2021

Computational Imaging Biomarkers For Precision Medicine: Characterizing Heterogeneity In Breast Cancer

Rhea Chitalia
University of Pennsylvania

Computational Imaging Biomarkers For Precision Medicine: Characterizing Heterogeneity In Breast Cancer

Abstract

In the United States, 1 in 8 women are diagnosed with breast cancer. Breast tumor heterogeneity is well-established, with intratumor heterogeneity manifesting spatially and temporally. Increased heterogeneity is associated with adverse clinical outcomes. Current critical disease treatment decisions are made on the basis of biomarkers acquired from tissue samples, largely under sampling the heterogeneous disease burden. In order to drive precision medicine treatment strategies for cancer, personalized biomarkers are needed to truly characterize intratumor heterogeneity. Medical imaging can provide anon-invasive, whole tumor sampling of disease burden at the time of diagnosis and allows for longitudinal monitoring of disease progression. The studies outlined in this thesis introduce analytical tools developed through computer vision, bioinformatics, and machine learning and use diagnostic and longitudinal clinical images of breast cancer to develop computational imaging biomarkers characterizing intratumor heterogeneity. Intrinsic imaging phenotypes of spatial heterogeneity, identified in dynamic contrast enhanced magnetic resonance imaging (DCE-MRI) images at the time of diagnosis, were identified and validated, demonstrating improved prognostic value over conventional histopathologic biomarkers when predicting 10-year recurrence free survival. Intrinsic phenotypes of longitudinal change in spatial heterogeneity in response to neoadjuvant treatment, identified in DCE-MRI were identified and leveraged as prognostic and predictive biomarkers, demonstrating augmented prognostic value when added to conventional histopathologic and personalized molecular biomarkers. To better characterize 4-D spatial and temporal heterogeneity, illuminated through dynamic positron emission tomography imaging, a novel 4-D segmentation algorithm was developed to identify spatially constrained, functionally discrete intratumor sub-regions. Quantifying the identified sub-regions through a novel imaging signature demonstrated the prognostic value of characterizing intratumor heterogeneity when predicting recurrence free survival, demonstrating prognostic improvement over established histopathologic biomarkers and conventional kinetic model derived parameters. Collectively, the studies in this thesis demonstrate the value of leveraging computational imaging biomarkers to characterize intratumor heterogeneity. Such biomarkers have the potential to be utilized towards precision medicine for cancer care.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Bioengineering

First Advisor

Despina Kontos

Keywords

Computer Vision, Imaging Biomarkers, Informatics, Radiology

COMPUTATIONAL IMAGING BIOMARKERS FOR PRECISION MEDICINE: CHARACTERIZING HETEROGENEITY IN BREAST CANCER

Rhea Devang Chitalia

A DISSERTATION

in

Bioengineering

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2022

Supervisor of Dissertation

_____

Despina Kontos
Matthew J. Wilson Associate Professor of Research Radiology II

Graduate Group Chairperson

_____

David Mankoff, Matthew J. Wilson Professor of Radiology

Dissertation Committee:

Joel Karp, Professor of Radiologic Physics in Radiology
Michael D. Feldman, Professor of Pathology and Laboratory Medicine

COMPUTATIONAL IMAGING BIOMARKERS FOR PRECISION MEDICINE:

CHARACTERIZING HETEROGENEITY IN BREAST CANCER

"To my mother, my grandmother, and the other 1 in 8 women diagnosed with breast cancer."

ACKNOWLEDGMENTS

I am enormously grateful to my advisor, Dr. Despina Kontos, for her support and guidance during my academic journey. In thanks to her mentorship, I have experienced immense personal and professional growth and have gained the confidence to become an independent scientist. Despina, you have always supported my academic and career goals, encouraged my intellectual curiosity, and during the most difficult research moments, reminded me to not to take it too seriously. I am confident that having you as a role model will help me be successful in the years to come.

I would like to thank the members of my thesis committee, Dr. David Mankoff, Dr. Joel Karp, and Dr. Michael Feldman for their invaluable feedback, guidance, and shared knowledge in all aspects of my thesis work and their direct mentorship on research projects. In particular, I would like to thank Dr. Mankoff for his continued excitement for and support of this thesis work, and for his advice for my career development.

To the past and present members of the Computational Biomarker Imaging Group (including those that were members of the Computational Breast Imaging Group), thank you. I am very fortunate and grateful to have you all as mentors, colleagues, and friends over the past many years. In particular, I would like to thank Michael Hsieh and Dr. Nariman Jahani for their invaluable assistance in all things image processing and for teaching me how to use the cluster. Thank you to Dr. Omid Haji Maghsoudi, Dr. Babak Haghighi, and Dr. Jose Marcio Luna for their friendship and career guidance. I am incredibly grateful to Dr. Aimilia Gastounioti, for her friendship and mentorship during my time at Penn. Much of this work would not have been possible without you. Thank you for always making time to discuss research projects, enjoy conference outings, and take coffee breaks. Thank you to Vivian Belenky for being a wonderful mentee and for contributing to studies in this thesis. A very special thank you to Lauren Pantalone for her friendship since day one of me joining CBIG. To my fellow PhD students, Apurva Singh,

Hannah Horng, and Alex Nguyen, my daily life in the lab was made richer through your friendship, intelligence, and inside jokes. I am grateful to the past and present members of CBICA who I have consider as mentors, colleagues and friends. In particular, thank you to Sarthak Pati for his friendship and for always helping me to debug my CaPTk issues.

This work would not be possible without the brilliant minds, both at Penn and afar, that I have been fortunate to have collaborated with. Thank you to Dr. Jenny Rowland, Dr. Elizabeth McDonald, and Dr. Austin Pantel for helping me segment tumors in the oldest of breast images. Thank you to Dr. Artemis Hatzigeorgiou, Marios Miliotis, and Dr. Spyros Tastsoglou at the University of Thessaly for your expertise in genomics and molecular profiling and for being great collaborators during a pandemic. Thank you to Dr. Mark Muzi and Dr. Lanell Peterson at the University of Washington for your invaluable insight into PET imaging. Special thanks to Dr. Varsha Viswanath for answering all of my PET questions, for designing all kinds of simulation, and for guiding me through the dissertation processes.

I have been very fortunate to have amazing mentors and teachers throughout my academic journey. I would like to thank my teachers at East Ridge High School in Clermont, Florida and my mentors at Duke University, in particular, Dr. Nimmi Ramanujam and Dr. Jenna Mueller. I am very grateful to Jenna for encouraging me to pursue a PhD and for showing me what a strong woman in STEM looks like.

I am extremely grateful to my many friends, including those not named here, that have been instrumental in my success and sanity during my PhD. Special thanks to Jonathan Galarraga for his support as we went through this final stretch together. Andrei Georgescu, Fabiana Zappala, Nick Perkons, Catherine Bautista, and Ryan Leiphart have been great friends since we started this program together; the dinners, game nights, and nights out have been bright spots during this PhD experience. To Sadhana Ravikumar, your friendship has been a true highlight of my time at Penn. Thank you for always listening to my experiment ideas and for accompanying

me to coffee breaks, Pottruck classes, and pandemic walks. To my roommate since day one, Divya Jain, thank you for being my family in Philadelphia. There's nothing like living in a tiny apartment during a pandemic lockdown to really test a friendship, and I'm proud that ours easily made it through. I am so grateful to have shared many years with Sadhana and Divya through this PhD and to have made life-long friendships. To Sruti Pisharody, Minali Nigam, Eshita Singh, Janice Yoon, Karina Guzman, and Kasey Cockefair, your friendship over the years is something I cherish the most.

I could not have done this without the love and support of my amazing family. My late grandfather, Hasmukhrai Meghani, taught me to always live and work with honesty. My grandmother, Manjula Meghani, has always showered me with pride and love from across an ocean. To my grandparents, Suresh and Hansa Chitalia, thank you for your boundless love and generosity. Your unconditional support and faith in everything I pursue has helped me to achieve this milestone.

Thank you to my sister-in-law, Haritha, for your constant optimism, pride, and guidance in everything I do. Most importantly, thank you for bringing Zion into my life. To my brother Nakul, thank you for always keeping me grounded and for being the person I have looked up to my entire life. I cannot articulate how important you both are to me.

Thank you to my fiancé, Ashwin Kommajesula. Your unfailing support, optimism, and ability to see the best in everything and everyone is something I learn from every day. The Megabus rides were worth it.

Lastly, and most importantly, the biggest thank you to my parents, Devang and Ushma Chitalia. Everything I have ever accomplished is both due to and for you.

ABSTRACT

COMPUTATIONAL IMAGING BIOMARKERS FOR PRECISION MEDICINE:

CHARACTERIZING HETEROGENEITY IN BREAST CANCER

Rhea Devang Chitalia

Despina Kontos

In the United States, 1 in 8 women are diagnosed with breast cancer. Breast tumor heterogeneity is well-established, with intratumor heterogeneity manifesting spatially and temporally. Increased heterogeneity is associated with adverse clinical outcomes. Current critical disease treatment decisions are made on the basis of biomarkers acquired from tissue samples, largely under sampling the heterogeneous disease burden. In order to drive precision medicine treatment strategies for cancer, personalized biomarkers are needed to truly characterize intratumor heterogeneity. Medical imaging can provide a non-invasive, whole tumor sampling of disease burden at the time of diagnosis and allows for longitudinal monitoring of disease progression. The studies outlined in this thesis introduce analytical tools developed through computer vision, bioinformatics, and machine learning and use diagnostic and longitudinal clinical images of breast cancer to develop computational imaging biomarkers characterizing intratumor heterogeneity. Intrinsic imaging phenotypes of spatial heterogeneity, identified in dynamic contrast enhanced magnetic resonance imaging (DCE-MRI) images at the time of diagnosis, were identified and validated, demonstrating improved prognostic value over conventional histopathologic biomarkers when predicting 10-year recurrence free survival. Intrinsic phenotypes of longitudinal change in spatial heterogeneity in response to neoadjuvant treatment, identified in DCE-MRI were identified and leveraged as prognostic and

predictive biomarkers, demonstrating augmented prognostic value when added to conventional histopathologic and personalized molecular biomarkers. To better characterize 4-D spatial and temporal heterogeneity, illuminated through dynamic positron emission tomography imaging, a novel 4-D segmentation algorithm was developed to identify spatially constrained, functionally discrete intratumor sub-regions. Quantifying the identified sub-regions through a novel imaging signature demonstrated the prognostic value of characterizing intratumor heterogeneity when predicting recurrence free survival, demonstrating prognostic improvement over established histopathologic biomarkers and conventional kinetic model derived parameters. Collectively, the studies in this thesis demonstrate the value of leveraging computational imaging biomarkers to characterize intratumor heterogeneity. Such biomarkers have the potential to be utilized towards precision medicine for cancer care.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 : Introduction to dissertation

Breast cancer is currently the most common malignancy and second leading cause of cancer-related death in women, with 1 in 8 women likely to develop breast cancer during their lifetime[1,2]. Cancer is a heterogeneous disease, with inter-tumor heterogeneity manifesting as variations between breast tumors across patients or as multifocal or multicentric tumors, and intratumor heterogeneity manifesting as variations within a single tumor[3,4]. Breast tumor heterogeneity is increasingly recognized as a key prognostic and predictive factor[4-8]. Genetic alterations and adaptive changes in response to dynamic, microenvironment specific stressors can result in sub-clonal populations with distinct stromal architecture and physiologic behavior[8,9]. This can continue to evolve throughout cancer progression and lead to functionally distinct sub-populations[8,10]. Increased intratumor heterogeneity is associated with adverse clinical outcomes[11,12] and tumor progression driven by aggressive subpopulations has been shown to be a mechanism for recurrence and therapy resistance[8].

Intratumor heterogeneity can manifest as spatial heterogeneity or temporal heterogeneity. Spatial heterogeneity can result in differences in blood flow, vessel permeability, and proliferation across the tumor volume and may be appreciated through pharmacokinetic dynamic variations. Temporal heterogeneity is due to longitudinal changes in the tumor that may arise in response to therapy leading to altered biomarker expression or to the acquired resistance by specific sub-clones during treatment[4,13].

Medical imaging is currently used during breast cancer screening[14,15], diagnosis[16,17], and treatment management[18,19]. Imaging modalities can allow for the non-invasive, whole tumor assessment of disease burden, including spatial and temporal heterogeneity, allowing for 4-D visualization[8]. Functional imaging modalities can yield parametric images illuminating various physiologic behaviors. In particular, dynamic contrast enhanced magnetic resonance imaging (DCE-MRI) is highly sensitive for primary lesion detection with the added ability to assess tumor vasculature[17,20]. Dynamic positron emission tomography (PET) imaging can also quantify specific facets of tumor molecular biology[7,19,21] and provide information beyond that of static imaging[22,23].

Advances in computer vision and machine learning have allowed for a transformation in the quantitative assessment of medical images[24-27]. Computational image analytic techniques and algorithms have been developed to leverage mineable, high dimensional information from medical images. The field of "radiomics" has introduced multi-parametric imaging features extracted with high throughput computational analysis from medical imaging data[28]. Such features, which may not be appreciable by the human eye, can provide complementary information to routine, qualitative clinical analysis. Medical imaging offers an opportunity to fully sample disease burden and characterize the extent of intratumor heterogeneity. Furthermore, leveraging computational analytic techniques can enable medical images to serve as novel non-invasive tumor assays.

Currently, critical disease treatment decisions for breast cancer patients are mainly made on the basis of markers typically acquired from core biopsy or surgical excision, which usually represents a small fraction of the tumor. Histologic assessment of such tissue samples determine common prognostic markers including tumor size, shape, grade, and nodal status, while estrogen receptor (ER) status, progesterone receptor (PR) status, human epidermal growth factor receptor 2 (HER2) status, and proliferative (Ki67) status are typically determined via immunohistochemistry[14,29,30]. Additionally, commercial prognostic molecular profiling tests such as MammaPrint (Agendia) and Oncotype DX Recurrence Score (Genomic Health) have been developed to measure mRNA and assess gene expression profiles respectively, in order to predict a tumor's risk of recurrence—however, they are expensive and are not always implemented in routine diagnosis[31]. Such conventional clinical prognostic biomarkers and prognostic molecular tests are largely dependent upon biopsy derived tissue, may be limited by spatial sampling, and therefore, fall short of characterizing a heterogeneous tumor volume. Furthermore, such biomarkers determined at the time of diagnosis may overlook temporal shifts due to dynamic breast cancer progression or response to therapy. As such, they may be inadequate for characterizing 4-D intratumor heterogeneity.

Qualitative descriptors of radiographic images are widely implemented in clinical settings, with observations of tumor morphology and "hot spots" utilized to classify malignant lesions[8,16,17]. Quantitative measurements describing static and kinetic lesion behavior in dynamic imaging are largely region-averaged prior to quantitative assessment, which assumes regions are functionally homogenous[32]. Beyond these clinical

uses of medical imaging, numerous radiomics studies have established relationships between conventional imaging biomarkers and clinically utilized prognostic markers from histology and genomics, as well as molecular subtypes[33]. While these studies have shown high prognostic capabilities, the derived imaging features are generally averaged over the entire tumor under the assumption that while heterogeneous, tumors are "well-mixed"[28,34]. Additionally, conventional radiomic features largely disregard 4-D differences in the tumor by utilizing spatial or temporal information, alone. Higher order radiomics features also often lack interpretability, thus limiting their clinical utility.

The goal of this work is to develop novel imaging biomarkers that characterize spatial and temporal 4-D functional tumor heterogeneity towards the development of personalized prognostic biomarkers and the non-invasive, *in vivo* characterization of functionally heterogeneous tumor biology. The first two studies discussed in this work focus on evaluating 4-D heterogeneity in DCE-MRI by assessing spatial differences across the tumor volume by combining whole-tumor assessment with dynamic imaging, as well as assessing the longitudinal, temporal changes in such 4-D heterogeneity. We then develop a novel method to identify functionally discrete, spatially constrained sub-regions in tumors and an imaging signature to characterize these sub-regions. We leverage the improved functional sampling afforded by dynamic PET imaging to evaluate 4-D spatiotemporal heterogeneity as a prognostic biomarker. Lastly, we explore extending the developed imaging biomarkers to pre-cancerous breast lesions and explore additional future directions. This work has been described further in the following chapters:

1. Explore spatial heterogeneity in DCE-MRI images of breast tumors at baseline by identifying imaging phenotypes. This study will utilize conventional radiomic features and evaluate the prognostic value of imaging phenotypes of tumor heterogeneity as a complement to conventional histopathologic biomarkers (Chapter 3).

2. Explore longitudinal changes in spatial heterogeneity in DCE-MRI images of breast tumors undergoing neoadjuvant chemotherapy through imaging phenotypes of change in intratumor heterogeneity. This study will evaluate the prognostic value of such longitudinal change as a complement to conventional histopathologic biomarkers and personalized molecular signatures (Chapter 4).

3. Develop a computational method to characterize 4-D pharmacokinetic spatial heterogeneity in dynamic PET imaging. This methodology leverages both spatial and temporal information and is evaluated using simulated dynamic PET image phantoms (Chapter 5).

4. Develop imaging features summarizing 4-D spatiotemporal heterogeneity and evaluate these features as a prognostic biomarker. This study is evaluated in dynamic PET images of women diagnosed with locally advanced breast cancer (Chapter 6).

5. Explore the feasibility of developing an unsupervised method to summarize 4-D pharmacokinetic spatial heterogeneity. This study will evaluate the performance of this methodology on dynamic PET image phantoms simulating tumor heterogeneity (Chapter 7).

6. Explore extended applications of quantifying lesion heterogeneity and examine additional future directions of this thesis work (Chapter 8).

Chapter 2 : Background

Subsections of this chapter have been adapted from the following:

Chitalia, R.D. and Kontos, D., 2019. Role of texture analysis in breast MRI as a cancer biomarker: A review. *Journal of Magnetic Resonance Imaging*, *49*(4), pp.927-938.

Liao, G.J., Henze Bancroft, L.C., Strigel, R.M., **Chitalia, R.D.**, Kontos, D., Moy, L., Partridge, S.C. and Rahbar, H., 2020. Background parenchymal enhancement on breast MRI: a comprehensive review. *Journal of Magnetic Resonance Imaging*, *51*(1), pp.43-61.

2.1. Introduction

Breast cancer is a malignancy originating through ductal hyper proliferation leading to *in situ*, invasive carcinoma, and potential metastatic disease[5]. Breast cancer displays the hallmarks of cancer shared across organ sites, defined as sustained proliferative signaling, evasion of growth suppressors, resistance of cell death, enablement of replicative immortality, increased angiogenesis, and invasion and metastasis[35]. Breast cancer has been established as a heterogeneous disease, and such heterogeneity can manifest both between tumors (inter-tumor heterogeneity) as well as within a single tumor (intratumor heterogeneity).

2.2. Tumor heterogeneity

Inter-tumor heterogeneity in breast cancer has been illuminated through the variations in seen across breast tumors[36], and has resulted in discrete molecular subtypes. Breast cancer molecular classification has been largely dictated by the presence or absence of the immunohistochemistry derived expressions of three main receptors: estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor

receptor 2+ (Her2). The increased presence of ER and PR suggests an increased dependency of the tumor on estrogen and progesterone steroids for growth, while Her2 tumors have an overexpression of the ERBB2 oncogene. As such, luminal breast tumors are positive for ER and PR, or hormone receptor (HR) status positive, Her2 enriched breast tumors are HER2+, and basal like breast tumors lack HR and HER2 and are considered to be triple-negative breast cancers (TNBC)[37,38]. Incidence rates, metastatic potential, disease progression, and patient outcome can all be stratified along these molecular subtypes[36,38].

Intratumor heterogeneity refers to the differential properties displayed across cancer cells, within a tumor[4]. Such properties and traits can be related to tumorigenesis including angiogenic, invasive, and metastatic potential[4]. Given the current understanding of breast tumor formation, intratumor heterogeneity may stem from the branched evolution of clonal cells, leading to the generation of multiple sub-clonal populations[39,40]. These sub-clonal populations may emerge during the malignancy conversion of non-malignancy cells, due to genetic and epigenetic alterations, or in response to microenvironment specific stressors as an adaptive response[4,6]. Intratumor heterogeneity remains dynamic through tumor progression as tumor properties may change in response to treatment due to de-novo or acquired resistance. Additionally, the acquired resistance by aggressive sub-clonal populations may drive recurrence or metastases[40]. As such, increased heterogeneity is associated with adverse clinical outcomes, with higher degrees of intratumor heterogeneity suggesting inferior responses to targeted treatments and anticancer therapies[39,41-45]. Various studies have demonstrated this heterogeneity through

single cell and whole-genome sequencing, RNA and microRNA sequencing, histopathology, and imaging modalities[8,46-50].

Intratumor heterogeneity can be spatial—resulting in variations in tumor properties across the volumetric disease burden, or temporal—resulting in variations in time seen through dynamic or longitudinal changes in tumor properties[4-6,39]. Spatial heterogeneity can often be appreciated through differences in tumor blood flow[9], metabolism[51], or proliferation[52] across the tumor volume. Temporal heterogeneity may include the dynamic heterogeneity demonstrated by tumors over shorter periods of time in response to a pharmacokinetic agent, or longitudinal changes in tumor behavior following natural disease progression or therapy.

