab="C-Index",xaxt="n",ylim = c(0,1))
legend("topright",c("CPFs including GTV & QIFs", "CPFs including GTV", "CPFs excluding GTV"),lty =
c(1,2,3),lwd = 3,cex=1.5)
axis(side = 1,at = seq(0,48,6))

###Identification of covariates selected in >50% of CV folds###
cnames<-table(termsOPTouter_clin)[order(table(termsOPTouter_clin))]
anames<-table(termsOPTouter_all)[order(table(termsOPTouter_all))]

```

```

clinvars<-cnames[which(cnames>length(out_stat)*0.5)]
allvars<-anames[which(anames>length(out_stat)*0.5)]

###development of "final" models###
clinform<-as.formula(paste("Surv(", "out_time, ", "out_stat", ")~", paste(names(clinvars), collapse="+"), sep = "))
allform<-as.formula(paste("Surv(", "out_time, ", "out_stat", ")~", paste(c(names(allvars), "Overall"), collapse="+"), sep
= "))

###Likelihood ratio test comparing models (need to be nested)###
anova(coxph(clinform, data=data), coxph(allform, data=data))

```

Appendix D: Relationship of Cardiothoracic Dosimetry with Disease Solidity

The delivered radiation therapy treatment plans for patients in Cohort 3 were retrospectively analyzed. Treatment planning contours were used to determine the mean lung dose, V20 (percent of the total volume of lung receiving 20 Gy), and the mean heart. We were able to obtain these metrics for 193/195 patients for mean lung dose and V20 and 190/195 patients for mean heart dose. Solidity was calculated on each patient's pretreatment FDG-PET using the primary and nodal tumor contours as previously described (3.5 Region of Interest Contouring on PET). We hypothesized that dispersed disease (i.e. low values of solidity) would increase the dose delivered to cardiothoracic normal tissues.