2.3. Clinical management of breast cancer

Breast cancer is most commonly detected through screening mammograms or through a palpable breast mass. Of breast lesions diagnosed in the United States, approximately 62% are confined to the breast, 31% have spread to regional lymph nodes, and 6% are metastatic at the time of diagnosis[53]. Breast cancer is most commonly diagnosed through histopathologic assessments from fine-needle aspiration, core biopsy, or surgical excision, and molecular biomarkers evaluated during this assessment often define treatment paradigms. For non-metastatic breast tumors, the goal of tumor eradication usually involves surgical resection and removal of axillary lymph nodes. In addition to this, therapy can include neoadjuvant (prior to surgery) or adjuvant (post-surgery) chemotherapy, radiation, and targeted therapies to prevent metastases or recurrence. Current conventional targeted therapies are specific to the molecular

receptors identified during histopathologic assessment. Endocrine therapy is suggested for HR+ tumors, such as anti-estrogen or aromatase inhibitory treatment. Antibody therapy is suggested for HER2+ tumors; Herceptin, a monoclonal antibody specifically targeted for HER2 receptor was developed to block intracellular signaling pathways, promote apoptosis, and promote cell proliferation[54]. A greater understanding of breast tumor biology has led to the identification of numerous molecular targets and targeted therapeutics, including inhibitors of growth factor tyrosine kinase, inhibitors of signaling pathways, and anti-angiogenic strategies[55,56].

Medical imaging allows for the non-invasive, whole tumor sampling of disease burden, with the ability to longitudinally monitor response to treatment[16,17]. Magnetic Resonance Imaging (MRI) is highly sensitive for primary lesion detection, particularly for high risk women[17]. Specific MRI sequences such as diffusion weighted (DW) MRI and dynamic contrast enhanced (DCE) MRI can provide further insight into tissue architecture and vascularization within and around the tumor[20,57]. In particular, DCE-MRI involves the injection of a contrast agent with small molecular weight, followed by sequential imaging over the first ten minutes to assess extravasation of the agent from vascular to interstitial spaces. Such imaging can illuminate tumor perfusion and microvascular permeability[58]. Current clinical analysis of MR images is largely qualitative, using DCE-MRI to identify tumor regions with contrast uptake or to monitor morphologic appearance[59]. DCE-MRI offers advantages over other imaging methods that can measure vasculature, including increased signal to noise ratio and spatial resolution, and does not involve radiation exposure[60].

Positron emission tomography (PET) imaging is most commonly clinically performed for breast staging of stage IIIB/C or IV disease[21]. PET imaging can also be used for staging patients with metastasis, or monitoring response to treatment[61]. PET imaging works to detect radiotracer—a radioisotope paired with a molecule—distribution in the body after injection. The most commonly used radiotracer in cancer imaging is ^{18}F-fluorodeoxyglucose, which pairs a glucose analog to the ^{18}F radioisotope. Use of this radiotracer can provide insight into tumor metabolism as tumors have been observed to utilize large amounts of glucose for energy generation through the anaerobic process of glycolysis, known as the Warburg effect[62]. This is in contrast to normal tissue which utilizes oxidative metabolism. Clinically, the most common quantitative parameter derived from PET imaging is the standardized uptake value (SUV), or the estimated amount of radiotracer per volume of radiotracer dose, scaled by patient weight. While static imaging is most commonly utilized in the clinic, recent advances in PET applications have demonstrated the added value of dynamic imaging of breast tumors[22].

While other imaging modalities such as X-Ray and Ultrasound are implemented in breast cancer clinical imaging[21], they are used largely for screening purposes and are not used as often as MRI for staging and treatment response monitoring due to radiation exposure or poor spatial resolution. Computed tomography (CT) imaging is often paired with PET imaging to provide a structural image in addition to the functional image.

2.4. Precision medicine and radiomic analysis

Precision medicine aims to tailor disease prognosis and treatment based on specific genotypic and phenotypic characteristics of an individual. Recent advances in

medical image analysis have highlighted the implementation of computer vision principles and analytic techniques used to quantify and describe medical images towards precision disease prognosis and prediction[24-27,33].

Image texture has been previously defined as repeating patterns of local variations in gray-level intensities[25,63]. Texture analysis has most broadly been used to characterize the spatial distribution of gray-level intensities within an image, capturing image patterns usually unrecognizable or undistinguishable to the human eye. The original utilization of texture analysis can trace back to computer vision applications for surface inspection and orientation, image and object classification, and shape determination, while current applications extend even beyond medical image analysis [63-65]. Within the scope of breast imaging, texture analysis has emerged as a quantitative, surrogate measure for breast parenchymal pattern when applied to images taken during mammographic and tomographic screenings serving to augment conventional measures of breast percent density in breast cancer risk assessment[66].

Many studies aiming to analyze imaging presentations of breast lesions for diagnostic, prognostic, and treatment applications have expanded the number and type of features extracted to include morphology, texture, and pharmacokinetic features, allowing for a thorough and quantitative characterization of all tumor properties. This has developed into the new field of radiomics, broadly defined as the extraction of high throughput quantitative features from images obtained from medical imaging modalities [28,67,68]. For the analysis of breast lesions, texture features are often extracted from a region of interest (ROI) selected within a segmented lesion, or from the whole lesion itself.

Additionally, recent studies have shown clinical associations of texture within the peri-tumor region as well, emphasizing the importance of the tumor microenvironment[69,70]. The most commonly used radiomics features can be stratified by the statistical order of the voxel information encoded within the image. Specifically, first order radiomic features include common statistical measures derived from a gray-level histogram, such as mean, median, and skewness. Second order radiomic features are often derived from the co-occurrence matrix, as determined by Haralick et al.[71], and the run-length matrix[72], while higher order radiomic features encode structural and frequency based texture information (Figure 2.1.).

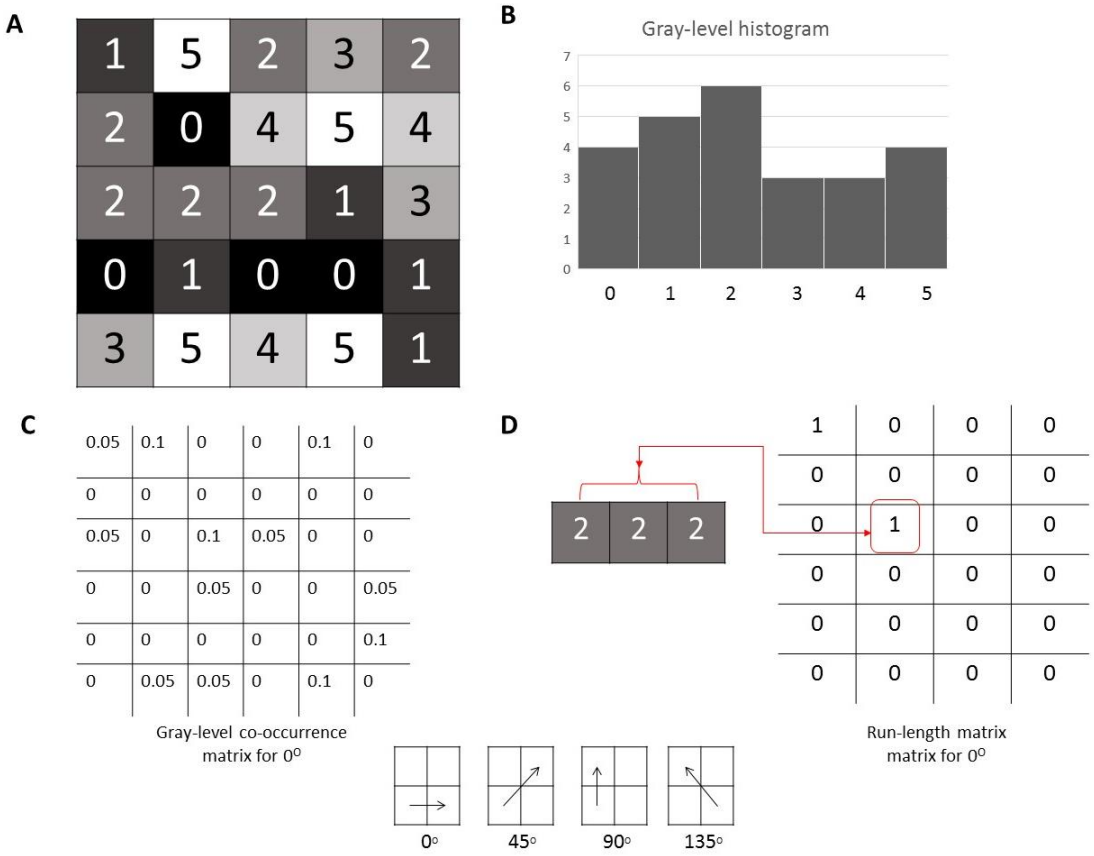


Figure 2.1. Visual example of radiomic features. Representative 5 x 5 pixel image with 6 possible gray-levels (0-5) (A). Gray-level histogram generated from representative image (B). Gray-level co-occurrence matrix generated for 0˚. The co-occurrence matrix encodes the frequency that two pixels are located a specific distance (1 pixel) away from one another (C). Run-length matrix generated for 0˚. Run-length matrix encodes the coarseness of an image in a specified linear direction (D). This figure was adapted from Chitalia et al. Journal of Magnetic Resonance Imaging, 2019.

2.5. First order: Gray-level histogram features

A gray-level histogram can be generated by calculating a frequency count of the number of voxels of each gray-level intensity value, where the total number of discretized gray-levels is often a user-selected parameter. From the resulting histogram, first order statistical features may be derived, including the mean, median, and variance. Higher moment features can also be extracted from the histogram including skewness, the measure of the histogram distribution symmetry, and kurtosis, a measure of the histogram

distribution shape. While many descriptors can be extracted from the histogram, they often provide cursory insight into the underlying texture, do not account for gray-level intensity spatial relationships within an image, and are dependent upon user-selected parameters. (Appendix Table A1)

2.6. Second order: Gray-level co-occurrence matrix and run length features

Gray-level co-occurrence matrix (GLCM) features are among the most commonly extracted radiomics features for MR imaging quantification. A gray-level co-occurrence matrix encodes the frequency that two voxels of specific gray-level intensities are positioned a specified distance away from each other in a specified image orientation [71]. GLCM features are most often quantified in the four diagonal image orientations of 0°, 45°, 90°, and 135°. Second order features can then be extracted from the co-occurrence matrix. Examples of such features include contrast, a descriptor of the intensity contrast between a pixel and its neighbor as determined by the distance parameter, correlation, a descriptor of the linear gray-level dependence, and homogeneity, a descriptor of the closeness of distribution in the co-occurrence matrix to the matrix diagonal. Other second order features such as energy, a descriptor of the certainty of gray-level co-occurrence respectively, and cluster shade, a descriptor of asymmetry in gray-level values, can also be extracted. Entropy, or the randomness of the GLCM, is another commonly extracted feature, often indicating image heterogeneity (Appendix Table A2). Run-length features measure the coarseness of an image in specified linear directions [72-74] (Appendix Table A3).

2.7. Higher order: Structural and transformation-based features

Structural features capture the intensity variations between central voxels and surrounding neighboring voxels[75] (Appendix Table A.4). Extending radiomic analysis beyond the spatial orientation of gray-level intensity values, transformation based radiomics features capture texture information encoded in a different space, such as the frequency space. Transformation based methods include the Fourier transform, Gabor transform, and Wavelet transform[76-78]. The wavelet transform, in particular, is commonly used due to its ability to capture MR images' frequency content both at varying image scales and multiple specified directions.

2.8. Morphology features

Morphologic radiomic features aim to quantify the lesion's physical characteristics, either in 2-D or 3-D space. Such features may include volume, area, and perimeter, as well as eccentricity, spiculation, or flatness. Variations of morphologic features may account the entire lesion volume or consider only enhancing portions of the tumor.

2.9. Radiomic analysis in breast lesions- MRI

Benefitted by the whole-tumor sampling and visualization of tumor vasculature afforded by MRI, one of the largest aims in radiomics is to accurately characterize and quantify intratumor heterogeneity [4,7,8]. As such, radiomic information extracted from medical images can have high clinical relevance [79-87]. As compared to a global or qualitative report of breast tumor appearance, radiomic analysis can provide a refined, local description of tumor complexity, heterogeneity, and kinetic behavior as seen in

medical imaging. This quantitative characterization can have specific applications towards the diagnosis, prognosis, and treatment of breast cancer. As MRI is the most commonly utilized imaging modality for breast cancer diagnosis and treatment management[16,88], most advances in breast tumor radiomics have been implemented in MR imaging[34].

2.9.1. Radiomics applications in breast computer aided diagnosis

Computer aided diagnosis (CAD) of breast tissue was one of the earliest applications of radiomic analysis in the breast [89]. Gibbs et al. [90] was one of the first to apply radiomic analysis towards classifying breast lesions as benign or malignant. The authors reported using 2D DCE-MR images from a cohort of 79 women, of which 45 were diagnosed with breast cancer. Within each lesion ROI, a co-occurrence matrix was determined for adjoining pixels in 0°, 45°, 90°, and 135° directions. Radiomic features of variance, sum entropy, and entropy were concluded to be the most significant when discriminating between benign and malignant lesions, suggesting that features quantifying image texture could be a useful tool in lesion delineation. Subsequent studies have followed this preliminary, yet promising, conclusion by utilizing more complex measures of image texture to diagnose breast lesions as benign or malignant [91-97]. This idea was extended by Chen et al. [96], who extracted 3D GLCM features from a 3D breast lesion segmentation. The 3D GLCM features yielded a higher diagnostic accuracy than 2D GLCM features extracted from a 2D ROI, when distinguishing between malignant and benign breast lesions, showing an advantage for 3D breast lesion characterization.

2.9.2. Beyond CAD: Radiomic analysis for histopathologic and molecular subtype classification

The promising conclusions of MRI radiomic analysis in breast cancer diagnosis suggests an architectural difference between the imaging presentation of benign and malignant lesions that can be quantified using radiomic features. Recent studies have begun extending this idea, hypothesizing that underlying tumor biological differences can be imaged using MRI and characterized using radiomic analysis. Consequently, many groups have employed radiomic analysis to distinguish between the heterogeneous histopathologic [70,93,95,98] and molecular [99-102] subtypes of breast cancer, with a larger goal of utilizing image texture features to provide a personalized diagnosis.

In an attempt to distinguish between Invasive Lobular Carcinoma (ILC) and Invasive Ductal Carcinoma (IDC), Holli et al. [95] extracted a total of 277 histogram, GLCM, run-length, and wavelet features. The authors found that of these, only GLCM related features characterizing lesion complexity and randomness were significantly different between ILC and IDC lesions. Similar conclusions were reported by Waugh et al.[98], who found that entropy, a measure of pixel distribution randomness, was significantly different between lobular and ductal lesions, suggesting a difference in underlying growth patterns and tumor heterogeneity.

Increasing the scope of radiomic analysis to include the surrounding microenvironment in additions to the breast lesion, Wang et al.[70] investigated the role of kinetic contrast uptake texture in differentiating between histopathologic subtypes of breast cancer. Radiomic features were extracted from pharmacokinetic parametric maps

generated from DCE-MR images of the tumor and surrounding parenchyma. Adding radiomic features characterizing heterogeneous uptake in the breast parenchyma to a model containing lesion texture features allowed for the identification of triple negative breast cancers (TNBC). The study concluded that the characterization of heterogeneity, both within the lesion as well as the surrounding parenchyma, could provide noninvasive insight towards heterogeneous tumor behavior associated with more aggressive subtypes. These results were similar to those found in previous studies [103,104], indicating the clinical value of lesion and peri-tumoral contrast uptake quantification. Radiomic features quantifying lesion heterogeneity have also shown to aid in delineating between molecular subtypes of breast cancer. Studies have shown a radiomic difference between the MR presentation of luminal A and luminal B subtypes, with luminal B lesions having a more quantifiably heterogeneous appearance[99,102].

2.9.3. Radiomic analysis for breast cancer prognosis and therapy response prediction

Recent studies have shown promising conclusions when exploring the relationship between breast lesion appearance and risk of recurrence, and the value radiomics as a non-invasive prognostic biomarker [93,105-109]. Kim et al. performed a retrospective analysis of 203 women diagnosed with invasive breast cancer, extracting histogram uniformity and entropy features from both T2-weighted MR images and T1-weighted DCE-MRI. Univariate and multivariate associations between these texture features and disease-free survival determined that increased tumor heterogeneity in T2-weighted MRI could be used to stratify patients more at risk for recurrence. This study suggested that tumor heterogeneity, as quantified by lesion texture, could be used in MR

imaging as an independent prognostic marker. Similar conclusions were drawn by Park et al. [107] who generated a multivariate feature vector based on morphologic, histogram texture, and GLCM texture features, from which specifically GLCM cluster tendency, GLCM variance, and GLCM sum variance were selected for in a model stratifying patients at risk for recurrence. Mahroongy et al. [108,109] extracted wavelet features from within tumor sub-regions partitioned by pharmacokinetic behavior and concluded that the spatial frequency texture pattern captured using wavelets within the heterogeneous sub-regions could serve as a strong prognostic biomarker for predicting risk of tumor recurrence (AUC= 0.88).

The potential for radiomic analysis in breast cancer treatment has been demonstrated in recent studies, showing the potential for MRI-extracted radiomic features to serve as non-invasive predictive biomarkers [69,73,104,105,110-120]. In order to predict response to treatment, some studies utilize first-order statistical measures extracted from the tumor ROI. Specifically, Johansen et al. [119] calculated three first-order statistical features of mean, standard deviation, and prediction from a relative signal intensity histogram generated from pre-chemotherapy DCE-MRI scans. Of these, skewness and kurtosis were found to be strongly correlated with complete response to therapy. Similarly, Padhani et al. [120] conducted a retrospective study of 25 women diagnosed with primary invasive cancer, imaged using DCE-MRI before and after the first cycle of treatment. Leveraging contrast enhancement, the authors generated a histogram from a pharmacokinetic parametric map of the full lesion ROI, and concluded

that responsive patients displayed a decrease in pharmacokinetic range, and proposed that this could be attributed to a decrease in heterogeneity after the first cycle of treatment.

While histogram texture can provide useful information regarding the distribution of gray-level intensity values, it is limited when capturing spatial heterogeneity within a lesion as it largely ignores the spatial relationships between voxels. Studies extracting higher-order radiomic features can further quantify the relation between tumor heterogeneity and response to therapy. To this end, Teruel et al. extracted second-order statistical GLCM features from pharmacokinetic maps generated from DCE-MRI images of women diagnosed with locally advanced breast cancer. Eight GLCM features were found to significantly differ between responders and non-responders, and GLCM sum variance was able to predict response to treatment with an AUC of 0.77. Similar conclusions were drawn by Thibault et al. [111], who expanded this 2D analysis by extracting 3D GLCM features from DCE-MR pharmacokinetic parametric maps in order to predict response to neo-adjuvant chemotherapy (NAC). The 3D GLCM features were particularly significant in identifying early responders to NAC, with results showing non-responders having higher microvascular heterogeneity. In a retrospective study of 36 women who underwent NACT, the change in tumor heterogeneity between pre-treatment and mid-treatment, as calculated by entropy and uniformity changes, was predictive of pCR with an AUC of 0.84 [113]. Comparing this performance to change in tumor size (AUC= 0.66) demonstrates a greater sensitivity for lesion texture in characterizing early response to pCR.

Current clinical predictions for achieving pathologic complete response (pCR) are based on tumor histopathologic characteristics. As intratumor heterogeneity is associated with adverse clinical outcomes, the limited tissue taken during biopsy may be inadequate for a whole tumor based prediction [7,8]. Michoux et al. [114] performed a retrospective analysis on the DCE-MR scans of sixty-nine patients diagnosed with IDC, undergoing NAC. For each woman, texture, kinetic, and morphology-based features were extracted from within the pre-treatment lesion ROI. The authors concluded that only four parameters—three features (GLCM inverse difference moment, grey-level nonuniformity, and long run high gray level emphasis), and the wash-in slope kinetic feature —were found to classify non-responders with 84% sensitivity. Of particular note, clinically utilized histopathologic predictive biomarkers such as estrogen receptor (ER) status, progesterone receptor (PR) status, Ki67 status, and human epidermal growth factor receptor 2 (HER2), along with tumor grade, were not significant when classifying early response, further highlighting the utility of radiomic analysis. Similarly, Golden et al. [112] extracted GLCM based features from pharmacokinetic parametric maps generated from DCE-MR images of women diagnosed with TNBC, in order to predict pCR, residual lymph node metastases, and residual tumor with lymph node metastases. The predictive performance of GLCM texture features was compared to 'patterns of response', a qualitative description of lesion appearance before and after chemotherapy, as determined by a radiologist. The GLCM features extracted from pre-chemotherapy MR images could predict pCR and residual lymph node metastasis with a reported AUC of 0.68 when classified in a logistic regression model. In contrast, the radiologist determined 'patterns of response' did not predict any of three outcome measures.

The conclusions drawn from studies implementing radiomic analysis for breast cancer diagnosis, prognosis, and treatment suggest that radiomic features demonstrating increased lesion heterogeneity are associated with aggressive growth, unfavorable prognosis, and poor treatment outcomes[69,105,109,115,121-123]. In addition, they propose a method for non-invasively quantifying the underlying biology of tumor sub-regions driving recurrence, response, and resistance to therapy (Figure 2.2).

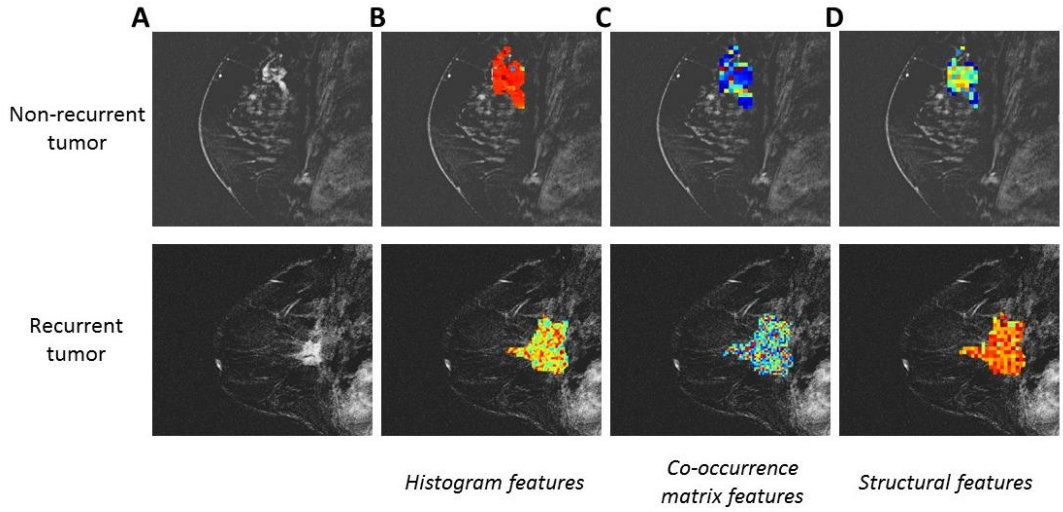


Figure 2.2 Representative images of a non-recurrent and recurrent breast tumor. (A). Examples of texture features maps showing distributions of histogram texture features (B), co-occurrence matrix texture features (C), and structural texture features (D). This figure was adapted from Chitalia et al. Journal of Magnetic Resonance Imaging, 2019.

2.10. Novel applications of breast radiomics

While most advances in radiomic analyses of breast cancer have been made utilizing MRI images, there are few studies investigating radiomic analysis of molecular imaging modalities such as PET[124]. For the diagnosis of breast lesion as benign or

malignant, Vogl et al. investigated MRI and PET radiomic features. The author found that only MRI-based texture features were predictive in lesion classification. Conversely, Ou et al. found that SUV and radiomic features from FDG PET could stratify breast cancer from breast lymphoma[125].

As PET imaging leverages biologically targeted radiotracers, radiomic analysis of PET imaging has been used for the biological characterization of breast tumors. Lemarignier et al. found that radiomic features of image texture extracted from FDG PET imaging were significantly associated with histological subtypes of breast cancer[126]. However, other studies looking at associations between radiomic features from FDG PET and histological subtypes found no significant associations[127-130]. Studies have also examined the correlation between PET radiomic features and molecular subtypes of breast cancer[128,130-132], finding associations between features and receptor status[131,133], and proliferative status[129].

PET radiomic analysis has also been explored for treatment response prediction. Numerous studies have found associations between PET-derived radiomic features and pathological complete response to NCT[129,132,134-136]. Lastly, PET radiomic analysis has been utilized to predict patient outcome including progression-free survival (PFS)[136], disease-free survival (DFS)[132], overall survival (OS)[127], and recurrence-free survival (RFS)[128].

While DCE imaging and PET imaging are not utilized as often as other modalities for breast cancer screening, recent studies have suggested associations between radiomic

analysis and characterizations of the whole breast, such as background parenchymal enhancement (BPE) from DCE-MRI[137-139]. With respect to personalized breast cancer diagnosis and treatment response monitoring, studies have demonstrated that increasing the scope of radiomic analysis to include surrounding BPE within the peri-tumoral region has allowed for improved diagnostic and predictive performance[69,70,140,141]. Mazurowski et al.[140] demonstrated that radiomic features characterizing the texture and pharmacokinetic behavior of breast lesions and the surrounding parenchyma were highly associated with the luminal B subtype, suggesting that quantitative characterization of BPE could provide personalized diagnoses. Similar conclusions were reported by Wang et. al [70], where features characterizing texture were extracted from pharmacokinetic maps generated from DCE images of the tumor and surrounding parenchyma. Augmenting a classification model of lesion radiomic features with BPE radiomic features improved performance for the identification of triple negative breast cancers (TNBC). Radiomic analysis has the potential to provide quantitative insight into parenchymal enhancement patterns to further understand the role of BPE in personalized clinical decision making.

Leveraging high-throughput genomic, molecular, and sequencing data alongside imaging data may provide complementary insight into patient-specific diseases for the improvement of precision cancer care. This has developed into the field of "radiogenomics", which aims to find associations between imaging and genomic streams of information for the combined prediction of patient outcomes and the biological interpretation of imaging features[142,143]. Advances in breast tumor radiomics has led to associations between radiomic signatures and molecular subtypes[144-146], combined

radiogenomic models for outcome prediction[101,147,148], and radiomic signatures correlated with biological pathways[123,149-151].

2.11. Radiomic analysis study designs

Effective use of radiomic analysis on medical images of breast cancer is highly dependent upon appropriate study design and statistical evaluation. There are numerous methods for radiomic feature extraction, resulting in a myriad of ways to quantify an image's texture. Consequently, having a high-dimensional radiomic feature set as compared to a relatively smaller sample size can result in the overfitting of a statistical learning model, resulting in false positive classification and over- or under- estimated statistical associations. Additionally, redundant radiomic features can often decrease performance accuracy. To alleviate this, methods such as principal component analysis or independent component analysis can be used to reduce the dimensionality of the radiomic feature set[152]. Feature selection methods can also be used to reduce radiomic feature redundancy and promote relevant features for analysis. Statistical correction methods, such as the Benjamini-Hochberg correction[153], can be used to reduce false positives in statistical association conclusions. Ideally, results should be validated using an independent dataset, to ensure veracity of the radiomic analysis. While finding comparable independent datasets is not always feasible, splitting the initial dataset into discovery and validation sets is an alternative solution to ensure repeatability. Similarly, cross-validation can be used to identify robust conclusions. Lastly, utilization of publicly available datasets and detailed methodology of specific radiomic parameters used during feature extraction, can allow for study repeatability.

2.12. Future Directions

Radiomic analysis is currently limited by user-defined feature parameters, such as selecting the number of discretized gray-levels within an image, image acquisition protocol, and image quality[25,96,154]. As such, literature largely lacks repeated studies performed on the same datasets with standardized radiomic feature extraction. Additionally, while radiomic features provide quantitative measures of breast tumor texture, the direct biological interpretation of specific radiomic feature values remains largely uncertain. Leveraging the specific properties of various functional MRI and PET imaging techniques, radiomic analysis can be used to quantify different tissue properties. Radiomic analysis applied to DCE-MRI or DW-MRI could provide insight on the distribution or longitudinal development of tumor vascularization and diffusion, respectively. Radiomic analysis applied to PET images utilizing biologically targeted radiotracers could provide insight into specific tumor physiology. Similarly, statistical associations between image derived radiomic features and histopathologic or genomic expression data could allude to a biological basis for tumor imaging presentations. Further work is needed to explore the relationships between specific radiomic features and underlying biology. Additionally, leveraging radiomic information from alternative imaging modalities in combination with that acquired from MRI could augment breast tumor characterization[155]. Advances in computational processing power such as increased memory and GPUs, have allowed for implementation of deep learning architectures for the analysis of breast tumors. Deep learning has the potential to use automated features and provide insights beyond “hand-crafted” radiomic features, which

are dependent upon user derived parameters. Using such technology however requires large datasets for training and may lack interpretability due to the higher order nature of resulting prognostic or predictive features.

Moving forward, computational medical image analysis has the potential for use in fully characterizing tumor biology to serve as a non-invasive quantitative tumor assay, complementing proteomic and genomic tumor analyses for a more comprehensive and personalized understanding of breast cancer.

Chapter 3 : Spatial heterogeneity in dynamic contrast enhanced magnetic resonance imaging

This chapter has been adapted from the following:

Chitalia, R.D., Rowland, J., McDonald, E.S., Pantalone, L., Cohen, E.A., Gastounioti, A., Feldman, M., Schnall, M., Conant, E. and Kontos, D., 2020. Imaging phenotypes of breast cancer heterogeneity in preoperative breast dynamic contrast enhanced magnetic resonance imaging (DCE-MRI) scans predict 10-year recurrence. *Clinical Cancer Research*, *26*(4), pp.862-869.

3.1. Introduction

This chapter provides details of exploring spatial heterogeneity in DCE-MRI images of breast tumors at baseline. We identify imaging phenotypes and aim to understand their relationship with conventional prognostic markers and patient outcomes to propose a non-invasive assessment of cancer.

Currently, critical disease treatment decisions are made on the basis of biomarkers acquired from tissue samples, typically obtained via core biopsy or surgical excision. Histopathologic assessment of this sample determines common prognostic markers including tumor size, shape, grade, nodal status, and metastasis. The prognostic and predictive markers derived from the limited diagnostic tissue samples may under-sample spatially heterogeneous breast tumors as well as overlook temporal shifts due to breast cancer progression or exposure to therapy. Therefore, there is a clinical need to develop prognostic and predictive markers of intratumor heterogeneity that may augment established biomarkers for personalized disease diagnosis, staging, management, and to assess treatment response to neoadjuvant therapy.

Medical imaging is currently used for breast cancer diagnosis, staging, and treatment response assessment, providing a means for longitudinal, non-invasive, whole-tumor evaluation of disease burden[16,17,57]. Dynamic contrast enhanced magnetic resonance imaging (DCE-MRI), in particular, is highly sensitive for primary lesion detection and staging, with the ability to assess tumor vascularization with contrast enhancement[17,20]. The field of "Radiomics" has shown promise in quantifying the imaging presentation of underlying tumor biology[8,28,101,109,116,156]. Identifying intrinsic radiomic phenotypes of breast cancer and understanding their relationship with patient outcomes and other histopathologic factors could complement conventional prognostic and predictive biomarkers. The purpose of this study was to identify and validate such intrinsic DCE-MRI radiomic phenotypes of breast cancer tumor heterogeneity and evaluate their independent prognostic performance in predicting 10-year recurrence, and their performance in augmenting established, histopathologic prognostic factors.

3.2. Discovery cohort: Study population and imaging protocol

Breast DCE-MRI scans were retrospectively analyzed from a previously completed, multimodality imaging trial conducted at our institution (2002- 2006; National Institutes of Health; P01CA85484) designed to evaluate an array of different breast imaging modalities in cancer staging, diagnosis, and screening. The study was Health Insurance Portability and Accountability Act-compliant, approved by the institutional review board at our institution, and in accordance with U.S. Common Rule. The trial originally recruited 901 women, including women with newly diagnosed breast cancer presenting for staging, women with a mammographically detected

suspicious finding or suspicious palpable mass directed to biopsy, and women eligible for high-risk screening. Informed written consent was obtained prior to trial participation. From these, 317 women were diagnosed with primary breast cancer, of which 231 were diagnosed with invasive breast cancer. From this subset of 231 women, 177 women had DCE-MRI images available for analysis. An additional 72 women were excluded for not receiving a consistent imaging protocol of fat-suppressed, T1-weighted DCE-MRI with at least two post contrast scans available for analysis. Lastly, 3 women were excluded on the basis of image quality, determined by biopsy artifacts and fiduciary markers, or the presence of diffuse disease in order to prevent inaccurate ROI segmentation, and 7 women were excluded due software algorithm output resulting in incomplete values during radiomic feature extraction. Therefore, a total of 95 women diagnosed with primary invasive breast cancer and imaged with a consistent T1-weighted DCE-MRI protocol with a first and second post-contrast acquisition, prior to any treatment, were included in our analysis. For this retrospective analysis, the requirement of informed consent was waived under institutional review board approval.

Following intravenous administration of gadolinium contrast, DCE-MRI images were acquired sagitally via a T1-weighted 3D protocol. Images were acquired with a 45-degree flip angle over a 16-18 cm field of view, with 2-2.5 mm slice thickness. Women subsequently underwent surgery for tumor removal. Histopathologic analysis of surgical specimens evaluated hormone receptor (HR) status, consisting of ER and PR status, HER2 status, clinical stage, size (cm) as determined from pathology sample, and surgical margins. Stage, Modified Bloom Richardson grade (MBRG), lymph invasion status, nuclear grade, and presence of ductal carcinoma in situ (DICS) were also documented.

Post-surgery therapy included a variable combination of chemotherapy, hormone therapy, and radiation. Recurrence-free survival (RFS) was monitored for all women over a 10-year follow-up period. Survival was determined as the date of breast cancer diagnoses to death or more recent follow-up. Patients without an event were censored at the date of last follow-up. In the discovery cohort, 11 women (12%) had recurrence events, and 84 women (88%) were event-free until their last available follow-up (Table 3.1). Clinical stage was statistically significantly associated with recurrence events ($p = 0.02$) (Table 3.1).

Table 3.1. Summary of patient characteristics from the discovery cohort.

Primary Invasive Cancers (*n*=95)			
	Non-recurrent cases at the time of last follow-up 84 (88% of total)	**Recurrent cases** 11 (12%)	Significance tested using Chi-square analysis
Malignant Pathology			**p=0.48**
Invasive Ductal Carcinoma (IDC)	65 (77% of NR)	8 (73% of R)	
Invasive Lobular Carcinoma (ILC)	8 (10%)	1 (7%)	
IDC/ILC	9 (11%)	2 (14%)	
Receptor Status			**p= 0/37**
Hormone Receptor Positive	61(73%)	9 (82%)	
HER2 Positive	20 (24%)	2 (18%)	
Triple Negative	11 (13%)	1 (9%)	
Clinical Stage			**p=0.02**
Early Stage (1)	35 (41%)	1 (9%)	
Advanced Stage (2-3)	45 (54%)	10 (91%)	
DCIS			**p=1**
Present	67 (80%)	9 (82%)	
Margins			**p=0.07**
Positive	40 (48%)	6 (55%)	
Negative	40 (48%)	5 (45%)	

3.3. Validation cohort: Study population and imaging protocol

An independent validation cohort was acquired from a subset of the ISPY-1/ACRIN 6657 trial (2002-2006)[157]. Women diagnosed with T3 breast tumors measuring 3cm or larger were enrolled in this trial, and underwent anthracycline-based neoadjuvant chemotherapy. DCE-MRI scans were acquired for women in this study as previously described[158]. The pre-treatment and pre-operative DCE-MRI images of 222 women were publicly available via The Cancer Imaging Archive[159]. From this, 15 women were excluded for having incomplete DCE acquisition or variability in imaging protocol. A further 43 women were excluded for having missing histopathologic data, RFS outcome, or pre-treatment DCE-MRI scans, and 1 woman was excluded due software algorithm output resulting in incomplete values during radiomic feature extraction. In all, 163 women were included in the validation cohort for this study; validation analysis utilized the scans that were both pre-treatment and pre-operative. Clinical information including HR status and HER2 status were available for each woman in the validation cohort. RFS status, defined as the time between first chemotherapy treatment and disease recurrence, was also available. A total of 44 women in the validation cohort (27%) had recurrent tumors (Table 3.2). A comparison between the two cohorts via chi-square analysis indicated a statistically significant difference between number of recurrent cases ($p = 0.02$), number of HR positive cases ($p = 0.02$), and clinical stage of tumors ($p < 0.001$) (Table 3.2).

Table 3.2. Summary of patient characteristics from the validation cohort (a). Statistical comparison between discovery and validation cohorts for covariates common in the two datasets (b).

a

	Non-recurrent cases 119 (73% of total)	**Recurrent cases 44 (27% of total)**	Significance tested using Chi-square analysis
Receptor status			**p=0.07**
Hormone receptor positive	67 (56%)	25 (56%)	
HER2 positive	34 (29%)	18 (40%)	
Triple Negative	29 (24%)	10 (23%)	

b

	Discovery n=95	**Validation** n= 163	Significance tested using Chi-square analysis
Recurrent cases	11 (12%)	44 (27%)	**p = 0.02**
Receptor Status			
Hormone receptor positive	70 (74%)	92 (56%)	**p = 0.02**
HER2 positive	22 (23%)	52 (32%)	**P=0.3**
Triple Negative	12 (13%)	39 (24%)	**p=0.4**
Clinical stage			**p < 0.001**
Early stage (1)	36 (38%)	0 (0%)	
Advanced Stage (2-3)	55 (58%)	163 (100%)	

3.4. Radiomic feature extraction

For each woman in the discovery cohort, the primary lesion was selected by a radiologist from the pre-treatment, pre-operative DCE-MRI scan, and manually segmented from the most representative slice, as determined by the largest tumor volume. This manual segmentation served as the initialization for 3-D tumor volume segmentation, which was performed using a previously validated, automated method[160] and visually verified by an expert after segmentation. Images were pre-processed using N3 bias-field normalization[161] and histogram normalization in order to correct for low frequency bias field signal or outliers that may induce artifacts within the image. Using the first (I_1) and second (I_2) post-contrast images, a signal enhancement ratio (SER) map

was generated(23) for the entire tumor volume, defined as the voxel-wise ratio between the first and second post contrast images:

$$(SER) = \frac{I_1 - I_0}{I_2 - I_0} \qquad \text{Eq. 3.1}$$

The first and second post-contrast images were acquired in succession at an average approximation of 90 seconds after contrast injection and first-post contrast scan, respectively. A multi-parametric, radiomic feature vector was extracted from the SER map for each woman, including a) previously validated morphologic features of tumor perimeter, area, ellipticity, and convexity, shown to be associated with disease progression[20,162], and b) radiomic features capturing structural[75], run-length[72-74], co-occurrence matrix[71], gray-level histogram, and gray-level size zone matrix textures[111], which were extracted and summarized over the primary lesion. Briefly, structural features capture intensity variations between central voxels and neighboring voxels. Run-length features measure the coarseness of an image in specific linear directions. Co-occurrence features analyze the spatial distribution of voxel intensity values by capturing frequency information of gray-level intensity values within a neighborhood of voxels in a specific linear orientation. Gray level histogram features are first order statistical features assessing the distribution of gray-level voxel intensities within an image. Gray level size zone features capture the connectedness of varying intensity levels within an image. Additionally, the mean and standard deviation of SER values of the tumor were calculated. Consequently, a total of 60 radiomic features were extracted. All features were extracted using the publicly available software, Cancer Imaging Phenomics Toolkit (CaPTk) (ver. 1.7.1, Univ. of Pennsylvania) (https://cbica.github.io/CaPTk/) [163].

3.5. Discovery of intrinsic imaging phenotypes

Prior to phenotype identification, the multi-parametric radiomic features extracted from the pre-operative DCE-MRI scans were z-score normalized[164]. Furthermore, z-scored features with extreme skewness or extremely low variations in distributions across women, defined as interquartile range (IQR) < 1 or kurtosis > 15, were excluded from further analysis to prevent a biased analysis. This resulted in a total of 22 features concatenated to form the final feature vector. Given the definition of each radiomic feature in the final feature vector, features were standardized such that a greater feature value indicates greater image heterogeneity. A tumor heterogeneity index was then generated for each woman, defined as the statistical average of z-score normalized, standardized features in the final feature vector. Thus, a higher heterogeneity index corresponds to higher intratumor heterogeneity whereas a lower heterogeneity index corresponds to increased intratumor homogeneity.

To identify intrinsic imaging phenotypes, unsupervised hierarchical clustering was performed on the extracted, multi-parametric feature vectors for women in the discovery cohort[36]. The *k* clusters obtained from the unsupervised hierarchical clustering algorithm are interpreted as intrinsic imaging phenotypes in the population. Briefly, an agglomerative approach was used to create a hierarchical clustering of women, using Euclidean distance for the distance between feature vectors and Ward's minimum variance method as the clustering criterion[165]. The optimal number of distinct phenotypes, *k*, was determined by assessing the stability and significance of each phenotype for each value of *k* that was considered. The optimal number of stable

phenotypes was determined using consensus clustering[166], where the dataset was subsampled and cluster arrangements were determined using varying numbers of *k*. For each number of *k* phenotypes, the proportion that two women occupied the same phenotype cluster out of the number of times they appeared in the same subsample was determined and stored in a symmetric consensus matrix, from which a cumulative distribution function (CDF) was determined. Cluster stability, as determined by the area under the CDF curve, was evaluated for each increase in k phenotype, with a change in stability less than 10% deemed insignificant. Statistical significance of the identified, stable phenotypes was evaluated using the SigClust method[167]. Here, the significance of the cluster index, defined as the sum of within- cluster sums of squares about the cluster-mean divided by the total sum of squares about the overall mean, was tested against a null distribution, simulated using 10,000 samples from a Gaussian distribution fit to the data. The test was performed at each phenotype split to determine statistical significance ($p<0.05$).

3.6. Independent validation of intrinsic imaging phenotypes

Tumor segmentation for cases from the validation cohort was performed per the ISPY-1/ACRIN 6657 protocol[158]. The 22 features identified from the discovery cohort were extracted from segmented tumors in the validation cohort using the same feature pre-processing steps outlined above, to form the final feature vectors for hierarchical unsupervised clustering analysis. These features were normalized using the mean and standard deviation values of each respective feature's distribution in the discovery cohort,

to standardize feature ranges. To validate identified phenotype reproducibility, women in the validation cohort were assigned to the discovery cohort- identified phenotypes by minimizing the Euclidian distance between each validation cohort feature vector and the discovery cohort-identified phenotype centroid. The significance and reproducibility of phenotype assignment in the validation cohort was assessed using Consensus Clustering and the SigClust methods (Figure 3.1).