Scatter plots were generated comparing GTV versus solidity in terms of mean lung dose, V20, and mean heart dose (Figure 39, Figure 40, and Figure 41, respectively). The first and third quartiles of the dosimetric variables were used as cutoffs for categorizing the value as low, medium, or high. This categorization was illustrated by the different point colors within the generated scatter plots.

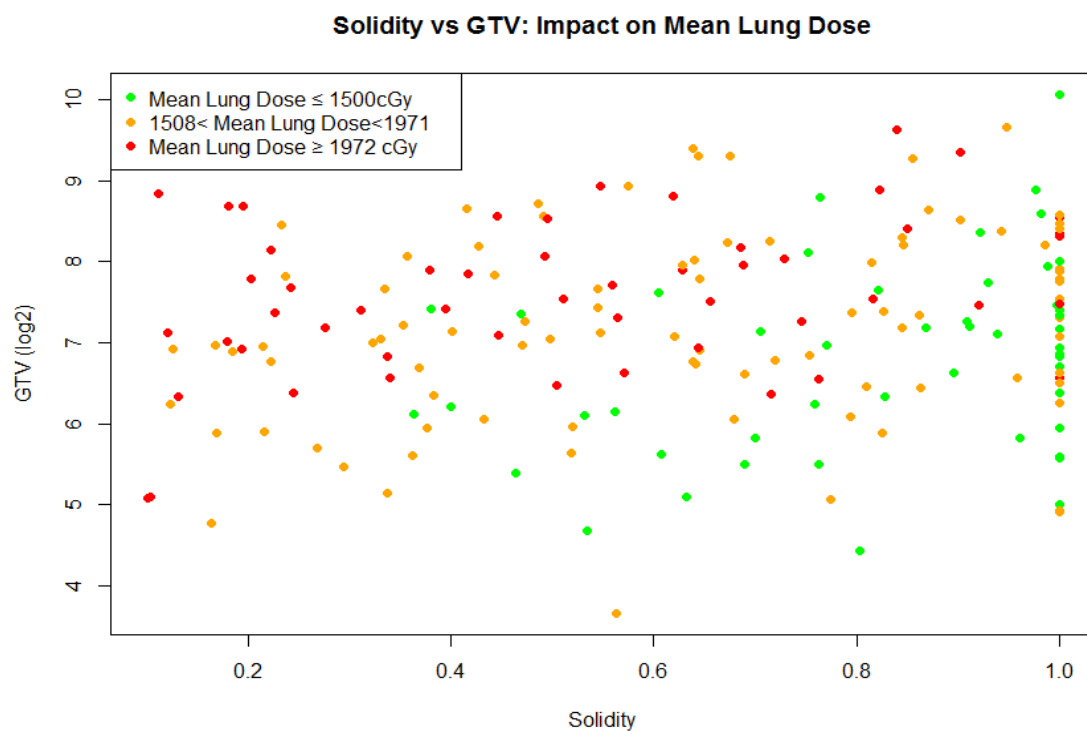


Figure 39. GTV versus solidity in terms of mean lung dose

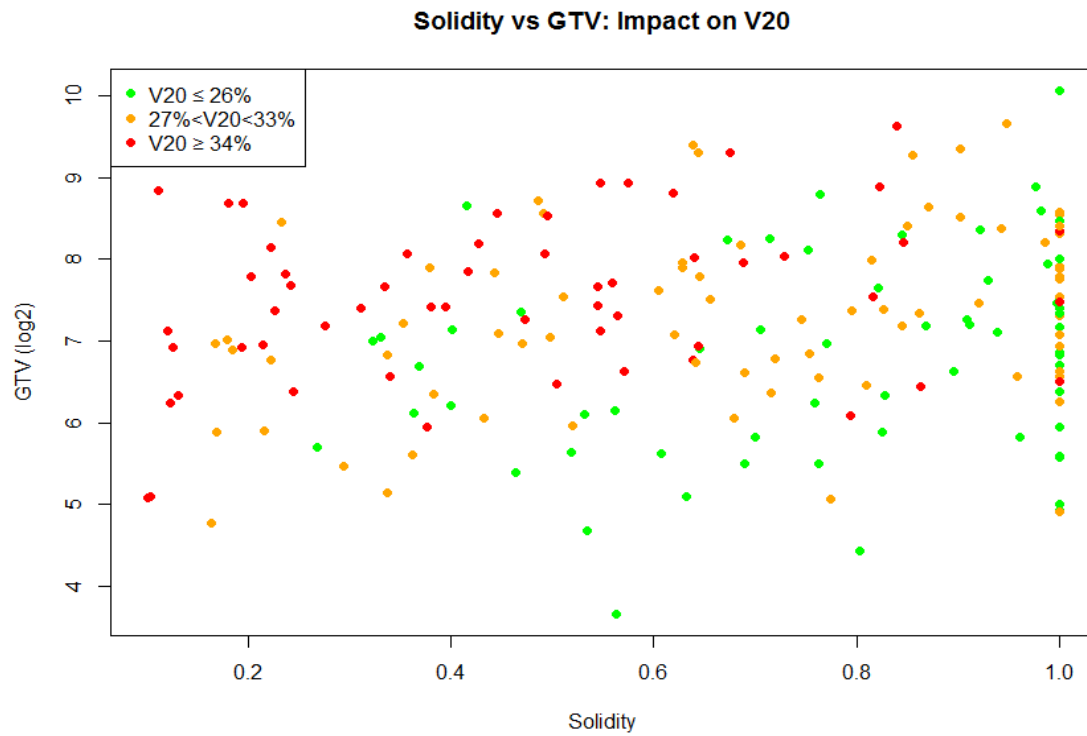


Figure 40. GTV versus solidity in terms of lung V20

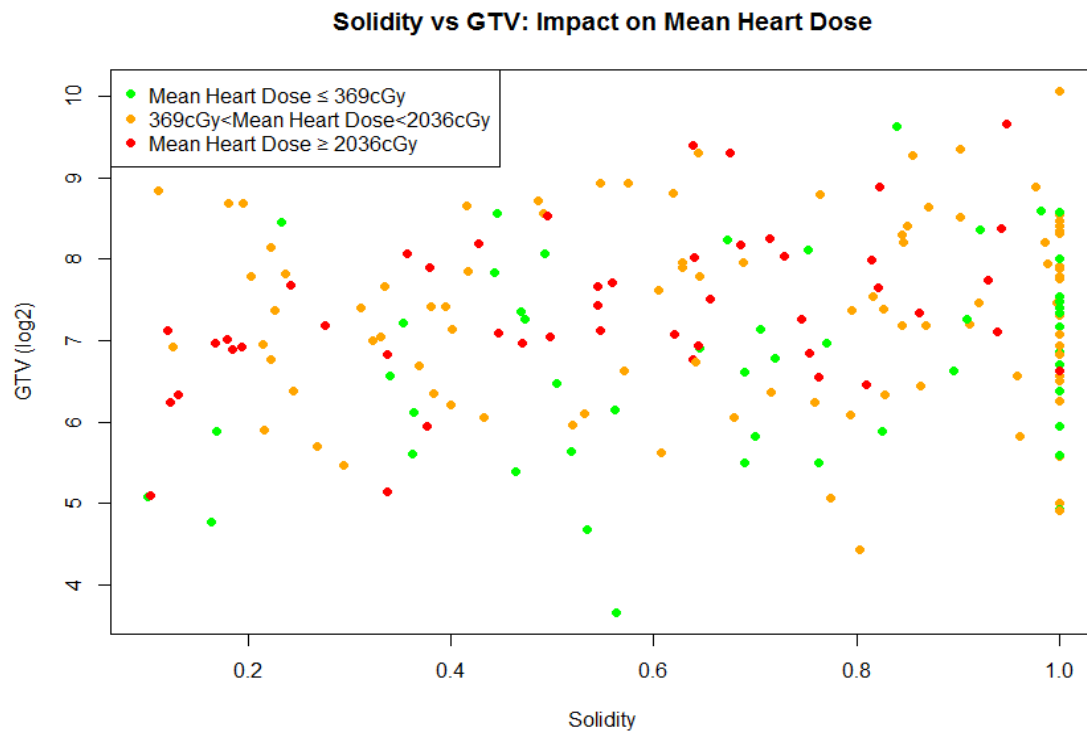


Figure 41. GTV versus solidity in terms of mean heart dose

For both the mean lung dose and V20, it can be qualitatively observed that patients in the upper quartile (red) are those with large volume and low solidity and those in the lower quartile (green) are those with low volume and high solidity. This pattern is not observed to the same extent in regard to mean heart dose. This is probably due to the increased importance of proximity to the heart which was not taken into account.

Boxplots were also generated using a risk score (Figure 42, Figure 43, Figure 44). Patients with a GTV greater than the median and a solidity value less than the median were assigned a score of 2, patients with either a GTV greater than the median or a solidity value less than the median were assigned a score of 1, and those with a GTV less than the median and a solidity value greater than the median were assigned a score of 0. Generally, patients with large **AND** dispersed disease were given a score of two, patient with either larger **OR** dispersed disease were given a score of one, and those with smaller, compact disease were given a score of zero. This was performed for mean lung dose, V20, and mean heart dose. Differences in values between risk score groups were assessed using unpaired t-tests.