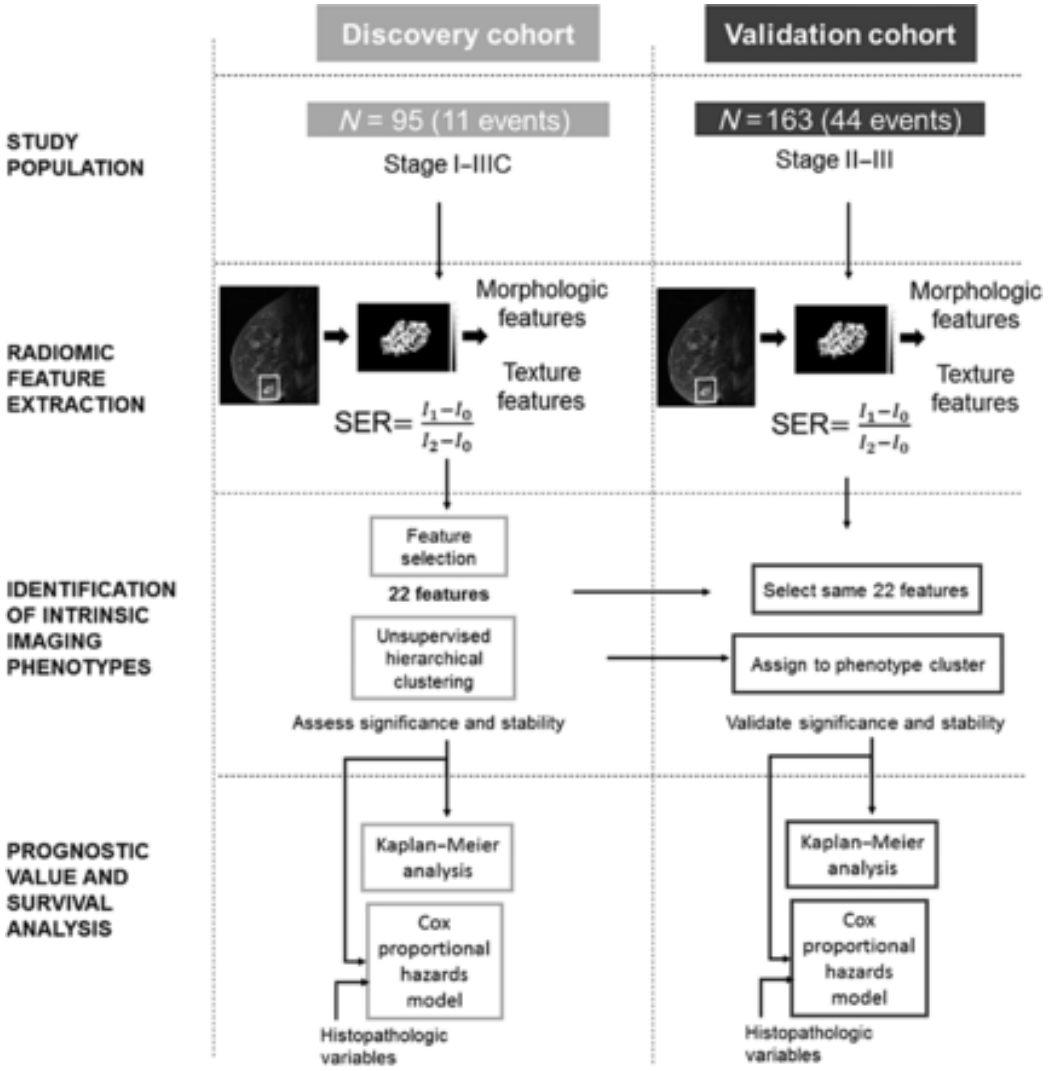


Figure 3.1. Study design. Radiomic features were extracted from SER maps generated from preoperative DCE-MRI scans from women in the discovery cohort (n = 95). Feature selection resulted in a 22-feature feature vector. Unsupervised hierarchical clustering was used to identify intrinsic imaging phenotypes, which were assessed for statistical significance and stability. The same 22 features were extracted from the preoperative DCE-MRI scans of 163 women in an independent validation cohort. Women in the validation cohort were assigned to a phenotype identified in the discovery cohort by minimizing the distance between their 14-feature feature vector and the corresponding phenotype centroids. The independent and additional prognostic values of heterogeneity phenotypes were assessed via Kaplan–Meier RFS analysis and Cox proportional hazards models when compared with a baseline model of established histopathologic biomarkers.

3.7. Prognostic value of imaging phenotypes

We assessed the distribution of histopathologic, prognostic covariate values for women assigned to each heterogeneity phenotype using chi-square tests for categorical biomarker values and Kruskal-Wallis tests for continuous biomarker values. The distributions of post-surgery therapy received by each woman (i.e., chemotherapy, hormone therapy, and radiation therapy) and recurrence-free survival were assessed across phenotypes to identify any associations between therapy and RFS or heterogeneity phenotypes.

RFS probabilities across heterogeneity phenotypes within the discovery and validation cohorts were evaluated using Kaplan-Meier curves, with log-likelihood statistical tests used to assess their significance and determine their independent prognostic value. To determine the additional value categorizing tumor heterogeneity phenotypes, a baseline Cox proportional hazards model was built using the established histological prognostic factors of HR and HER2 status. Performance of this model in predicting RFS was tested both with and without phenotype cluster assignment, coded as a categorical variable.

3.8. Study results: Discovery of intrinsic imaging phenotypes

Three statistically significant phenotypes were identified in the discovery cohort via unsupervised hierarchical clustering and found to be statistically significant via the SigClust methods ($p < 0.01$) (Figure 3.2). Ordering the heterogeneity indices of the corresponding centroids for the identified phenotypes in ascending order allowed for the identified phenotypes to be interpreted as phenotypes of low, medium, and high

intratumor heterogeneity (Figure 3.3). The number of recurrences were statistically significantly different across the heterogeneity phenotypes via chi-square analysis ($p = 0.01$).

Kaplan-Meier RFS curves for women stratified by heterogeneity phenotype assignment were found to be statistically significantly different, as determined using the log-rank test ($p < 0.05$). A baseline Cox-proportional hazards model consisting of HR status and HER2 status resulted in a c-statistic of 0.55 when predicting 10-year RFS. Adding heterogeneity phenotype assignment to the baseline model resulted in a c-statistic of 0.73 A log-likelihood test showed statistically significant improvement in the augmented model performance ($p = 0.007$) (Figure 3.2).

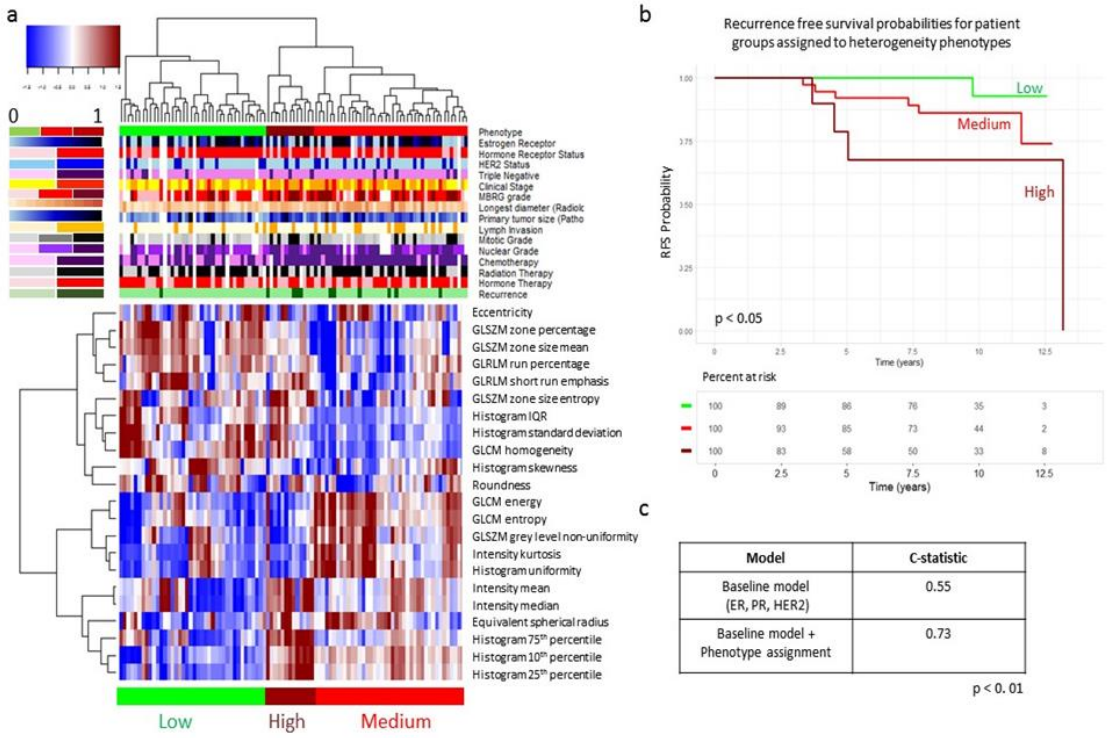


Figure 3.2. Identification of intrinsic imaging phenotypes of tumor heterogeneity. Unsupervised hierarchical clustering of SER features identifies three intrinsic phenotypes in the discovery cohort (**A**). RFS curves for women stratified by imaging heterogeneity phenotype show that heterogeneity phenotype is statistically significant ($P < 0.05$) when predicting RFS (**B**). Adding phenotype information to Cox regression model shows an improvement in c-statistic when predicting recurrence events (**C**).

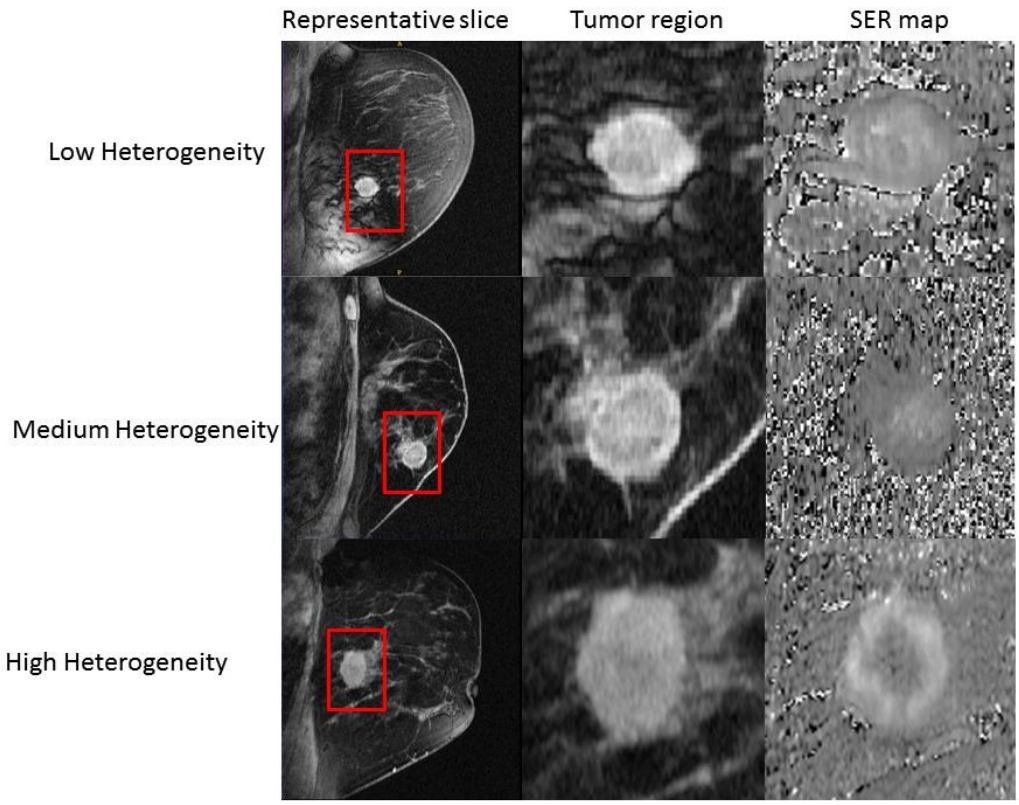


Figure 3.3. Representative cases from heterogeneity phenotypes. Representative slice, tumor region, and SER map from a pre-menopausal woman diagnosed with primary stage-I, HR+, HER2- , node-negative breast cancer assigned to the low heterogeneity phenotype (top), a peri-menopausal woman with primary stage-II, HR-, HER2+, node- negative breast cancer assigned to the medium heterogeneity phenotype (middle), and a pre-menopausal woman with primary stage-II, ER-, HER2-, node-positive breast cancer assigned to the high heterogeneity phenotype (bottom).

Analysis of clinical covariate significance across heterogeneity phenotype status showed that differences in tumor MBRG, estrogen receptor percentage, and tumor mitotic stage were statistically significant across heterogeneity phenotypes ($p = 0.03$, $p=0.001$, and $p=0.02$, respectively). Of poorly differentiated (MBRG 8-9) tumors, 80% were assigned to the medium or high heterogeneity phenotypes, and 20% were assigned to the low heterogeneity phenotype (Figure 3.4a). Tumors assigned to the low and medium heterogeneity phenotypes had median estrogen receptor percentages of 75% and

70% respectively, while tumors assigned to the high heterogeneity phenotype had median estrogen receptor percentages of 40% (Figure 3.4b). Of tumors with high mitotic stages, 81% were assigned to the medium or high heterogeneity phenotypes, and 19% were assigned to the low heterogeneity phenotype (Figure 3.4c).

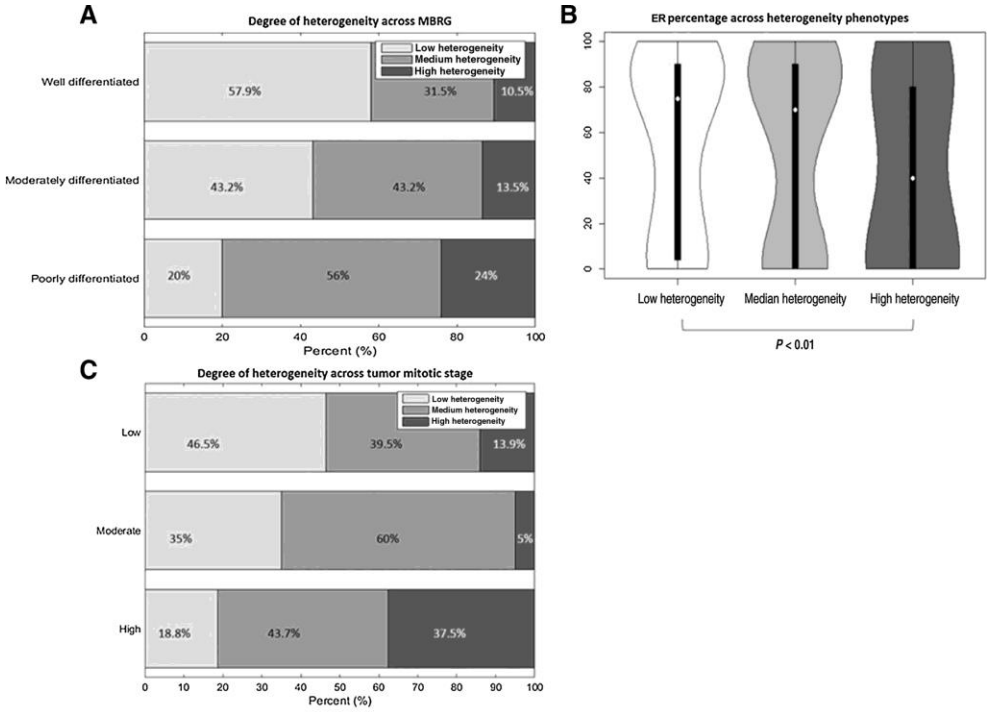


Figure 3.4. Associations between histopathologic prognostic markers and heterogeneity phenotypes. Associations between histopathologic prognostic markers and heterogeneity phenotypes identified in the discovery cohort. Degree of phenotypic heterogeneity in well, moderately, and poorly differentiated tumors (**A**). Percent of estrogen receptor distribution for women in low, medium, and high heterogeneity phenotypes (**B**). Degree of phenotypic heterogeneity in tumors with low, moderate, and high mitotic stage (**C**).

3.9. Study results: Validation of intrinsic imaging phenotypes

Women in the discovery cohort assigned to the low, medium, and high heterogeneity phenotypes had average heterogeneity indices of -0.09, -0.03, and 0.16, respectively (Figure 3.5a). Women in the validation cohort assigned to the low, medium, and high heterogeneity phenotypes had average heterogeneity indices of -0.05, -0.24, and 0.21, respectively (Figure 3.5b).

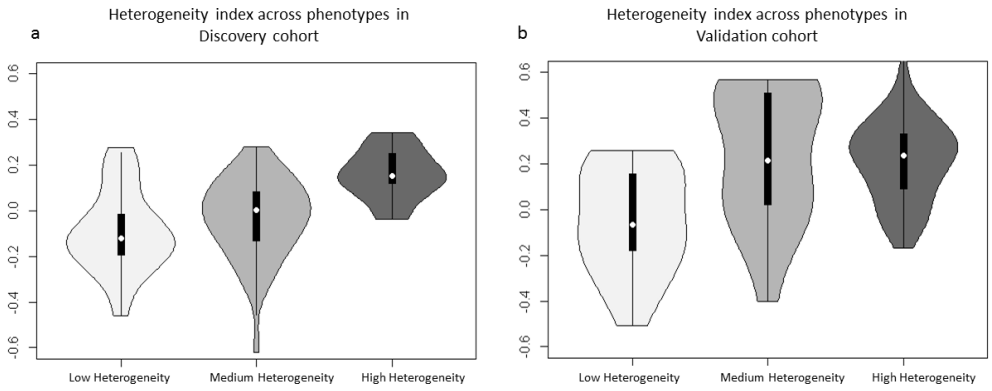


Figure 3.5. Heterogeneity index* distributions of women in low, medium, and high heterogeneity phenotypes in the discovery (A) and validation (B) cohorts. *Defined as the statistical average of z-score normalized, heterogeneity standardized features in the final feature vector for each tumor in the discovery and validation cohorts.

The heterogeneity phenotypes were found to be reproducible and statistically significant in the validation set using the SigClust methods ($p = 0.01$). Kaplan-Meier RFS curves for women stratified by phenotype clustering assignment were statistically significantly different ($p = 0.01$) (Figure 3.6). A baseline Cox-proportional hazards model consisting of HR status and HER2 status resulted in a c-statistic of 0.61 when predicting 10-year RFS. Adding heterogeneity phenotype assignment to the baseline model resulted in a c-statistic of 0.67 A log-likelihood test showed statistically significant improvement in the augmented model performance ($p = 0.01$).

Analysis of clinical covariate significance across heterogeneity phenotype status showed that differences in progesterone receptor status were statistically significant across heterogeneity phenotypes (p = 0.03).

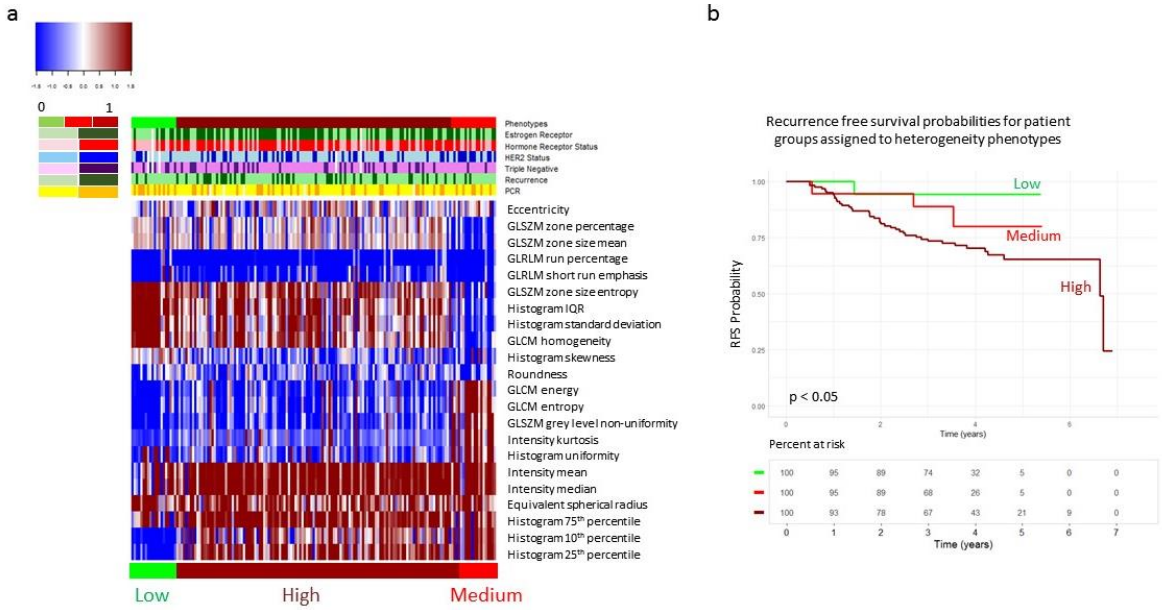


Figure 3.6. Independent validation of intrinsic imaging phenotypes of tumor heterogeneity. Phenotypes identified in the discovery cohort are significantly reproducible in the validation cohort (A). RFS curves for women stratified by imaging heterogeneity phenotype show that heterogeneity phenotype is statistically significant (p = 0.01) when predicting RFS (B).

3.10. Discussion

Our results indicate that distinct imaging phenotypes exist within invasive breast tumors which correspond to different degrees of intratumor heterogeneity, suggesting that radiomic features can non-invasively characterize such heterogeneity patterns. The identification and validation of distinct image heterogeneity phenotypes show that the phenotype clusters identified are both interpretable and meaningful. Most notably, the

validated heterogeneity phenotypes show independent prognostic value when predicting 10-year RFS, indicating that intrinsic imaging phenotypes can potentially identify intratumor heterogeneity features driving aggressive tumor behavior. Women assigned to the high heterogeneity phenotype demonstrated decreased probabilities of RFS over the 10-year follow-up period, corroborating the hypothesis that heterogeneous tumors are associated with aggressive tumor behavior and treatment resistance.

Of particular note, the medium heterogeneity phenotype encompasses a wide range of tumors, as evidenced by the differences in Kaplan-Meier survival analysis between women assigned to the medium heterogeneity phenotype in the discovery and validation cohorts. While both cohorts represent populations of women diagnosed with invasive breast cancer, all women in the validation cohort were diagnosed with advanced stage disease, and therefore eligible for neoadjuvant therapy. As the discovery cohort consists of more diverse disease stages, women in the validation cohort assigned to the low heterogeneity phenotype may have higher degrees of tumor heterogeneity as compared to women in the discovery cohort also assigned to the low heterogeneity phenotypes and may therefore be more similar to tumors assigned in the discovery cohort as having a medium heterogeneity phenotype. This is supported by the similarity of the average heterogeneity indices for women in the validation cohort assigned to the low heterogeneity phenotype versus women in the discovery cohort assigned to the medium heterogeneity phenotype (-0.05 and -0.03, respectively). Consequently, survival probabilities for women in the validation cohort assigned to the low heterogeneity phenotype may be more similar to women in the discovery cohort originally assigned to

the medium heterogeneity phenotype, as compared to women in the low heterogeneity phenotype. A statistical comparison between cohorts suggests women in the validation cohort had a statistically significantly higher proportion of HR negative tumors, thereby suggesting more aggressive tumor behavior and outcome (Table 3.2).

Imaging phenotypes of intratumor heterogeneity also provide additional prognostic value when augmenting established histopathologic prognostic biomarkers. The independent and additional prognostic value of phenotype assignment suggests that imaging phenotypes can provide unique information about underlying tumor behavior, and therefore, complement clinically utilized prognostic markers for personalized prognosis and decision making.

Higher degrees of imaging phenotype heterogeneity were shown to be associated with poorly defined tumors as per the MBRG and tumors with higher mitotic grades. The increased mitotic grade of tumors in the higher heterogeneity phenotypes may contribute to the genetic diversity and sub-clonal evolution thought to increase intratumor heterogeneity[39]. These results suggest that tumor characteristics such as nuclear pleomorphism and increased mitotic rates, which are characteristic of aggressive tumor behavior, may be captured by the imaging phenotypes of tumor heterogeneity. Additionally, tumors with lower percentages of estrogen receptors displayed statistically significantly higher imaging phenotype heterogeneity, correlating with established hypotheses that estrogen positive tumors are associated with more positive prognoses[168,169]. The statistically significant associations of histopathologic prognostic

covariate distributions across heterogeneity phenotypes indicate that image heterogeneity may correlate with underlying tumor biology.

We have identified intrinsic imaging phenotypes of intratumor heterogeneity in primary invasive breast cancer that can independently predict 10-year recurrence and have validated these findings in an independent cohort. While previous studies have utilized hierarchical clustering analysis to identify imaging phenotypes or have investigated relationships between radiomic features and histopathologic and genomic tumor characteristics, most of these studies have used surrogate measures of recurrence or were limited by a lack of independent validation[101,147].

Limitations to our study should be noted. For this exploratory analysis, we chose a fixed set of radiomic features. The independent validation cohort utilized for this study consisted of only advanced stage tumor diagnoses with a limited availability of histopathologic prognostic biomarkers, as opposed to the discovery cohort which consisted of both early and advanced stage tumors and a wide array of histopathologic prognostic biomarker information available for each woman. As both cohorts included women diagnosed with invasive breast cancer and follow-up information, we determined that validating our heterogeneity phenotypes with a more niche cohort can still demonstrate the added value and generalizability of tumor imaging heterogeneity phenotypes. Additionally, we limited the histopathologic prognostic biomarkers used for baseline model regression analysis of the discovery cohort to match those covariates available in the validation cohort. Future work will aim to expand this analysis to a larger

cohort. Lastly, we will seek to further explore the added prognostic value of imaging phenotypes to that of emerging prognostic molecular profiling assays.

In conclusion, our results demonstrate that intrinsic imaging phenotypes of tumor may heterogeneity exist, with independent and additional prognostic value in predicting RFS. Additionally, these heterogeneity phenotypes show associations with established histopathologic prognostic biomarkers, suggesting that image heterogeneity phenotypes non-invasively capture underlying tumor biology.

Chapter 4 : Longitudinal changes in spatial heterogeneity in dynamic contrast enhanced-magnetic resonance imaging in response to breast neoadjuvant chemotherapy

This work is under review.

Rhea Chitalia[1,2], Marios Miliotis[3,4], Nariman Jahani[2], Spyros Tastsoglou[3,4], Elizabeth S. McDonald[2], Vivian Belenky[2], Eric A. Cohen[2], David Newitt[5], Laura J. van 't Veer[6], Laura Esserman[6], Nola Hylton[5], Angela DeMichele[7], Artemis Hatzigeorgiou[3,4], Despina Kontos[2§]

Departments of [1]Bioengineering, [2]Radiology, and [7]Medicine, Division of Hematology/Oncology, University of Pennsylvania, Perelman School of Medicine 3400 Spruce Street, Philadelphia, PA 19104, USA, [3]Department of Computer Science and Biomedical Informatics, University of Thessaly, Lamia, Greece [4]DIANA-Lab, Hellenic Pasteur Institute, Athens, Greece, [5]Department of Radiology and Biomedical Imaging, [6]Department of Surgery and Oncology, University of California at San Francisco, USA.

4.1. Introduction

Breast cancer heterogeneity is well-established, with intratumor heterogeneity arising due to genomic and transcriptomic variations leading to heterogeneous subpopulations driving prognosis and response to therapy[4,11,39,170,171]. As such, increased heterogeneity is thought to be associated with adverse clinical outcomes[6].

Neoadjuvant chemotherapy (NACT) is an established course of treatment for locally advanced breast cancer (LABC) and can promote breast conserving surgeries by reducing tumor size[21]. Additionally, women achieving pathologic complete response (pCR) after completing neoadjuvant chemotherapy may have improved survival outcomes[172,173]. Early prediction of response to neoadjuvant treatment can allow for personalized changes to treatment plans, including targeted therapies, and early

discontinuation of inactive therapies[174,175]. Intratumor heterogeneity is thought to change in response to neoadjuvant chemotherapy, leading to altered biomarker expressions[176]. Such changes may arise due to the acquired resistance of specific sub-clones during treatment[177]. Early, noninvasive characterization of such changes may indicate response versus resistance to treatment, enabling early treatment changes prior to treatment completion.

Personalized gene expression based molecular assays, such as the 70-gene MammaPrint microarray assay (Agendia BV) and the 50-gene PAM50 risk of recurrence score assay (ROR-S), provide risk stratification for future recurrence[178,179]. p53 mutation status is an established predictor for more aggressive tumor biology and therefore a worse prognosis in terms of recurrence free survival (RFS)[180]. Such precision medicine predictors may improve clinical decision making by deviating from the "one size fits all" approach to treating breast cancer. However, as such assays, mutation statuses, and established histopathologic biomarkers are determined largely from selective tissue sampling acquired by biopsy, they may fall short in fully capturing heterogeneous disease burden.

Dynamic contrast enhanced magnetic resonance imaging (DCE-MRI) can allow for longitudinal, non-invasive monitoring of heterogeneous tumors during the course of neoadjuvant chemotherapy. Previous studies have demonstrated the role of longitudinal patterns for tumor response during neoadjuvant chemotherapy and have examined their associations with treatment response and overall survival[69,104,181,182]. Hylton et al. demonstrated the prognostic and predictive value of measuring functional tumor volume

(FTV) at various longitudinal time points during neoadjuvant chemotherapy[183]. Jahani et al. developed registration-based biomarkers for the early prediction of pCR and recurrence free survival (RFS) in tumors from baseline to early treatment time points[184]. While much progress has been made, these studies may be limited by not examining the associations between aggregate changes in intratumor heterogeneity that arise in response to therapy and the complementary information provided by genomics-based information[113].

The purpose of this study was to identify imaging phenotypes of early changes in intratumor heterogeneity in DCE-MRI and evaluate their prognostic value in augmenting FTV measures and molecular profiling signatures scores for predicting RFS after breast NACT.

4.2. Discovery cohort

DCE-MR images of women enrolled in the ACRIN 6657/I-SPY 1 trial, diagnosed with advanced invasive breast cancer from May 2002 through March 2006, were retrospectively analyzed[157,159]. Per the inclusion criteria of ACRIN 6657/I-SPY 1, women diagnosed with stage 2 or 3 breast cancer were selected for the study and underwent an anthracycline-cyclophosphamide NACT. Longitudinal DCE-MRI was performed using a 1.5 T scanner at four time points: prior to the start of neoadjuvant therapy (T1), at least 2 weeks after the first cycle of chemotherapy (T2), between treatments (T3), and after the completion of chemotherapy, before surgery (T4). Data acquisition was as described in

the ACRIN 6657/I-SPY 1 protocol[182]. The first and second post-contrast images were acquired 2.5 and 7.5 minutes after contrast injection.

Of the 222 trial participants with publicly available data[157,159], we retained the 143 women for whom both complete clinical data and T1 and T2 DCE-MR imaging were available. For analyses involving gene expression, we used the subset of 100 women for whom gene expression information was available through the Gene Expression Omnibus[185,186], under the accession number GSE22226[187]. Clinical and histopathologic data including age, hormone receptor (HR) status, human epidermal growth factor receptor 2 (HER2) status, and pCR status were available for each woman. Functional tumor volume at T2 (FTV_2), previously shown to have significant association with RFS[183], was also calculated for each woman. RFS times were available, defined as time to recurrence (event), or time to death or last follow-up (censor).

4.3. Validation cohort

A validation cohort of 92 women was formed from the remaining 43 women from the original cohort (n=143) for whom gene expression data was not publicly available, and a separate dataset of 49 women from the publicly available Breast MRI NACT Pilot study [188]. This study had similar inclusion criteria as the I-SPY 1 trial, and participants underwent a similar treatment and imaging protocol as the I-SPY 1 trial. Clinical information on age, HR status, and HER2 status and 3-year RFS information was available for each woman in the validation cohort.

4.4. Approximation of gene expression based molecular profiling signatures

Molecular profiling of the I-SPY 1 enrolled women with gene expression information was built as previously described[189]. Specifically, we re-created three gene signatures in order to classify tumors regarding their metastatic potential, risk of recurrence, and p53 oncogene mutation status: the 70-gene signature (MammaPrint)[190,191], PAM50 risk of recurrence (ROR-S)[192,193], and p53 mutation signature[194], respectively. Briefly, MammaPrint classification was achieved by calculating the cosine similarity of the expression of the 70-gene signature for each sample against a "good prognosis" sample set[190], using thresholds as defined in the original study[191]. ROR-S sample categorization was determined by computing the weighted sum of the correlation coefficients[193] of each sample against the intrinsic subtype sample sets of the PAM50 gene signature study[192]. Lastly, p53 mutation status was estimated by calculating the proximity of the I-SPY 1 samples and the p53 mutation signature centroids (wildtype *vs.* mutant) as Spearman's correlation values, as described in the p53 gene signature study[194]. The integrity of our classification was examined by comparing our results with the original results of the Esserman et al. study[187]. We confirmed that our recreated results corresponded to the original results by comparing the numbers of individuals attributed to each class in the overall cohort.

4.5. Delta radiomic feature extraction

For each woman in the discovery cohort, the 3-D primary lesions at pre-treatment (T1) and early-treatment (T2) time points were selected by first identifying the functional

tumor volume (FTV) within the publicly available bounding region, as previously reported[158]. The largest contiguous volume of voxels included in the FTV was selected as the location for the primary lesion; this volume was then further refined using manual segmentation to remove isolated voxels and include voxels within the primary tumor lesion volume which were not initially selected by the FTV threshold[158]. Final tumor segmentations for T1 and T2 were visually confirmed by a board-certified and fellowship trained breast imaging radiologist (ESM). Images were preprocessed by N3 bias-field normalization to correct for bias field signal[161].

For each woman in the discovery cohort, at T1 and T2 time points, four voxel-wise kinetic image maps were calculated within the segmented tumor, the peak enhancement (PE) (Eq. 4.1), signal enhancement ratio (SER) (Eq. 4.2), wash-in slope (WIS) (Eq. 4.3), and wash-out slope (WOS) (Eq. 4.4) images, to quantify the enhancement patterns over the dynamic scans using the signal intensity for the pre-contrast, first post-contrast, and second post-contrast time points (I_0, I_1, and I_2, respectively).

$$PE = \max_{t=t_{PE}} \frac{I_t - I_0}{I_0} \quad \text{Eq. 4.1}$$

$$SER = \frac{I_1 - I_0}{I_2 - I_0} \quad \text{Eq. 4.2}$$

$$WIS = \begin{cases} \dfrac{PE}{t_{PE} - t_0} & if\ t_{PE} \neq 0 \\ 0 & otherwise \end{cases} \quad \text{Eq. 4.3}$$

$$WOS = \begin{cases} \dfrac{I_2 - I_1}{t_2 - t_{PE}} & if\ t_2 \neq t_{PE} \\ 0 & otherwise \end{cases} \quad \text{Eq. 4.4}$$

All kinetic image maps and tumor segmentations were resampled by linear interpolation to a spatial resolution of 256 x 256 voxels, the lowest resolution of the data cohort, to ensure consistent resolution across all scans. A total of 104 radiomic features characterizing lesion intensity, texture patterns, and morphology were extracted from the entire tumor region, from each kinetic map at each treatment time point, resulting in a total of 416 features at each time point for each woman. All features were extracted using the publicly available Cancer Imaging Phenomics Toolkit (CaPTk; v.1.7.1; University of Pennsylvania; https://cbica.github.io/CaPTk/)[163]. Features at each treatment time point (f_{T1} and f_{T2}) were subsequently sign-adjusted such that increasing feature values corresponded to increasing lesion heterogeneity as per each feature's definition. Subsequently, the change in each radiomic feature between the baseline and early treatment time points, or delta feature Δf, was calculated as:

$$\Delta f = \frac{f_{T2} - f_{T1}}{f_{T1}} \quad \text{Eq. 4.5}$$

These delta features were subsequently z-score normalized and features with extreme skewness or low interquartile range (i.e., skewness>5, IQR<1) were excluded

from further analysis. Features characterizing tumor texture or morphology in only 2-D image dimensions were also excluded to allow for a whole-tumor, 3-D analysis. This resulted in a total of 42 delta features included in our final analysis. To reduce dimensionality and identify correlated delta features, features were clustered in an agglomerative hierarchical manner using Pearson's correlation as the distance metric, with highly correlated features being grouped together. Consensus clustering was used to determine the optimal number of stable delta feature groups, with each feature group consisting of highly correlated delta features. Within each feature group, principal component analysis (PCA) was performed and principal components (PCs) totaling greater than 85% explained variance were retained to represent each feature group[195]. As higher values for each delta radiomic feature prior to PCA indicated increasing heterogeneity from T1 to T2, higher values of a PC incorporating primarily positive contributions of features were interpreted as increasing heterogeneity, and one with negative contributions were interpreted as decreasing heterogeneity[196]. The PCs found, and their subsequent use in identifying imaging phenotypes of tumors, could serve to characterize tumors as having radiomic signatures indicating increasing or decreasing heterogeneity.

4.6. Identifying imaging phenotypes of early change in tumor heterogeneity

To identify imaging phenotypes of early changes in tumor heterogeneity, tumors in the discovery cohort were classified via unsupervised hierarchical clustering, using the retained principal components to represent each tumor. The clusters identified through unsupervised clustering were interpreted as phenotypes of changes in heterogeneity seen

in the study population. An overview schematic for how imaging phenotypes were generated can be found in Figure 4.1. An agglomerative hierarchical approach was used to cluster tumors, using Euclidean distance as the distance metric between the retained principal components for each tumor. Ward's minimum variance method was used as the clustering metric[165]. To determine the optimal *k* number of clusters, consensus clustering[166] was used to determine the number of stable phenotypes by repeatedly subsampling the data, performing unsupervised hierarchical clustering, and noting the proportion of subsamples in which every pair of tumors occupied the same cluster when they appeared in the same dataset. As such, a cumulative distribution function (CDF) was determined for each increase in *k*, and the stable number of clusters was determined to be the *k* at which the area under the CDF increased less than 10%. SigClust[167] methods were used to determine the number of significant phenotypes by calculating the significance of the cluster index, a metric defined as the sum of within cluster sum of squares about the overall mean, tested against a null distribution at each cluster division. The significance of each phenotype split was tested at $p < 0.05$ (Figure 4.1).

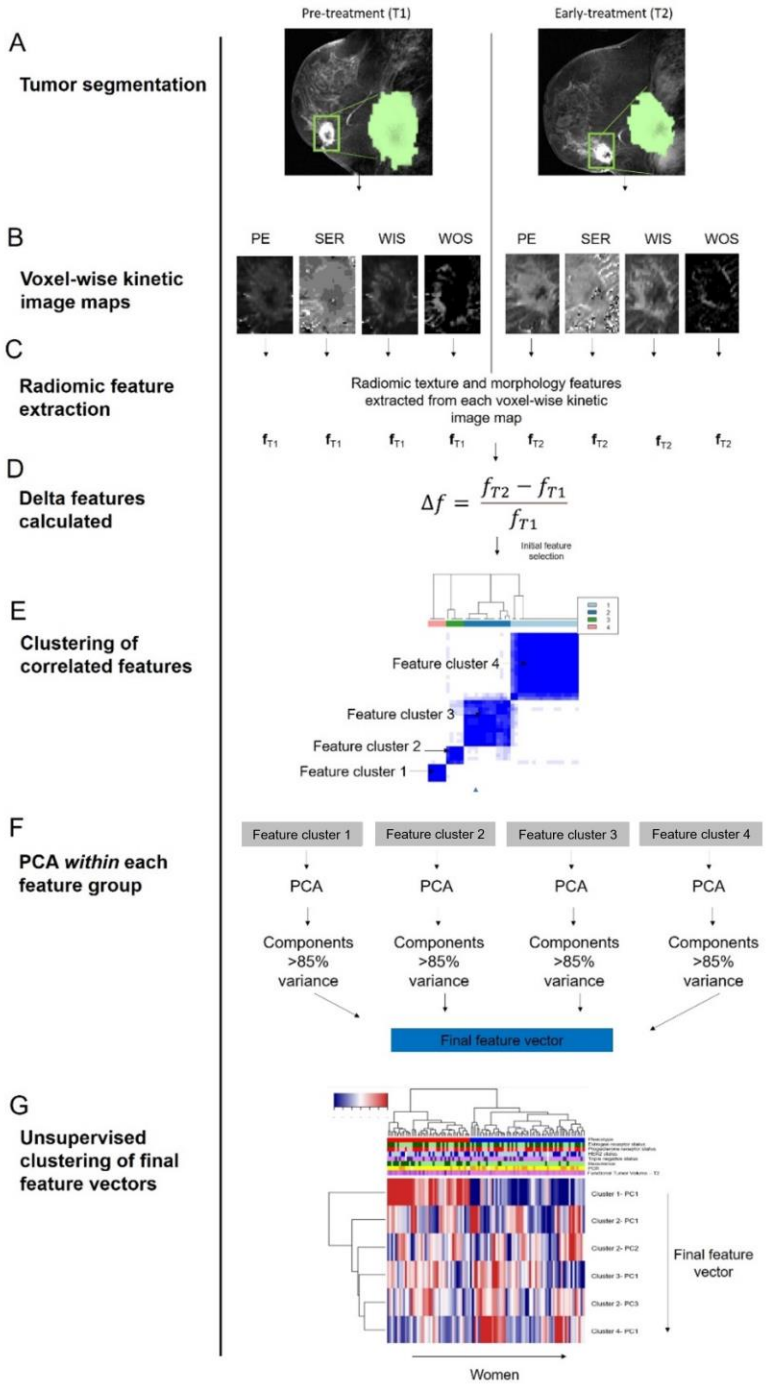


Figure 4.1. Identification of change in heterogeneity phenotypes. Primary lesion volume segmented from pre-treatment (T1) and early-treatment (T2) DCE-MRI images (A). Four voxel wise kinetic image maps are created for peak enhancement (PE), signal enhancement ratio (SER), wash-in slope (WIS), and wash-out slope (WOS) images to quantify enhancement patterns over each dynamic scan from T1 and T2 images (B). Radiomic features are extracted from each kinetic image map from both T1 and T2 images (C). Delta radiomic features are calculated for each extracted feature (D). Delta radiomic features are clustered based on correlation, and consensus clustering is used to determine the optimal number of stable feature clusters

(E). Within each feature cluster, PCA is performed and the number of principal components is selected to account for >85% of explained variance (F). Selected principal components are concatenated across all feature clusters to form the final feature vector for each woman. Unsupervised hierarchical clustering of the final feature vectors is performed to identify imaging phenotypes seen in the study population (G).