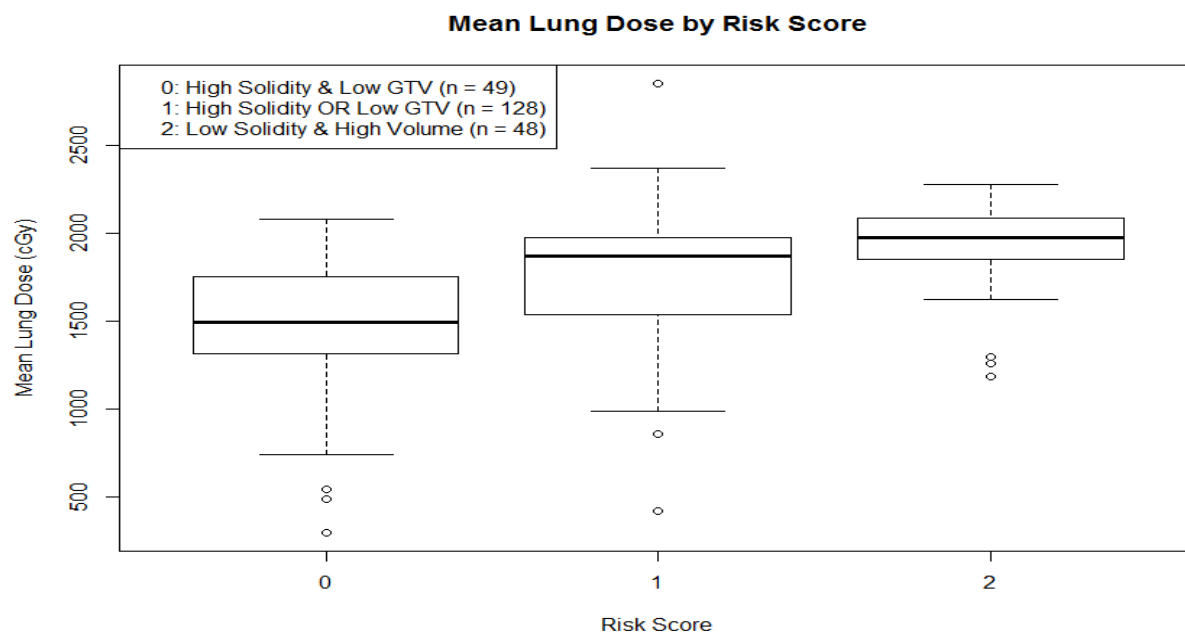


Figure 42. Mean lung dose stratified by risk score

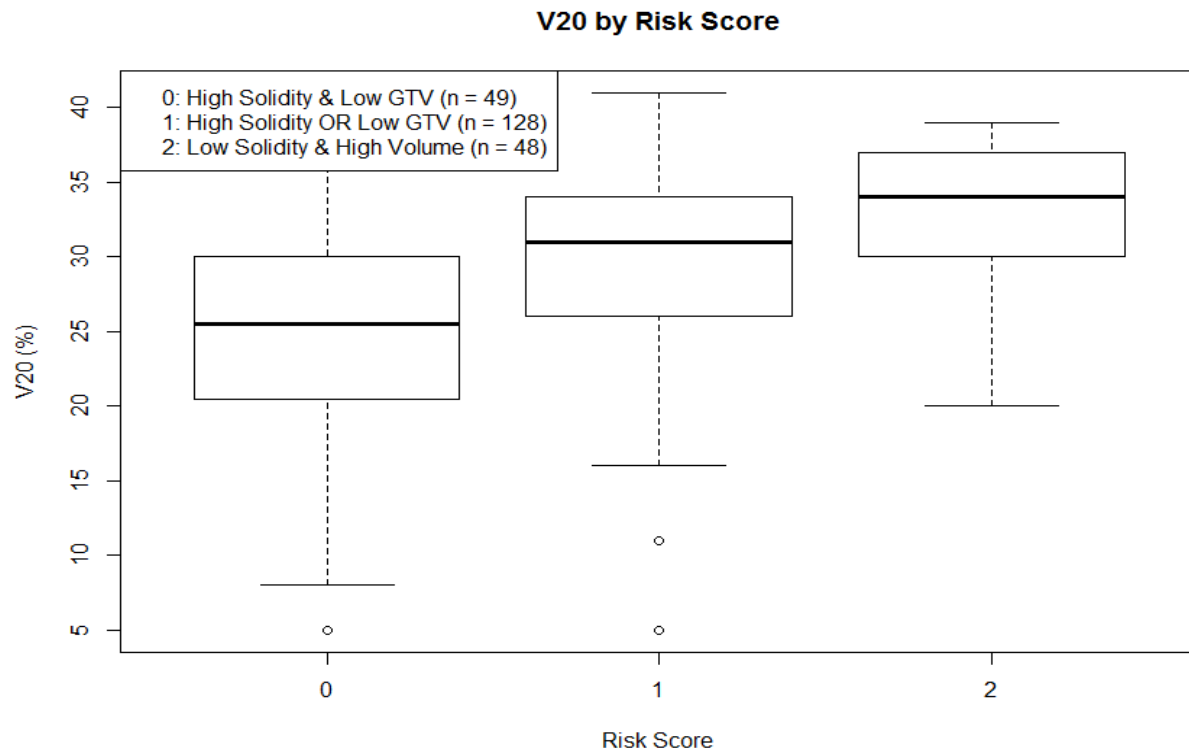


Figure 43. Lung V20 stratified by risk score

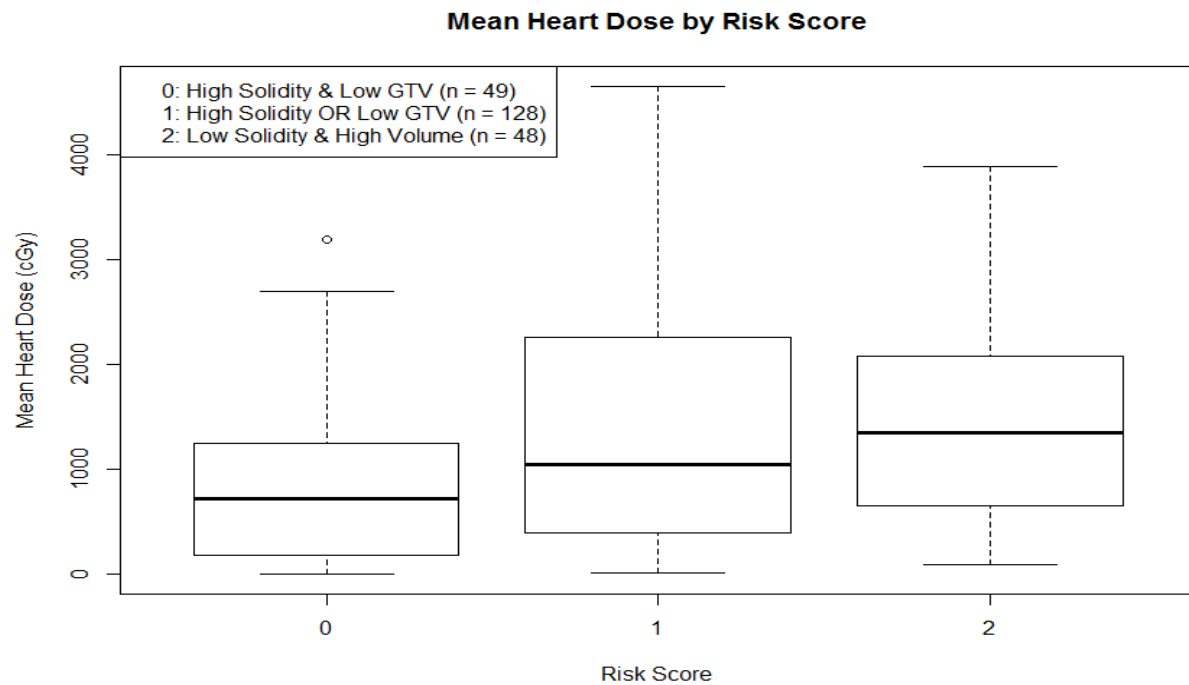


Figure 44. Mean heart dose stratified by risk score

For mean lung dose and V20, significant differences ($p < 0.05$) were seen between all three risk score groups. For mean heart dose, there was a significant difference between patients with risk scores of 0 versus 1, but there was no statistically significant difference between patients with risk scores of 1 versus 2.