4.7. Prognostic value of early change in heterogeneity phenotypes-statistical analysis

Distributions of clinical and histopathologic covariate values and molecular profiling scores were assessed for differences across radiomic phenotypes using Chi-square and Kruskal-Wallis tests for categorical and continuous covariates, respectively. Statistical corrections for multiplicity were made using the Bonferroni correction[197].

RFS times across phenotypes were evaluated using Kaplan-Meier survival curves, in both the whole cohort and within strata of HR status, HER2 status, TN status, and greater than and less than median FTV_2 values, with the log rank test used to determine statistical significance. RFS was also modeled via Cox proportional-hazards regression. Eight models were evaluated: univariable models for each molecular signature; the baseline model — using the covariates age, HR status, and HER2 status; baseline + FTV_2; baseline + FTV_2 + radiomic phenotype; and baseline + FTV_2 + all molecular signatures, both with and without the addition of radiomic phenotype. All models were evaluated using 5-fold cross validation and averaged over 100 replicates.

The prognostic value of radiomic phenotypes was further evaluated by generating a risk score for each woman, defined as the prediction score of covariates weighted by the corresponding Cox-proportional hazard's coefficients. Kaplan-Meier survival was analyzed split on the median risk calculated by the Cox model using baseline factors and FTV_2.

Lastly, confusion matrices for the categories of RFS event/censor were generated to assess the predictive performance of radiomic phenotypes compared to MammaPrint scores, ROR-s, and p53 mutation status.

4.8. Validation of early change in heterogeneity phenotypes

Tumor segmentations for cases in the validation cohort were generated similarly to those in the discovery cohort. Delta radiomic features were calculated using the same feature preprocessing methods used in the discovery cohort. The same delta features selected in the discovery cohort were also selected for the validation cohort. These resulting delta features were normalized using the mean and standard deviation values from the delta feature values in the discovery cohort to standardize feature ranges.

Features were subsequently grouped together based on the cluster assignment of correlated features determined from the discovery cohort. Within each validation feature cluster, features were projected into the discovery cohort feature groups' principal component space to determine component values. The same numbers of PCs summarizing each feature group retained in the discovery cohort were selected from the validation cohort to form the validation cohort principal-component vectors.

To determine phenotype assignment in the validation cohort, each tumor was assigned to the discovery cohort-identified phenotypes by minimizing the Euclidean distance between each validation cohort principal component vector and the discovery cohort phenotype centroid, defined as the average of the principal component vectors across all tumors in each phenotype.

4.9. Results

4.9.1. Discovery cohort

A total of 28 (28%) women included in the discovery cohort had future events of recurrence while 72 (72%) women did not have future events of recurrence (Table 4.1). Median RFS time was 3.9 years (range, 0.5–6.9 years)[158]. Neoadjuvant and radiation therapy information was available for women in the discovery cohort (Table 4.2).

Table 4.1 Selected patient characteristics for discovery cohort.

	No future event of recurrence (n=72)	**Future event of recurrence (n=28)**
Hormone Receptor positive	28 (53%)	17 (61%)
HER2+ positive	23 (32%)	11 (39%)
pCR	23 (32%)	4 (14%)
Age (min-max)	48.15 (33.18-64.33)	46.31 (28.76-65.39)

Table 4.2. Selected treatment characteristics for discovery cohort.

Locally advanced cancers (n=100)

	No future event of recurrence (n=72)	**Future event of recurrence (n=28)**	**p-value**
Neoadjuvant Chemotherapy			> 0.99
Anthracycline-Cyclophosphamide (AC) only	1(1.4%)	0 (0%)	
AC + Tamoxifen	62 (86%)	24 (86%)	
AC + Tamoxifen + Herceptin	8 (11%)	3 (11%)	
AC + Tamoxifen + Other	1 (1.4%)	1 (4%)	
Herceptin	8 (11%)	3 (11%)	> 0.99
Radiation Therapy	58 (81%)	19 (68%)	0.21

4.9.2Validation cohort

Of the women included in the validation cohort, 27 (29%) women had future events of recurrence while 65 (71%) did not (Table 4.3). Median RFS time was 4.13 years (range, 0.28-8.79 years).

Table 4.3. Selected patient characteristics for validation cohort.

	No future event of recurrence (n=65)	**Future event of recurrence (n=27)**
Hormone Receptor positive	37 (57%)	10 (37%)
HER2+ positive	19 (29%)	10 (37%)
pCR	16 (25%)	4 (15%)
Min. age	27.85	31.01
Max. age	71.47	63.8
Mean age	48.58	46.18

4.9.3. Gene expression signatures classification

Recreated classifications closely approximated the original results, considering minor differences regarding the sample cohorts (Appendix Tables B1-B3). Following that, gene expression data were matched to the available imaging data for each patient. The recreated methods were then utilized to classify each tumor in the discovery cohort. Classifications are shown in Table 4.4. Further details regarding the recreated analysis are available in Appendix B.

Table 4.4. Molecular profiles in the discovery cohort.

Gene signature	Distribution rates (n = 100)		
MammaPrint	13 (low risk)		87 (high risk)
p53 score	37 wildtype)		63 (mutant)
PAM 50 ROR-S	32 (low risk)	31 (int. risk)	37 (high risk)

4.9.4. Delta radiomic feature extraction

Four stable groups of correlated features were determined by consensus clustering. Selecting the PCs totaling greater than 85% explained variance from each group, a total of six principal components were identified to summarize change in heterogeneity for each primary lesion.

4.9.5. Imaging phenotypes of early change in tumor heterogeneity

Two radiomic phenotypes of early change in intratumor heterogeneity were identified using unsupervised hierarchical clustering and shown to be statistically significant using the SigClust method ($p<0.01$). Comparing the average of the six radiomic PC values observed for tumors in each phenotype allowed the two phenotypes to be interpreted as decreasing (Phenotype 1, n=58) and increasing (Phenotype 2, n=42) intratumor heterogeneity from T1 to T2 (Figure 4.2). A Bonferroni statistical correction resulted in a p-value of less than or equal to 0.007 to signify statistical significance in clinical covariate distribution across phenotypes. The number of future recurrences was

significantly different across phenotypes ($p < 0.001$), with proportionally more recurrences in Phenotype 2 (increasing heterogeneity) than Phenotype 1 (decreasing heterogeneity), via the Chi-square test. Other clinical and histopathologic covariates, and molecular signatures, were not significantly different across phenotypes (Figure 4.3). Additionally, neoadjuvant treatment paradigms and targeted treatment paradigms were not significantly associated with radiomic phenotypes. Kaplan Meier RFS curves were also significantly different between phenotypes ($p < 0.001$).

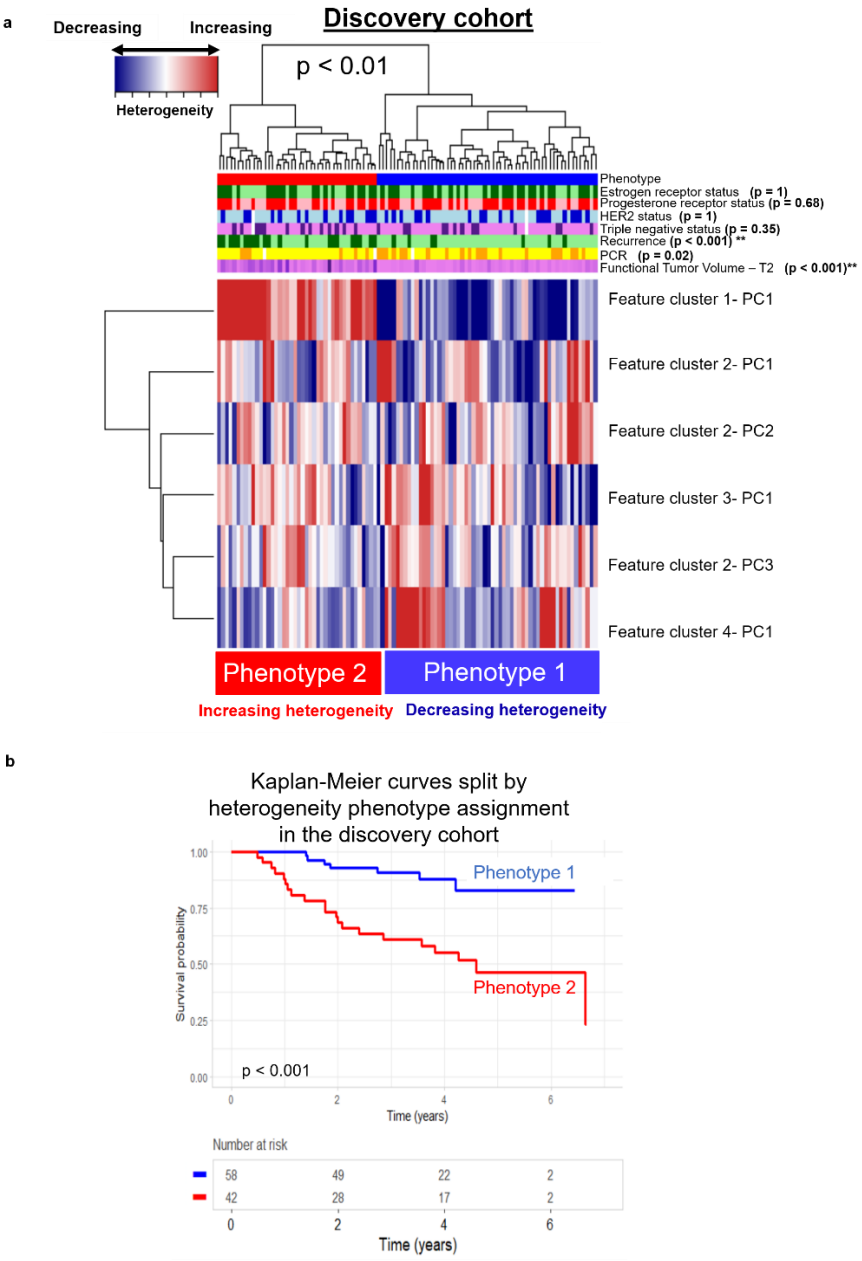


Figure 4.2. Unsupervised hierarchical clustering of tumors in the discovery cohort identified two phenotypes of early changes in intratumor heterogeneity: decreasing heterogeneity from T1 to T2 (Phenotype 1, in blue) and increasing heterogeneity from T1 to T2 (Phenotype 2, in red) (A). Kaplan-Meier curves for RFS of patient groups split by phenotype show significant separation, with tumors showing increase in intratumor heterogeneity after initiation of neoadjuvant therapy (Phenotype 2) having worse recurrence outcomes (B).

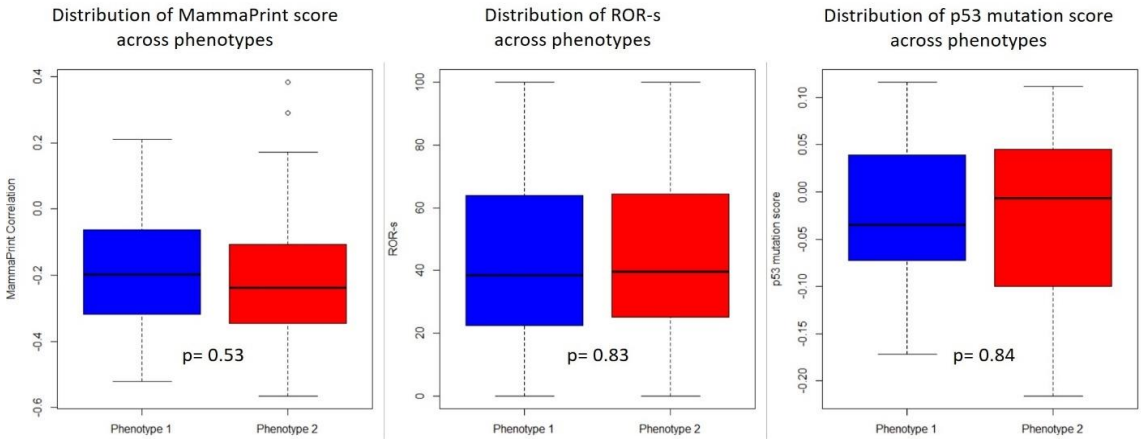


Figure 4.3. Distribution of molecular profiling scores across phenotypes. Significant distributions across the phenotypes were not seen in the three molecular profiling scores.

Within molecular subtypes of breast cancer, splitting women by radiomic phenotype assignment showed no significant difference in RFS for the HR+/HER2- subgroup (p = 0.3) and significant differences within the HER2+ and Triple Negative subtypes (both p = 0.02) (Figure 4.4).

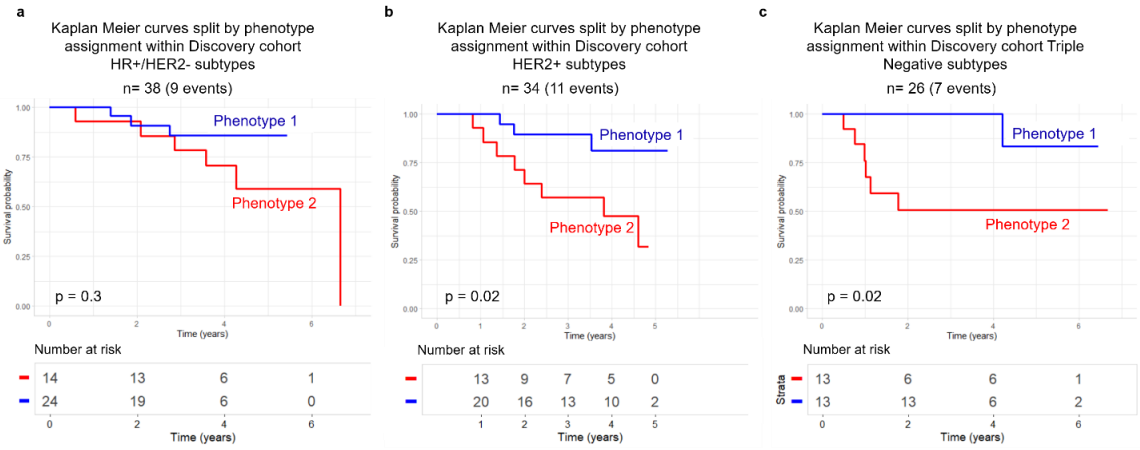


Figure 4.4. Kaplan Meier RFS curves split by phenotype assignment for (A) HR+/HER2-, (B) HER2+ and (C) Triple Negative molecular subtypes of breast cancer in the discovery cohort.

Kaplan Meier RFS curve separation by phenotype, within strata of FTV_2 value (less than/greater than median FTV_2) also demonstrated significant differences; curve separation on FTV_2 itself was not significant (Figure 4.5).

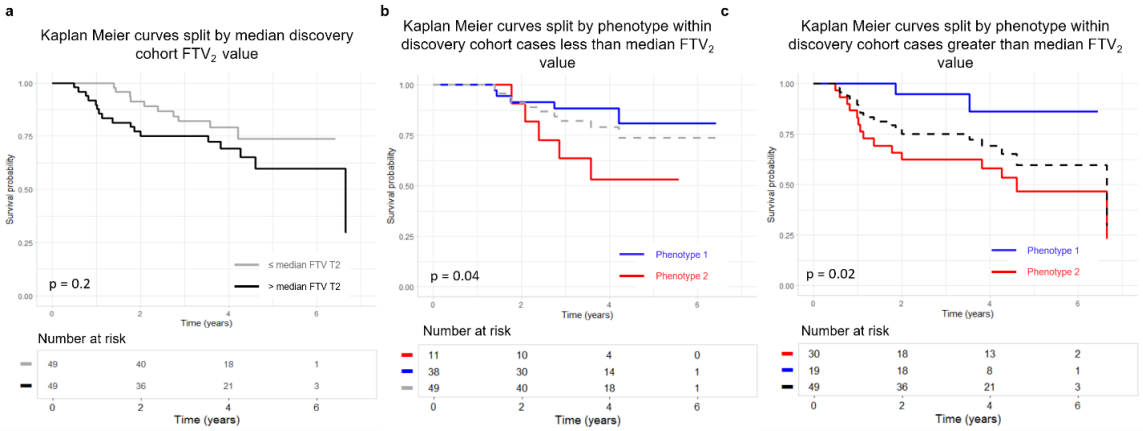


Figure 4.5. Kaplan Meier RFS curves for the discovery cohort split by median FTV2 value (A) versus split by phenotype within strata of less than median FTV_2 (B) and greater than median FTV_2 (C). RFS split by above/below median FTV_2 does not show $p < 0.05$ for separation. Within each stratum of FTV_2, the split on phenotype is significant (B and C).

Kaplan-Meier curve separation of tumors split on the median risk score generated from a Cox proportional hazards model using baseline model covariates (age, HR status, and HER2 status) and FTV_2 was significant (p=0.04). Within the low-risk tumors, further separation on phenotype demonstrated no significant curve differences. For high-risk tumors, separation by phenotype was significant ($p < 0.01$) (Figure 4.6).

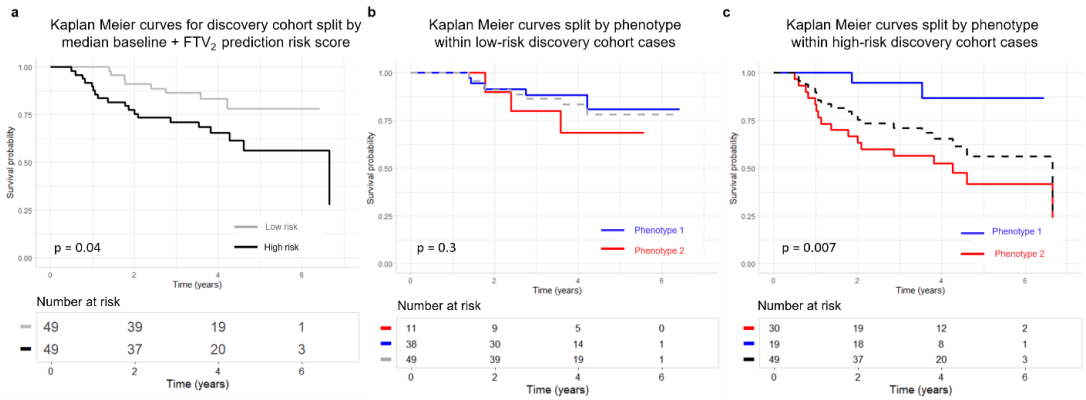


Figure 4.6. Survival versus risk score for the discovery cohort calculated by a Cox model using baseline model covariates (age, HR status, and HER2 status) and FTV_2. Split on above versus below median risk (A). Split on phenotype within the low-risk stratum (B). Split on phenotype within the high-risk stratum (C).

Univariable Cox regression models based on each of MammaPrint, ROR-S, and p53 scores resulted in c-statistics of 0.64, 0.64, and 0.66, respectively. Kaplan Meier survival curves for MammaPrint, ROR-S, and p53 scores were not significant (Figure 4.7).

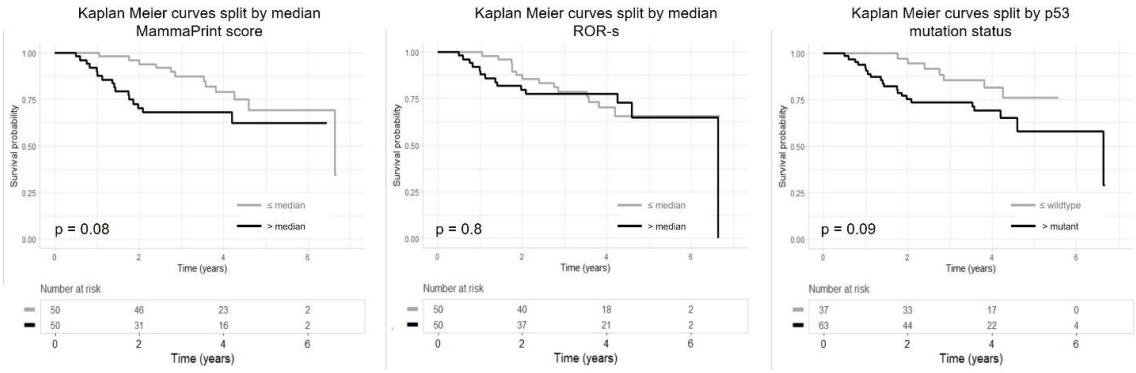


Figure 4.7. Kaplan Meier survival curve for molecular profiling scores.

A baseline model (model 1) based on age, HR status and HER2 status resulted in a cross validated, averaged over 100 replicates, c-statistic of 0.55. Adding FTV_2 to the baseline model (model 2) improved the c-statistic to 0.67 and adding molecular signatures to the baseline and FTV_2 model (model 3) resulted in a c-statistic of 0.61. A

model of baseline, FTV_2, and radiomic phenotype assignment (model 4) resulted in a c-statistic of 0.73 and a combined model of baseline, FTV_2, molecular profile scores, and radiomic phenotype assignment (model 5) demonstrated improved discriminatory capacity with a c-statistic of 0.79. The improvement in the final combined model was significant compared to the baseline, FTV_2, and molecular signature score model, as determined by the log-likelihood test ($p<0.01$) (Table 4.5).