Lastly, we generated linear regression models relating solidity to mean lung dose, V20, and mean heart dose while controlling for GTV. The p-values for decreasing solidity leading to an increase in cardiothoracic dosimetry values while controlling for GTV in the regression models were 2.5×10^{-9} , 2.8×10^{-11} , and 0.002 for mean lung dose, V20, and mean heart dose, respectively. Solidity is therefore considered to be an independent predictor of all three cardiothoracic dosimetry values even when controlling for the volume of disease. This confirmed our hypothesis that more dispersed disease would increase the dose to cardiothoracic normal tissues.

Appendix E: Comparison of FDG-PET Delineation Methods

A variety of methods exist for delineating tumors on FDG-PET. We examined common semi-automatic methodologies ($\text{SUV} \geq 2.5$, 40% SUV_{max} , 50% SUV_{max}) and compared the values obtained in terms of volume, uniformity, COM energy, and SUV_{mean} compared to the PETedge method used in our work using the concordance correlation coefficient (CCC). We examined 20 patients with a volume distribution similar to our entire 195 patient cohort in terms of range and median/mean of tumor volumes. A summary table of the CCC values is shown below.

Table 22. CCC Values for Comparison of Delineation Methodologies

| | Volume | Uniformity | COM Energy | SUV_{mean} |
|---|---------------|-------------------|-------------------|--|
| $\text{SUV} \geq 2.5$ | 0.89 | 0.91 | 0.94 | 0.7 |
| 40% SUV_{max} | 0.82 | 0.77 | 0.77 | 0.88 |
| 50% SUV_{max} | 0.65 | 0.64 | 0.75 | 0.78 |

It was observed that delineating using $\text{SUV} \geq 2.5$ consistently overestimated the volume of the tumors and underestimated the SUV_{mean} compared with when PETedge was used. Delineating using 40% and 50% SUV_{max} consistently underestimated the tumor volume, overestimated the uniformity (in terms of uniformity and COM energy) and overestimated the SUV_{mean} . The analyzed segmentation methods all yielded reasonably different results than those obtained by PETedge for the assessed metrics (volume, uniformity, COM energy, and SUV_{mean}). Furthermore, these segmentation methods may also have overly optimistic results due to the caveats described below.

There are some caveats to this analysis. First, I generated not only the PETedge contours but also the other contours from various delineation methods. While semi-automatic, the other methods are still very reliant on the bounding box applied. Since I had a general notion of the contour result from PETedge, I may have biased the size and location of the bounding box to tightly conform to what I believed the contour should look like based on PETedge. Furthermore, MIMvista automatically adjusts

the window/level of the images based on the values within the image and therefore may result in more consistent values than software this feature is not available and the window/level settings are more user dependent. Overall, the results may overestimate the accuracy of these methods due to the caveats mentioned. There is substantial evidence in the literature regarding the insufficiencies of the methods tested.^{42, 67-69}

Appendix F: Sequential FDG-PET Analysis

Ninety seven patients at the time of analysis were enrolled in a protocol entitled “A Bayesian Randomized Trial of Image-Guided Adaptive Conformal Photon vs Proton Therapy, with Concurrent Chemotherapy, for Locally Advanced Non-Small Cell Lung Carcinoma”. Patients were excluded for small (<5cc) initial primary tumor volume, if they did not have an FDG-PET scan taken either during treatment (~30 days from initiation) or post treatment, or if their scans were performed with 2D reconstruction. We wanted to assess whether we would be able to observe changes in QIFs across different time points. Having only twenty two patient did not really allow for us to sufficiently correlate changes in QIFs with any outcomes. In general, it was observed that tumors decreased in volume, SUVmax, and SUVmean while becoming more uniform in terms of uniformity and COM energy (see Figures, below).

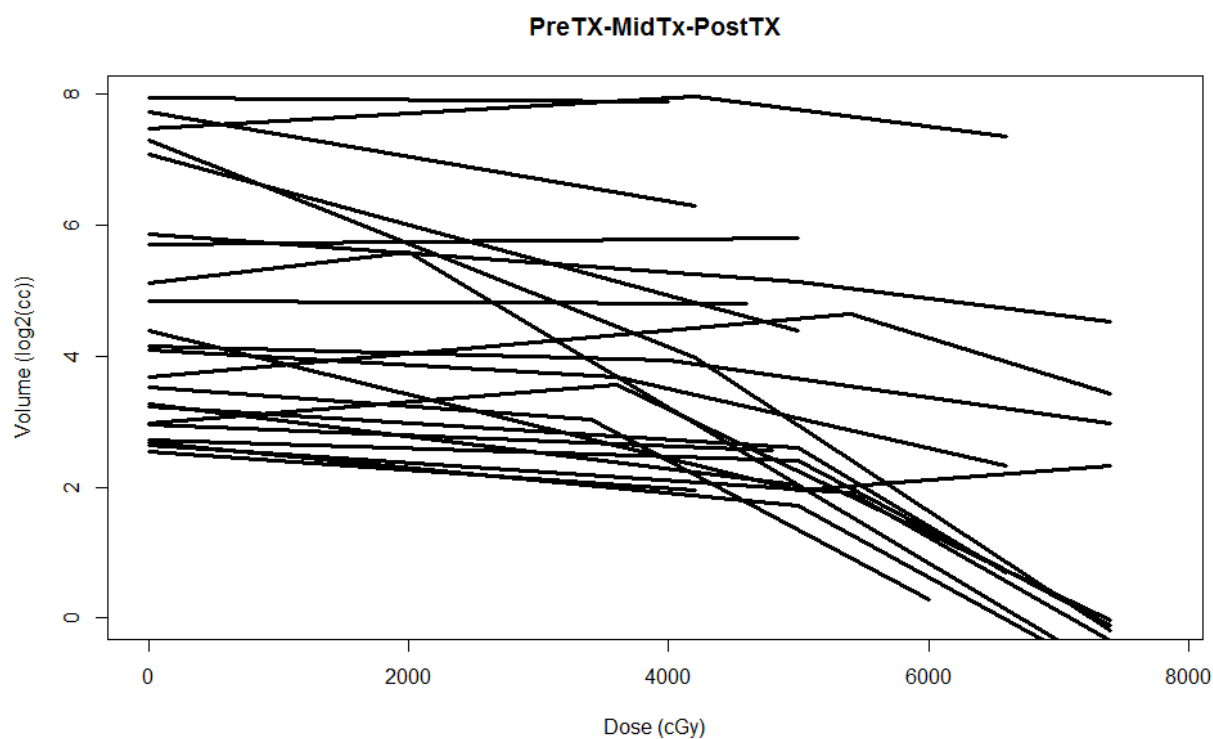


Figure 45. Primary volume changes during treatment

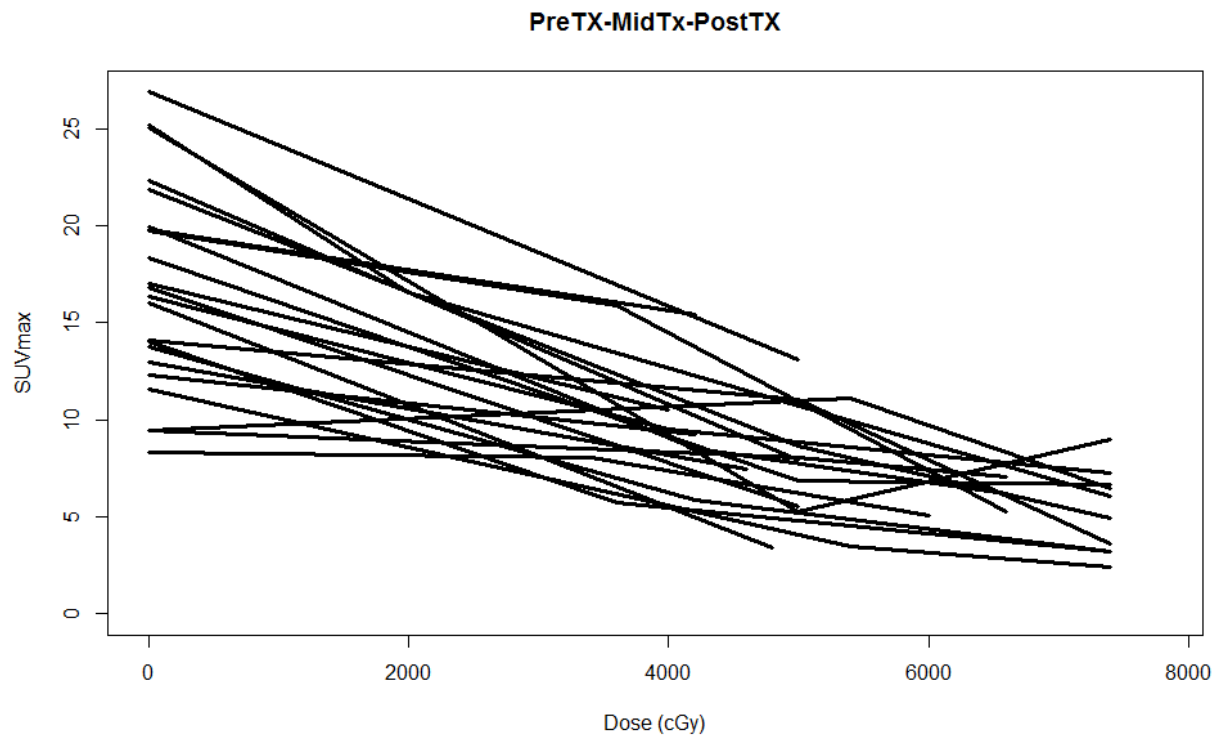


Figure 46. SUVmax changes during treatment

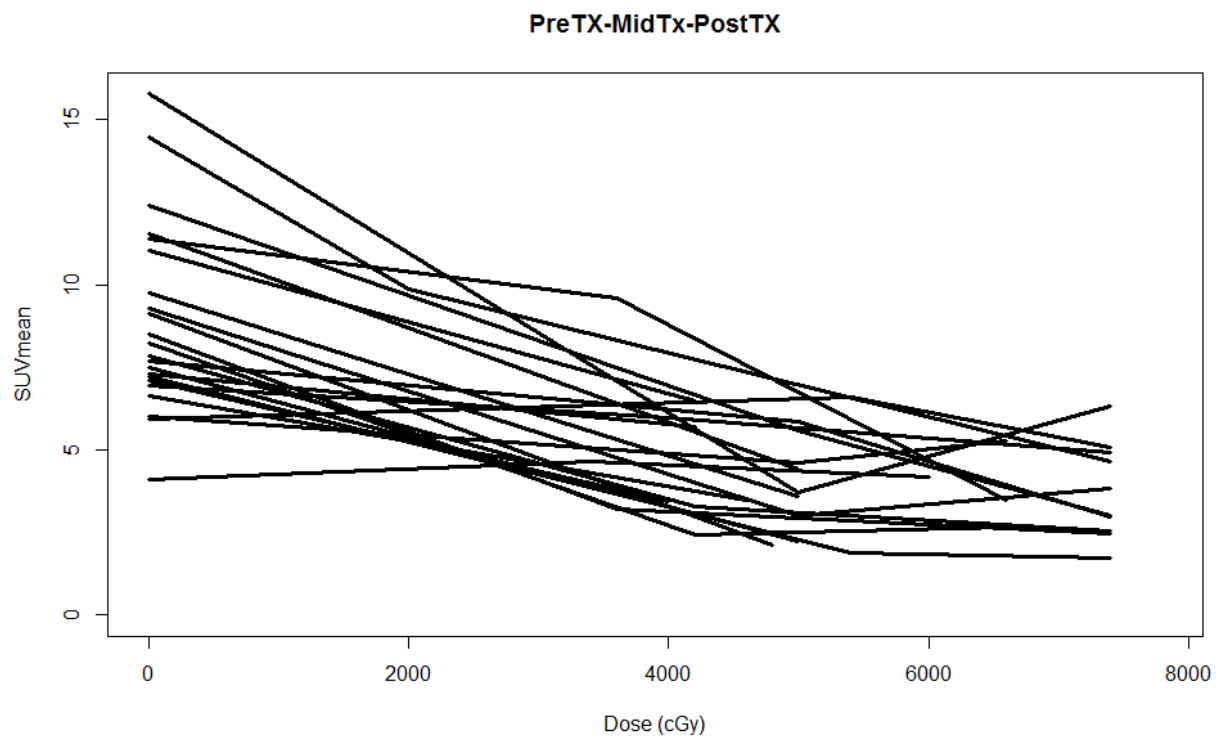


Figure 47. SUVmean changes during treatment

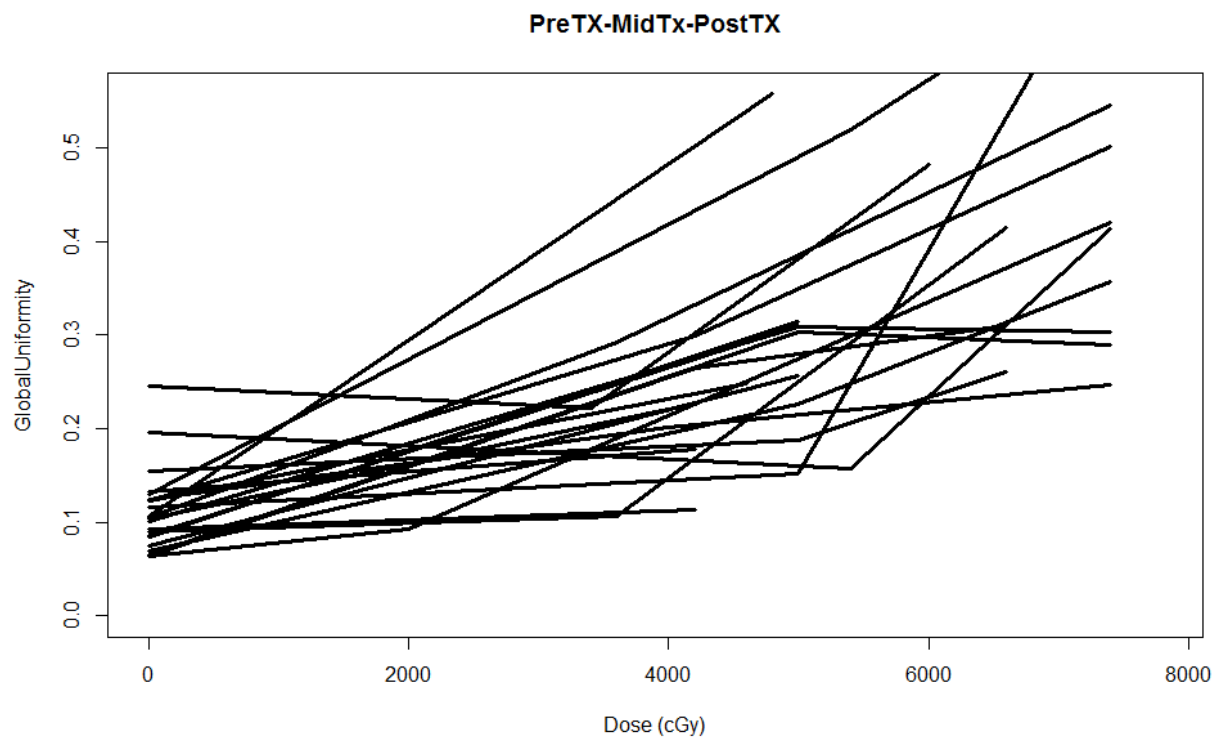


Figure 48. Uniformity changes during treatment

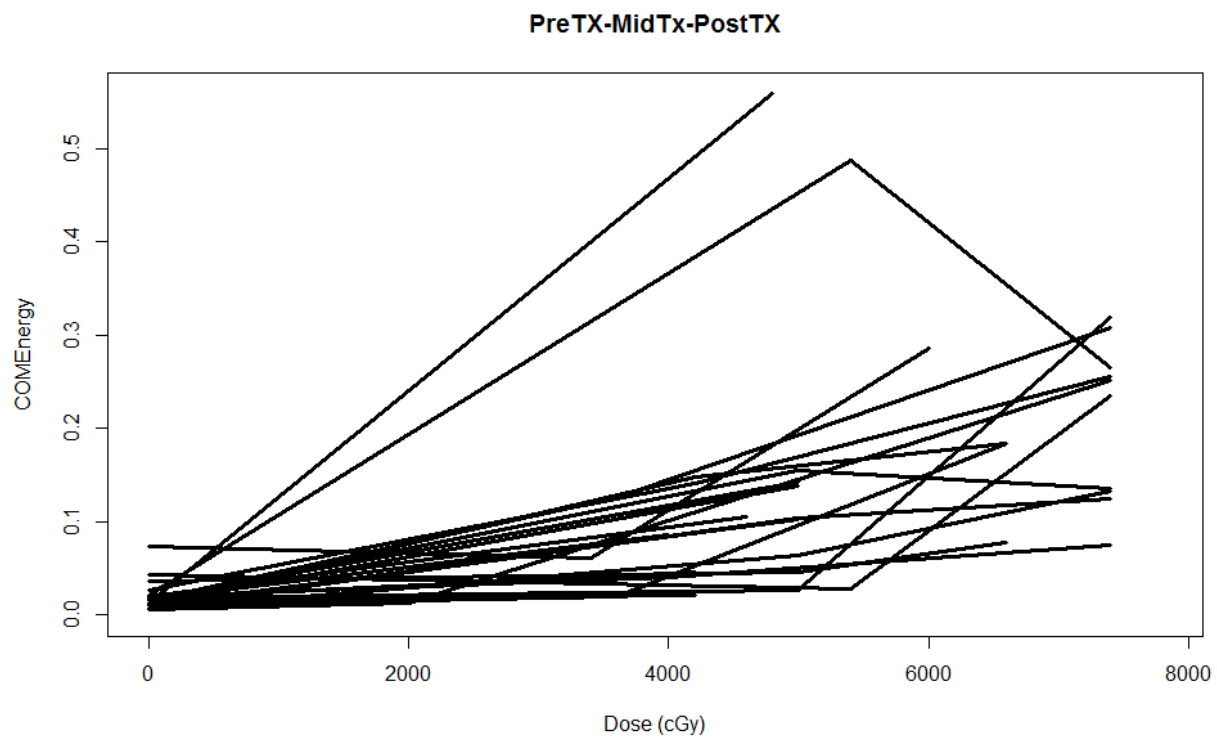


Figure 49. COM Energy changes during treatment

A summary of the primary tumor feature changes from pre to mid treatment and pre to post treatment are shown in Table 23.

Table 23. Summary of Changes in Features between Pre, Mid, and Post Treatment

| Feature | Average Pre- Treatment | Average Mid- Treatment | Average Post- Treatment | Average % Change Pre to Mid | Average % Change Pre to Post |
|----------------|---------------------------------------|---------------------------------------|--|--|---|
| Volume (cc) | 56.5 | 39.9 | 16.0 | -114 | -2077 |
| SUVmax | 16.9 | 9.0 | 5.5 | -117 | -223 |
| SUVmean | 8.9 | 4.6 | 3.8 | -127 | -161 |
| Uniformity | 0.11 | 0.25 | 0.43 | 46 | 70 |