Table 4.5. Univariable and multivariable Cox models of RFS within the discovery cohort.

Model	c-statistic	95% CI for c-statistic	Model p[1]	p versus nested model
MammaPrint	0.61	0.59-0.63	0.2	
ROR-S	0.64	0.63-066	0.1	
p53 score	0.58	0.56-0.60	0.06	
Model 1: Baseline (age, HR status, HER2 status)	0.55	0.55-0.56	0.7	
Model 2: Baseline, FTV_2	0.67	0.66-0.68	0.06	0.005[2]
Model 3: Baseline, FTV_2, molecular signatures	0.61	0.59-0.62	<0.05	0.13[3]
Model 4: Baseline, FTV_2, phenotype	0.73	0.72-0.74	< 0.01	0.01[3]
Model 5: Baseline, FTV_2, molecular signatures, phenotype	0.79	0.78-0.81	<0.001	0.002[4]

1. p versus null model of equal hazard for all patients.

2. p versus Model 1, log-likelihood test.

3. p versus Model 2, log-likelihood test.

4. p versus Model 3, log-likelihood test.

Confusion matrices for associations between molecular profile scores and radiomic phenotypes and RFS event/censor were generated. Overall positive predictive values (PPV) and negative predictive values (NPV) for MammaPrint, ROR-S, p53

mutation status and radiomic phenotype assignment demonstrated that radiomic phenotype status had the highest PPV and NPV out of the four models (Figure 4.8).

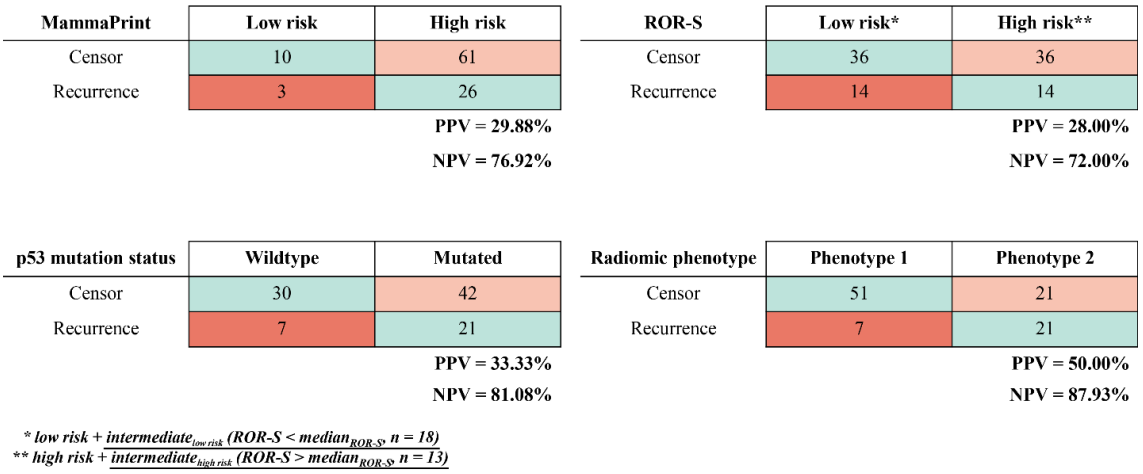


Figure 4.8. Confusion matrices for RFS prediction models within the discovery cohort using MammaPrint score (A), ROR-S (B), p53 mutation status (C), and radiomic phenotype (D).

The two radiomic phenotypes identified in the discovery set were replicated in the validation cohort and found to be statistically significant via the SigClust method ($p=0.04$). Kaplan-Meier curves of tumors in the validation cohort split by phenotype also had a statistically significant difference ($p<0.01$). A Bonferroni statistical correction resulted in a p-value of 0.008 to mean statistical significance in clinical covariate distribution across phenotypes. The proportional number of recurrences was significantly different across phenotypes ($p=0.004$) using the Chi-square test, with Phenotype 2 (increasing heterogeneity) having proportionally more recurrence evens than Phenotype 1 (decreasing heterogeneity) (Figure 4.9).

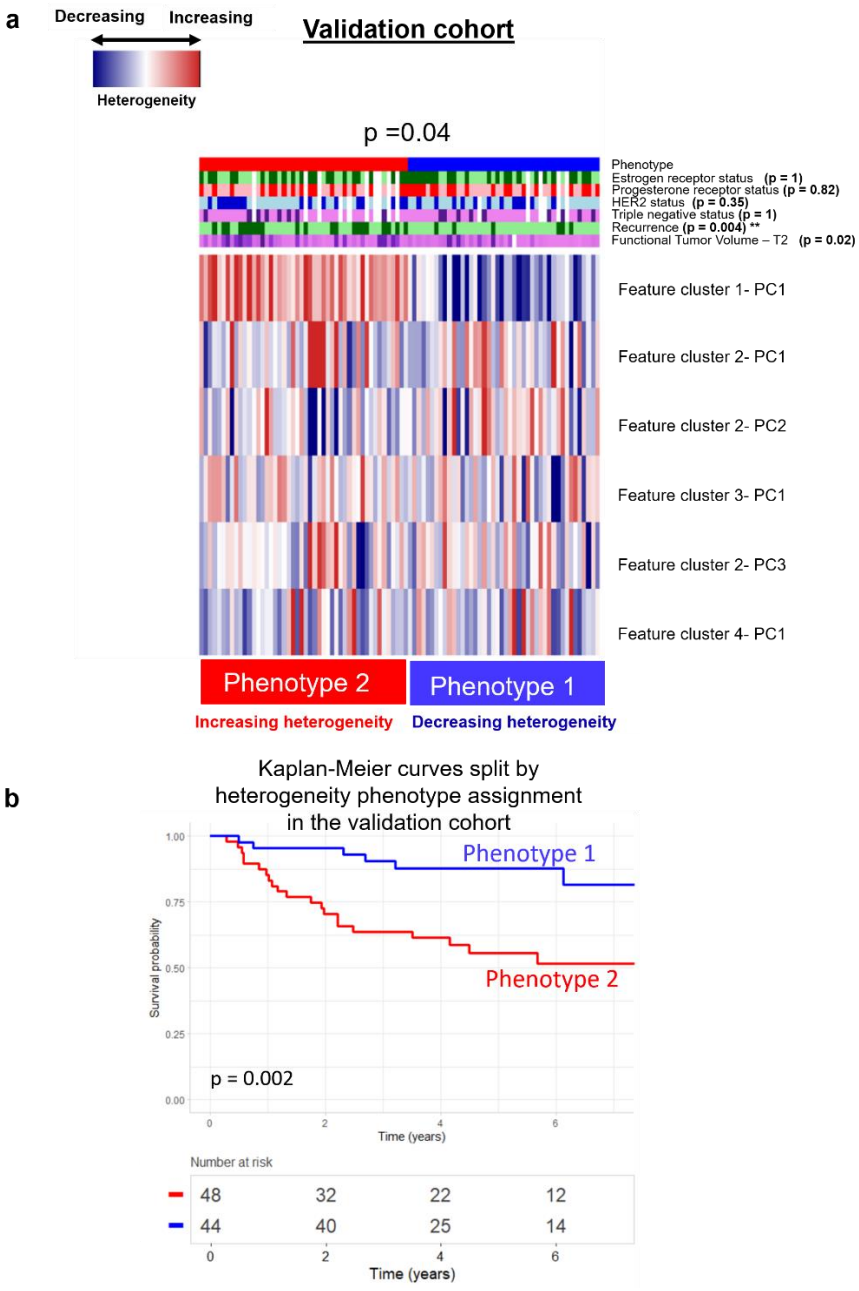


Figure 4.9. Replication of radiomics phenotypes in the validation cohort found to be significant (p=0.04) (A). Kaplan-Meier curves for RFS split on radiomic phenotype show significant separation (p = 0.002) (B).

4.10. Discussion

Two intrinsic radiomic phenotypes of early change in intratumor heterogeneity in response to neoadjuvant chemotherapy for locally advanced breast cancer were identified and validated. Interpretation of the two radiomic phenotypes as capturing an increase and decrease in intratumor heterogeneity from pre-treatment to early-treatment showed that tumors assigned to the phenotype with increasing intratumor heterogeneity had a greater number of future recurrences. This was further supported by significant separation in Kaplan Meier curves when stratifying women by phenotype assignment. Additionally, the stratification of women within FTV subgroups by phenotype demonstrates the added value of radiomic analysis in modeling prognosis (Figure 4.5). Augmenting established clinical and histopathological prognostic factors with molecular signature scores and radiomic phenotypes resulted in better prediction of RFS. This suggests that leveraging the complementary information provided by genomic and radiomic data can allow for a more comprehensive assessment of tumors and personalized therapy selection.

There may be certain plausible explanations for our observations. By capturing changes in kinetic maps of the DCE-MRI data, the identified phenotypes could reflect changes in tumor composition and angiogenic properties in response to neoadjuvant chemotherapy. Increased heterogeneity may in turn reflect tumor plasticity, which can lead to acquired resistance. The imaging phenotype demonstrating increased heterogeneity from baseline to early-treatment exhibited an increased number of recurrence events, thus supporting the hypothesis that more heterogeneous tumors may result in more adverse clinical outcomes. In contrast, the radiomic phenotype

demonstrating decreasing heterogeneity from pre-treatment to early-treatment included a higher number of tumors achieving pCR, which may suggest a relationship between decreased intratumor heterogeneity and an improved response to neoadjuvant chemotherapy (Figure 4.2).

Interpreting the radiomic phenotypes of change in tumor heterogeneity through the lens of tumor biology may provide further insight into the biologic changes occurring within the tumor in response to neoadjuvant chemotherapy. As an example, two representative tumors from women with similar age, receptor status, FTV_2 values, and ROR-S scores indicating low risk of recurrence were assigned to separate imaging phenotypes based on their early change in heterogeneity. The tumor assigned to Phenotype 2, with an increase in intratumor heterogeneity after initiation of treatment, actually had a future event of recurrence while the tumor assigned to Phenotype 1, having a decrease in intratumor heterogeneity, did not have a future event of recurrence (Figure 4.10). For these two representative cases, both women were of similar age with similar histopathologic status (HR+/HER2-). While ROR-S scores for both women characterized their tumors as “low risk of recurrence”, the women had different MammaPrint and p53 status and were assigned to separate phenotypes based on their early change in their intratumor heterogeneity. In this particular example, the woman classified as “high risk” by MammaPrint score, went on to have no future recurrence event, while the woman classified as “low risk” did have a future event of recurrence. As MammaPrint status and p53 status assignments were not significantly associated with phenotype assignment across the cohort (Figure 4.3), this suggests that the complementary information provided

by radiomic and genomic analysis could allow for increased confidence in treatment planning and clinical decision making. Furthermore, examining the principal component feature values for each woman suggests that quantitative imaging characterizations could reflect differences in these two tumors that may predict future outcomes. Of the six principal components used to cluster all tumors into the two phenotypes, C1-PC1, C4-PC1, and C3-PC1 distributions were found to be statistically significant tested against a p-value of 0.05 using Significance Analysis of Microarrays Test[198]. Examining the delta radiomics features comprising each feature cluster from which the principal components were generated could provide more insight into the specific quantitative differences in tumors in each phenotype. Specifically, as all radiomic features were extracted from the voxel-wise kinetic images, they provide a quantitative characterization of tumor angiogenesis and perfusion related properties. C1-PC1 consists largely of features characterizing changes in tumor morphology across all kinetic images, including ellipse diameter and sphericity. In the representative images, the tumor assigned to Phenotype 2 has a greater value of this feature, suggesting that it had an increase in ellipse diameter and more irregular volume moving from T1 to T2. C4-PC1 consists of features characterizing changes in mean contrast intensity, specifically from the WOS image. As this image quantifies the rate of "wash-out" of contrast agent, the representative image in Phenotype 2 may have an increase in tumor wash-out from T1 to T2, suggesting an increase in leaky vasculature due to increased angiogenesis, a characteristic of more aggressive tumors[199]. Lastly, C3-PC1 consists of features summarizing morphologic flatness across all four kinetic images. Both representative tumors have similar values for

this feature suggesting that both tumors decreased in morphologic flatness from T1 to T2 (Table 4.6).

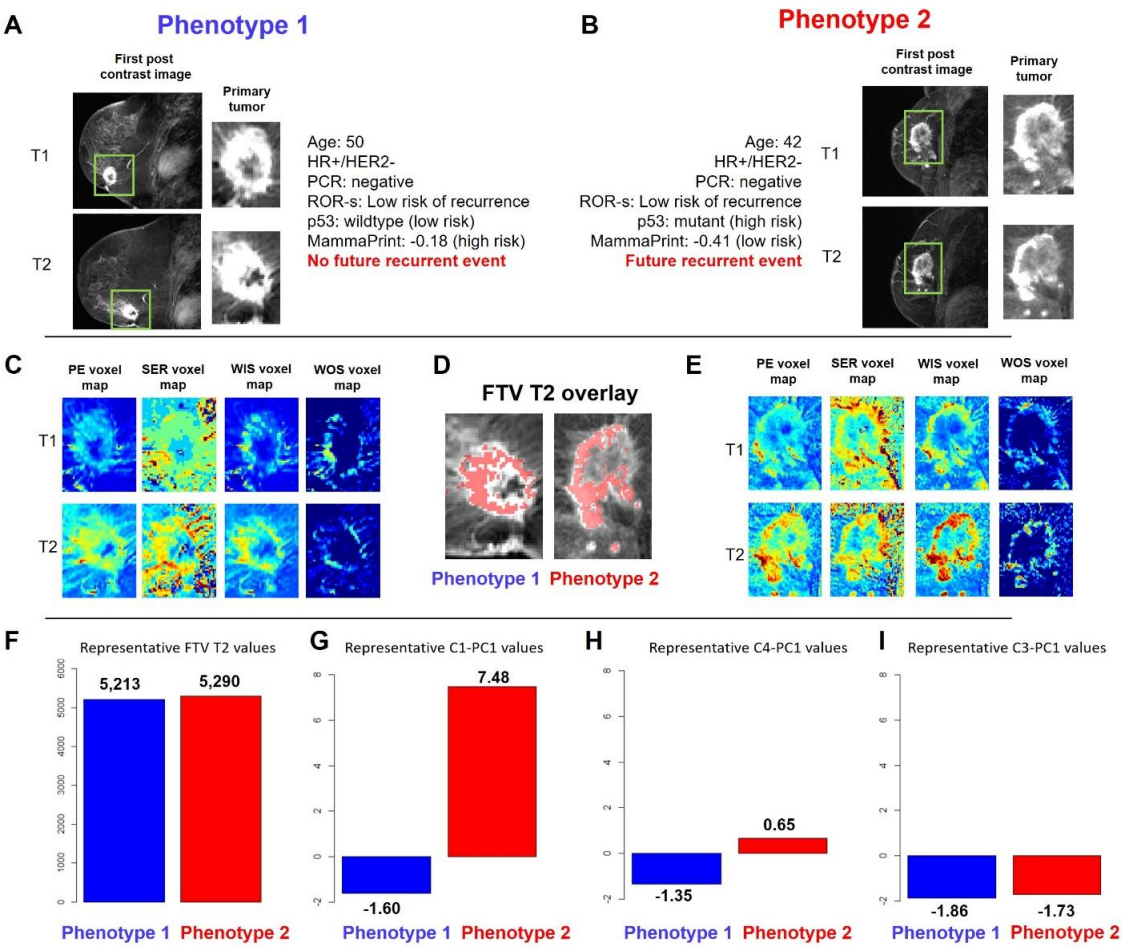


Figure 4.10. Representative tumors from Phenotype 1 (early decrease in intratumor heterogeneity) and. Phenotype 2 (early increase in intratumor heterogeneity) shown in DCE-MRI scans at T1 and T2 from the discovery cohort. (A) Representative 2D DCE-MRI slice and tumor region for T1 and T2 images from a woman, age 50, with an HR+/HER2-, ROR-S low risk, p53 wildtype (low risk), and MammaPrint score of -0.18 (high risk) tumor with no pCR and no future event of recurrence assigned to Phenotype 1. (B) Representative 2D DCE-MRI slice and tumor region for T1 and T2 images from a woman aged 42, with an HR+/HER2-, low risk ROR-S, p53 mutant (high risk), and MammaPrint score of -0.41 (low risk) tumor with no pCR and a future event of recurrence assigned to Phenotype 2. (C) Representative 2D images of PE, SER, WIS, and WOS voxel-wise maps for T1 and T2 for the tumor in phenotype 1. (D) FTV at T2 overlay for these representative tumors from phenotype 1 and 2. (E) Representative 2D images of PE, SER, WIS, and WOS voxel-wise maps for T1 and T2 for the tumor in phenotype 2. (F) FTV at T2 values for each representative tumor. Values for features (G) C1-PC1, (H) C4-PC1, and (I) C3-PC1, for each representative tumor. These representative cases provide an example where imaging characterizations of changes in each tumor's heterogeneity provided a stratification related to future outcomes. In this example, established clinical covariates did not provide such stratification.

Table 4.6. Radiomic features comprising significant feature cluster principal components.

Cluster 1- PC1	**Cluster 4- PC1**	**Cluster 3- PC1**
PE Intensity Quartile Coefficient of Variation	WOS Histogram Ninetieth Percentile	PE Morphologic Flatness
PE Morphologic Ellipse Diameter Axis 0	WOS Histogram Root Mean Square	SER Intensity Interquartile Range
PE Morphologic Ellipse Diameter Axis 1	WOS GLCM Entropy	SER Morphologic Flatness
PE Morphologic Ellipse Diameter Axis 2	WOS GLRM Short Run High Grey Level Emphasis	WIS Morphologic Flatness
PE Morphologic Equivalent Spherical Radius	WOS GLSZM Grey Level Mean	WOS Morphologic Flatness
SER Morphologic Ellipse Diameter Axis 0		
SER Morphologic Ellipse Diameter Axis 1		
SER Morphologic Ellipse Diameter Axis 2		
SER Morphologic Equivalent Spherical Radius		
WIS Intensity Quartile Coefficient of Variation		
WIS Morphologic Ellipse Diameter Axis 0		
WIS Morphologic Ellipse Diameter Axis 1		
WIS Morphologic Ellipse Diameter Axis 2		
WIS Morphologic Equivalent Spherical Radius		
WOS Morphologic Ellipse Diameter Axis 0		
WOS Morphologic Ellipse Diameter Axis 1		
WOS Morphologic Ellipse Diameter Axis 2		
WOS Morphologic Equivalent Spherical Radius		
WOS GLSZM Zone Size Nonuniformity		

Significant separation of women by radiomic phenotype assignment by Kaplan-Meier curves for women with HER2+ and triple-negative breast cancers may further highlight the known sub-clonal diversity within these subtypes (Figure 4.4)[200-202]. Our findings suggest that tumors within these subgroups that become more heterogeneous as an early

response to neoadjuvant chemotherapy may be more aggressive, resulting in increased likelihood of recurrence.

Confusion matrices for RFS prediction using molecular signatures and radiomic phenotype assignment demonstrate a greater PPV and NPV when using radiomic phenotypes (Figure 4.8). However, a limitation of using only radiomic phenotypes can be seen when comparing the predictive value of radiomic phenotypes alone against the MammaPrint assay. Seven women in Phenotype 1 went on to have recurrence despite decreasing heterogeneity on imaging whereas only 2 women, identified as a MammaPrint "low risk", had a recurrence. Leveraging the complementary information from both personalized molecular signatures and incorporating longitudinal data about tumor heterogeneity resulted in the most accurate predictive model in our study.

Limitations to our study should be noted. First, our exploratory analysis included a relatively small sample size, as we restricted it to publicly available data from the ACRIN 6657/I-SPY 1 trial with both DCE-MRI and gene expression data available. In addition, the validation cohort utilized for this study did not include gene expression data which prevented us from validating the prognostic benefit of the molecular profiling scores. The publicly available microarray data used to generate the molecular profiling scores was also limited by older acquisition protocols and technology. Additionally, image analysis was limited by the older image acquisition protocol and technology used in this study. However, the scan duration used for the dataset deriving from the I-SPY 1 trial was 4.5 minutes, which is similar to the current American College of Radiology (ACR) recommendation of <= 4 minutes[203]. Moreover, the datasets used in this study for

discovery and validation are among the only publicly available datasets with true long-term follow up available following NAC. Ultimately, given the encouraging, exploratory, and proof of concept results with these older MRI protocols, we can hypothesize that the performance of the proposed radiomic features may be better with newer MRI protocols. Future work will include expanding our analysis to larger cohort sizes with images acquired with newer, more clinically utilized MRI acquisition protocols, as well as exploring relationships between early changes in tumor heterogeneity via radiomic phenotyping and differentially expressed genes with related molecular pathways. Additionally, utilization of Next Generation Sequencing (NGS) techniques which, in contrast with microarrays, do not depend on specific probes for the quantification of the expression of pre-specified genes, will allow for deeper analyses.

In conclusion, our exploratory results demonstrate that early changes in intratumor heterogeneity in response to neoadjuvant chemotherapy as captured by radiomic analysis of DCE-MRI may provide improved prediction of RFS for locally advanced breast cancer. Longitudinal non-invasive assessment of tumor phenotypes via imaging may allow for monitoring of heterogeneity and underlying tumor biology. Augmenting clinical, histopathologic, and molecular covariates with imaging phenotypes may allow for personalized risk stratification and early adaptation of treatment strategies.