| COM energy | 0.02 | 0.12 | 0.20 | 62 | 87 |

Table 23 shows that tumors became smaller, less FDG-avid (in both max and mean SUV), and more homogeneous (in terms of uniformity and COM energy) from pre-treatment to mid-treatment as well as pre-treatment to post treatment. The volume, SUVmax and SUVmean decreased on average by approximately 110-130 percent while uniformity and COM energy increased by approximately 50 percent from pre-treatment to post treatment. These trends increased from pre-treatment to post treatment with even larger reductions in primary volume, SUVmax, and SUVmean and increases in uniformity and COM energy. These analysis are inherently selecting tumor that do not have a complete response mid or post treatment as lesions would not be evaluable for quantitative analysis. Therefore, this type of assessment may be useful in characterizing lesions with a partial response, stable disease, or progressive disease. Quantifying changes in tumor uniformity during treatment alongside changes in volume and FDG-avidity may provide complimentary information that may be useful assessing response to therapy. Larger analyses would be needed to generate evidence relating changes in tumor uniformity to response or patient outcome.

Appendix G: Assessment of Volumetric Stability

Primary tumors from FDG-PET scans in cohort 5 were used to determine volumetric stability (i.e. reproducibility of QIFs across changes in tumor volume) via resampling (3.7.7 Analysis of PET Tumor Resampling (Cohort 5)). An example using histogram entropy is shown in Figure 50. Figure 50A-E show plots of the resampled entropy values versus the original entropy values with the corresponding CCC. Figure 16F displays a plot of the CCC values with respect to the approximate number of voxels. Table 24 contains the CCC values for a variety of features and their association with the number of voxels post resampling. It can be observed that the reproducibility decreases as the number of voxels post resampling is reduced. Certain features such as mean and standard deviation are stable regardless of the number of voxels while other features degrade by varying amounts as the number of voxels is reduced.

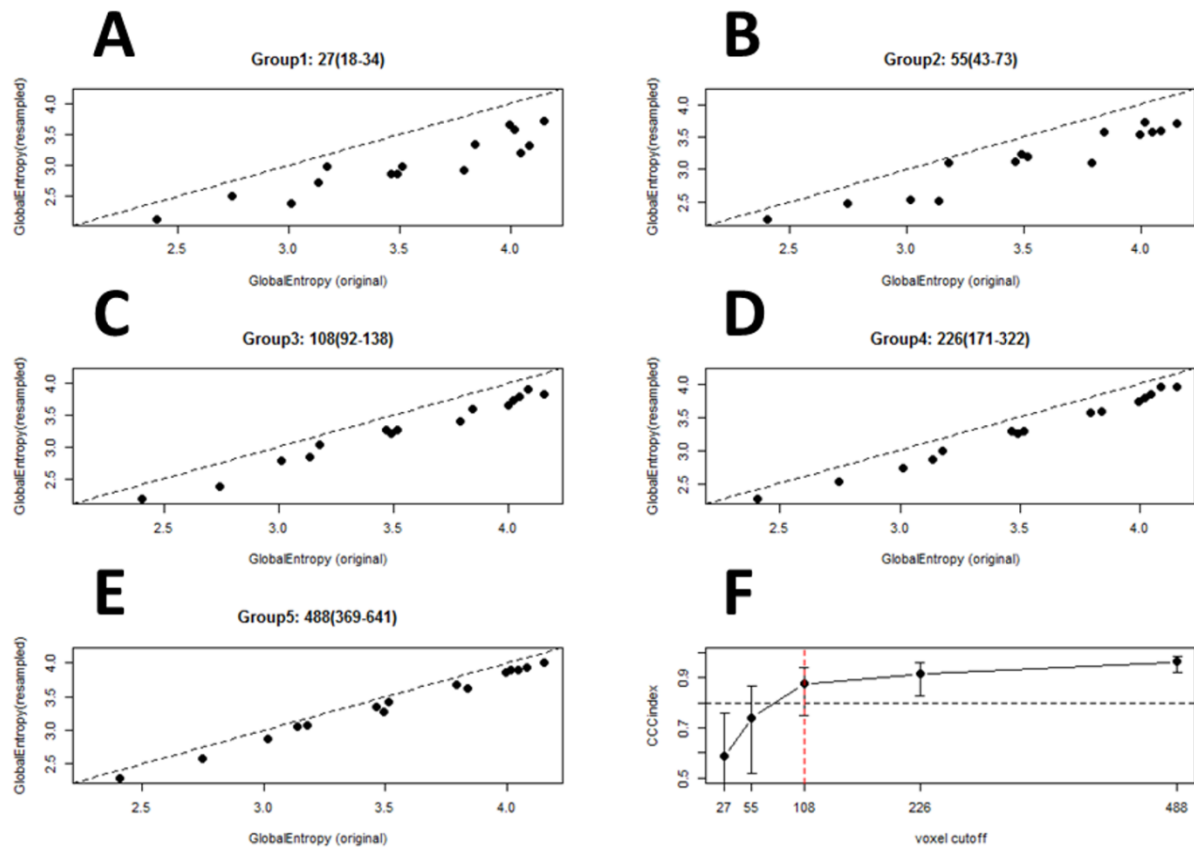


Figure 50. Plots of original versus resampled entropy values and associated CCC values

Table 24. CCC values of features with respect to the resampled number of voxels

| | Voxel Number Groups | | | | |
|-------------------|----------------------------|-----------|------------|------------|------------|
| QIFs | 27 | 55 | 108 | 226 | 488 |
| Volume | 0.92 | 0.94 | 0.94 | 0.98 | 0.99 |
| cumHistogram | 0.35 | 0.28 | 0.68 | 0.67 | 0.91 |
| TLG | 0.97 | 0.97 | 0.98 | 0.99 | 0.99 |
| GlobalEntropy | 0.59 | 0.74 | 0.87 | 0.92 | 0.96 |
| COV | 0.63 | 0.57 | 0.79 | 0.90 | 0.98 |
| Global Max | 0.73 | 0.73 | 0.87 | 0.88 | 0.93 |
| Global Mean | 0.98 | 0.98 | 0.97 | 0.98 | 0.97 |
| Global Std | 0.97 | 0.94 | 0.90 | 0.93 | 0.96 |
| Global Uniformity | 0.63 | 0.76 | 0.89 | 0.93 | 0.97 |
| Kurtosis | 0.43 | 0.49 | 0.52 | 0.80 | 0.91 |
| Skewness | 0.61 | 0.46 | 0.78 | 0.92 | 0.97 |
| COM Contrast | 0.48 | 0.59 | 0.76 | 0.85 | 0.94 |
| COM Correlation | 0.04 | 0.03 | 0.05 | 0.11 | 0.28 |
| COM Energy | 0.62 | 0.86 | 0.97 | 0.99 | 0.99 |
| COM Homogeneity | 0.56 | 0.73 | 0.73 | 0.86 | 0.93 |

A volume threshold below which quantitative features cannot be accurately/reproducibility measured is an area that requires further investigation. A couple publications have examined this issue but a consensus on size limitation has not been reached in order to ensure adequate sampling.^{53, 70} We found that features from tumors consisting of ~55 voxels (~5cc) yielded similar feature reproducibility to much larger lesions and that reproducibility suffers when resampling the same tumors to a smaller size. While improving reproducibility across tumor volumes is important it is not the only factor one should consider. Excluding larger and larger tumors due to non-ideal reproducibility also means reducing cohort sizes and the general applicability of quantitative techniques. For example, in our 225 patient cohort with FDG-PET scans 93% of patients have primary tumors with at least 27 voxels (lowest cutoff used in analysis). However, the percent of eligible patients decreases to 86%, 79%, 61%, and 46% when having a cutoff of at least 55, 108, 226, and 488 primary voxels, respectively.

VITA

David Vincent Fried was born in Durham, North Carolina on February 8, 1987 to Gary Fried and Susan Watts. He attended Green Hope High School in Cary, NC. In May 2009, David received his Bachelor of Science in Applied Sciences with a concentration in Biomedical Engineering from The University of North Carolina at Chapel Hill. Over the next three years, he worked as a research assistant in the Department of Radiation Oncology at UNC Hospitals. In August 2012, he entered the Medical Physics Ph.D. program at The University of Texas Graduate School of Biomedical Sciences with The University of Texas MD Anderson Cancer Center. He conducted his dissertation research under the guidance of Laurence Court, Ph.D.

Permanent address:

213 Draymore Way
Cary, NC 27519