Chapter 5 : Developing a 4-D segmentation method to characterize intratumor heterogeneity

5.1. Introduction

Given the current hypotheses of linear and branched evolution for breast tumor progression, sub-clonal populations are thought to be spatially contiguous regions of shared physiologic function. Partial volume effects introduced during medical image reconstructions result in an image blurring effect, caused by the finite spatial resolution of an imaging system following a 3D convolutional operation of the source with a point spread function[204]. Additionally, voxel sampling as a 3-D grid suggests that neighboring voxels may share underlying tissue structure, or that a single voxel may summarize the signal of multiple tissue structures, a phenomenon known as the "tissue fraction effect"[204]. As such, it is hypothesized that imaging representations of intratumor heterogeneity would result in spatially contiguous voxels sharing similar physiological behavior.

Conventional radiomic features largely aim to quantify the spatial distribution of voxel intensities[28]. While informative when paired with imaging modalities with high spatial resolutions, some imaging modalities, such as PET, may be limited in spatial resolution. Instead, novel methods to characterize intratumor heterogeneity by leveraging the molecular specificity provided by PET imaging should be leveraged.

Preliminary work suggests that established unsupervised parcellation techniques[205] are inadequate for identifying spatially constrained, functionally similar sub-regions. Such methods (e.g. k-means clustering, hierarchical clustering) assume that dynamic data extracted from each voxel is independent from its surrounding voxels. While many segmentation techniques exist, a novel paradigm is required to perform an unsupervised segmentation of spatially constrained, non-rigid, and arbitrarily shaped regions with discrete functional behavior. Other established segmentation techniques are designed for segmenting well defined anatomical structures or require manual initialization or well-defined control points[206-208]. Probability based clustering methods[209] require an user-defined threshold for segmentation and therefore may not account for partial volume effects[204]. We hypothesize that a segmentation method incorporating both spatial and temporal image information would have improved segmentation accuracy over established methods utilizing only temporal dynamics or spatial information alone. Incorporation of spatial constraints to a segmentation method may account for tissue fraction and partial volume effect

We have developed a method to characterize 4-D functional tumor heterogeneity (FTH) by aiming to capture aspects of both spatial and kinetic tumor heterogeneity as seen in dynamic imaging. Identifying functionally discrete sub-regions within a single tumor region may allow for characterizing the extent of intratumor heterogeneity and may lead to studies monitoring sub-region growth and response to treatment. Leveraging the improved dynamic sampling and molecular specificity afforded by dynamic PET imaging may allow for non-invasive, novel prognostic and predictive markers to

characterize tumor molecular biology. The developed method is agnostic to the specific radiotracer utilized and does not depend on complex kinetic modeling assumptions. Instead, the approach is data driven in terms of identifying intrinsic 4-D patterns of molecular tumor heterogeneity.

5.2. Simulated image phantoms

Algorithm development and training was carried out using simulated 4-D PET data. With the goal of developing a method with broad applicability, agnostic to radiotracer or disease site, we chose to perform these algorithm trainings using a tracer with broadly similar tracer kinetics to fluorodeoxyglucose (^{18}F-FDG) (FDG), a commonly used radiotracer in cancer staging. Simulated dynamic PET images based on data from fluorothymidine (^{18}F-FLT) (FLT) PET[210] were utilized for method development and validation. FLT was chosen as a representative tracer with kinetics similar to other trapped cancer-relevant radiotracers to assess the generalizability of our method. All simulations were done using Geant4 Application for Tomographic Emission (GATE)[211]. The scanner simulations were based on the PennPET Explorer[212] with 70-cm axial field of view. Based on prior human studies of FLT data from patients with lung cancer[210], kinetic parameters were selected to emulate low, medium, and high K_{FLT} lesions (ml/cm3/sec) and the blood input curve was derived from an FLT PET patient dataset and fit to a tri-exponential model. Details on simulated image generation have been previously described[213]. The simulated images were cropped to a region-of-interest (ROI) comprising the simulated regions and surround background area to a total size of 64x69x9 voxels x 45 frames. The simulated images consisted of two regions modeling

low tracer uptake (10mm and 13mm sphere diameter), two regions modeling medium tracer uptake (10mm and 13mm sphere diameter), two regions modeling high tracer uptake (10mm and 13mm sphere diameter), a blood region, and a background region (Figure 5.1).

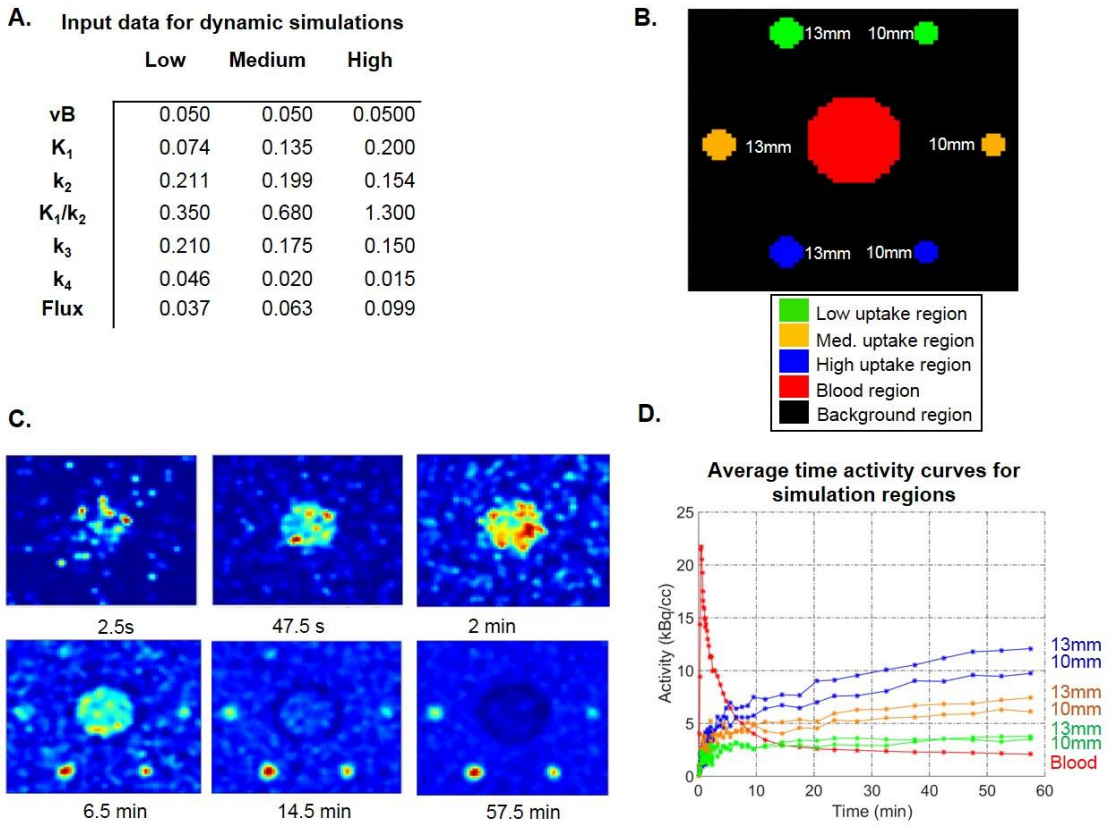


Figure 5.1. Dynamic FLT-PET simulation images used for method development and validation. (A) Input data used to generate dynamic simulations. (B) Labels for all regions in simulated image. (C) Center slice images from representative simulation frames. (D) Average TACs for simulation sphere and blood curve regions.

5.3. Limitations of existing methods

Unsupervised segmentation analysis allows for the intrinsic grouping of 4-D functional data without prior knowledge of voxel labels. Mapping histopathologic ground

truth to imaging representations of sub-clonal regions is often limited by the non-isotropic resolution of voxels and tissue deformation that can occur between resection and fixation for histology analysis[7,21].

Developing a method to identify functionally discrete sub-regions began by evaluating the performance of established unsupervised clustering algorithms for time activity clustering and by adding spatial constraints to these established methods. Use of the simulated image phantoms allowed for evaluating method performance when segmenting spatially constrained, functionally discrete sub-regions. Initial method performance was evaluated using visual inspection of the resulting segmentation of simulated spheres image, allowing for iterative improvements to be made.

The first evaluated unsupervised technique was hierarchical clustering, grouping together voxels within the image based on the Euclidean distance between time activity curves. Ward's method for minimum variance was used as the clustering criteria. Qualitative assessment of the resulting segmentation suggested a need for adding a spatial weight to ensure that minimum variations in time activity behavior within a region, causing neighboring voxels to be classified with different labels, could be offset by spatial proximity.

A spatially weighted clustering method for hierarchical clustering was then implemented to promote within cluster minimization based on voxel spatial location and functional behavior, as seen in Eq. 5.1.

$$within\ Cluster_k\ minimization = (1-\alpha)\sum_i \sum_j\ d^2_{f\{i,j\}} + \alpha \sum_i \sum_j\ d^2_{s\{i,j\}} \qquad \text{Eq. 5.1}$$

Here, d_f represents the within cluster Euclidean distance between time activity curves between voxels *i* and *j*. Similarly, d_s represents the within-cluster Euclidean distance between spatial coordinates for voxels *i* and *j*. The term α is used to control the degree of influence either functional similarity or spatial similarity has on the minimization. Use of this minimization function was implemented with hierarchical clustering and assessed qualitatively.

Next, dynamic time warping (DTW) was used instead of Euclidean distance, to identify alignment between voxels' time activity curve behaviors[214]. DTW is a technique used to find the optimal alignment between two time-dependent sequences that have been sampled at equidistant points in time. Briefly, the pairwise distance between time series, X and Y, are first calculated. Within this, the algorithm aims to find and alignment path through low-cost areas of the distance matrix[215]. The resulting segmentation performance of using hierarchical clustering with DTW as a distance metric in the time domain was assessed qualitatively.

Spectral clustering is a clustering algorithm utilizing graph-based segmentation techniques. This algorithm calculates the distance between all voxels and then uses the top eigenvectors of this distance matrix to form groupings. Briefly, an adjacency matrix is created in which each element is the pairwise adjacency between two voxels, x_i and x_j, defined as:

$$w(x_i, x_j) = e^{-\left(\frac{d_f(x_i,x_j)}{var(d(I))}\right)}$$ Eq. 5.2

Here, d_f represents the within cluster Euclidean distance between time activity curves Between voxels x_i and x_j. From this, a degree matrix, a diagonal matrix with degrees $d_1, \ldots, d_n$ calculated along the diagonal is calculated as:

$$D_{ii} = \sum_i w_{ij}$$ Eq. 5.3

For each voxel *i*, summing over all voxels adjacent to *i*. The non-normalized Laplacian *L*, is then calculated as L= D- W. From this, the first *k* eigenvectors summarizing *L* are calculated, and then k-means clustering is used to cluster the eigenvectors into sub-groups[216]. Spectral clustering was used to segment the simulated image with voxels represented as time activity curves, and segmentation performance was qualitatively assessed.

A spatial constraint was then added to the spectral clustering adjacency matrix as previously identified, by multiplication of an exponential spatial weight:

$$w(x_i, x_j) = e^{-\left(\frac{d_f(x_i,x_j)}{var(d(I))}\right)*} \; e^{-\left(\frac{d_s(x_i,x_j)}{var(d(I))}\right)}$$ Eq. 5.4

where d_s represents the within-cluster Euclidean distance between spatial coordinates for voxels x_i and x_j. This exponential spatial weight was added to penalize clustering assignment by spatial location, such that voxels grouped together according to functional

data would also be spatially contiguous. Segmentation performance was qualitatively assessed.

Lastly, wavelet decomposition analysis was explored to reduce the dimensionality of time activity curves prior to clustering. The segmentation performance of representing voxels by their first wavelet decomposition was qualitatively assessed. Representative segmentations of the central 2-D slice of the simulated image phantoms from all evaluated techniques can be seen in Figure 5.2. These evaluated methods did not qualitatively result in successful segmentation of the simulated image.

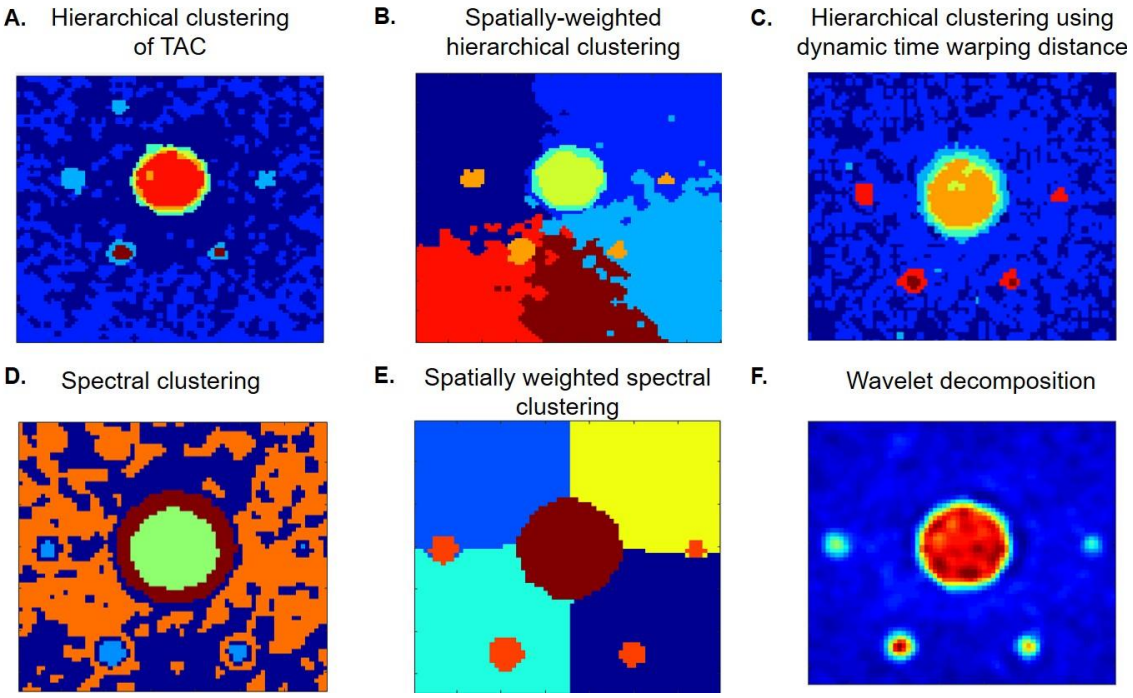


Figure 5.2. Segmentation performance for various evaluated techniques applied to the simulated phantom images. Segmentation performance was qualitatively assessed by visually comparing performance to the true segmentation labels.

Overall, these experiments suggested the need for improved representation of similarities in time activity behavior along with a spatial weight in order to segment the simulated images in 4-D.

5.4. Radiomic functional intratumor (Rad-FIT) clustering

Due to the limitations of existing unsupervised voxel parcellation methods and in order to better assess both the spatial and temporal heterogeneity seen in breast tumors, Radiomic Functional Intratumor (Rad-FIT) clustering was developed as a 4-D-segmentation method to segment tumors into spatially constrained, functionally discrete sub-regions (Table 5.1).

In characterizing intratumor heterogeneity, the assumption of voxel independence[8], upheld in many conventional unsupervised clustering studies, does not hold due to overwhelming evidence suggesting that neighboring voxels share physiologic similarity[39]. Therefore, a method incorporating a voxel's spatial neighborhood in its own label assignment was incorporated. Further, Rad-FIT clustering does not require manual initialization or defined control points. Instead, we utilized a randomly initiated k-means clustering for initialization. Compared to other clustering algorithms that determine hard, spherical or compact clusters (e.g. k-means clustering), or prevent adjustment once a cluster merge or split decision has been executed (e.g. hierarchical clustering), Rad-FIT clustering allows for an iterative, probability-based label assignment with no prior assumption about sub-region shape. Lastly, Rad-FIT clustering allows for easier interpretation of within sub-region similarities and between sub-region differences, as

compared to other graph-based segmentation approaches, due to summarizing time activity information into dominant modes of variation.

For a given voxel ***v*** within a 3-D ROI**,** the 4-D functional behavior can be represented as $v_{(x,y,z,y,t)}$ given a 3-D spatial location *x*,*y*,*z* and *t* representing the temporal signal from a set of dynamic frames. The temporal signal of an ROI over dynamic PET scans is first summarized using functional principal component analysis (FPCA). Functional principal component analysis (FPCA) is a method used to determine modes of variation in functional data. Principal component analysis (PCA) is a well-established method for dimensionality reduction and capturing dominant modes of variation in high dimensional data through a set of orthonormal vectors. Similarly, FPCA extends this idea to continuous functions in time, where the dominant modes of variation in the time curves are represented as eigenfunctions, and data points are projected into this space represented by functional principal components. Use of FPCA allows for the dimensionality reduction of time series data for each voxel while retaining the variance of its temporal information. The dynamic scan is then reduced, with each voxel represented using functional principal components (FPC) capturing greater than 85% of the variance seen in its dynamic behavior ($v_{x, y, z, f}$).

A Markov Random Field (MRF) segmentation paradigm is applied to the voxel data. Here, ignoring the independence assumption between image voxels, the prior probability of a label *z* for pixel $v_{x,y,z,f}$ is then modeled using a simple state prior model (Eq. 5.5) [217].

$$p(z_{x,y,z} = k \,|z^{l}, l \in N_{x,y,z}) = \frac{e^{\beta \delta_{x,y,z(k)}}}{\sum_{k=1}^{K} e^{\beta \delta_{x,y,z(k)}}} \quad \text{Eq. 5.5}$$

where k is the number of possible labels, β is a floating parameter controlling the influence of neighboring voxels, and $N_{x,y,z}$ defined as a 3x3x3 voxel grid surrounding $v_{x,y,z,f}$. The likelihood model of $v_{x,y,z,f}$ is modeled as a multivariate Gaussian distribution (Eq 5.6).

$$p(v_{x,y,z,f})|z_{x,y,z} = k\,, \phi) = \frac{1}{2\pi^{m\backslash 2}\ |\Sigma_k|^{\frac{1}{2}}} e^{-\frac{1}{2}(v_{x,y,z,f}-\mu_k)^T \Sigma_k^{-1}(v_{x,y,z,f}-\mu_k)} \quad \text{Eq. 5.6}$$

Initialization of mean and standard deviation values for each label k is performed using K-means clustering with a predetermined number of k labels. The posterior distribution for all labels, k, and all n pixels within the ROI is modeled by,

$$p(z^{t}_{x,y,z} = k|v_{x,y,z,f}) = \frac{p(v_{x,y,z,f}|z^{t}_{x,y,z} = k)P(z^{t}_{x,y,z} = k)}{\Sigma^{n} p(v_{x,y,z,f}|z^{t}_{x,y,z} = k)P(z^{t}_{x,y,z} = k)} \quad \text{Eq. 5.7}$$

and parameters $\theta^{(t+1)}$ are updated by the following equations (Eq 5.8, Eq 5.9):

$$\mu_k^{t+1} = \frac{\Sigma^n p\left(z_{x,y,z}^t = k \,|v_{x,y,z,f}\right) v_{x,y,z,f}}{\Sigma^n p(z_{x,y,z}^t = k \,|x_{i)}} \qquad \text{Eq. 5.8}$$

$$\left(\sigma_{kf}^2\right)^{t+1} = \frac{1}{m} \frac{\Sigma^n p\left(z_{x,y,z}^t = k \,|v_{x,y,z,f}\right)\left(v_{x,y,z,f} - \mu_{kf}^{t+1}\right)^2}{\Sigma^n p\left(z_{x,y,z}^t = k | v_{x,y,z,f}\right)}, f = 1,2,..m \qquad \text{Eq. 5.9}$$

where *j* is the number of FPCs selected. The optimization problem and parameter estimation is then solved using the expectation maximization (EM) algorithm[218].

Table 5.1. Radiomic functional intratumor (Rad-FIT) clustering

Algorithm 1 — Radiomic functional intratumor (Rad-FIT) clustering

Input: $\mathbf{V_i} \in R^{3\times t}$, β (floating parameter), k (number of sub-regions)

Output: $\mathbf{z} \in [0,1]^K$ (Clustering Assignment)

Initialization: Initialize $\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\sigma}$ by K-means clustering algorithm

Step 1: Functional principal component analysis (retain components with 85% of variance:

$$\mathbf{V_{(x,y,z,t)}} \gg \mathbf{V_{(x,y,z,f)}}$$

Loop- Repeat (t) until convergence

- Fix $\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\sigma}$ — Solve for $p\left(v_{x,y,z,f}\right)\left|z_i = k\right., \boldsymbol{\mu}, \boldsymbol{\sigma})$
- Fix $p\left(v_{x,y,z,f}\left|z^t_{x,y,z} = k\right.\right)$ —
- Solve for $z^t = \underset{z \in Z}{argmax}\left[P\left(v_{x,y,z,f}\left|z_{x,y,z}, \boldsymbol{\mu}, \boldsymbol{\sigma}^t\right.\right)P(Z)\right]$

5.5. Initialization of Rad-FIT clustering

K-means clustering was chosen as the initialization technique for Rad-FIT clustering to determine an initial parameter estimation for the conditional distribution. K-means clustering is well established as an initialization method for probability-based clustering using the EM algorithm. Use of k-means clustering as an initialization method over random initialization resulted in improved performance of Rad-FIT clustering (Figure 5.3).

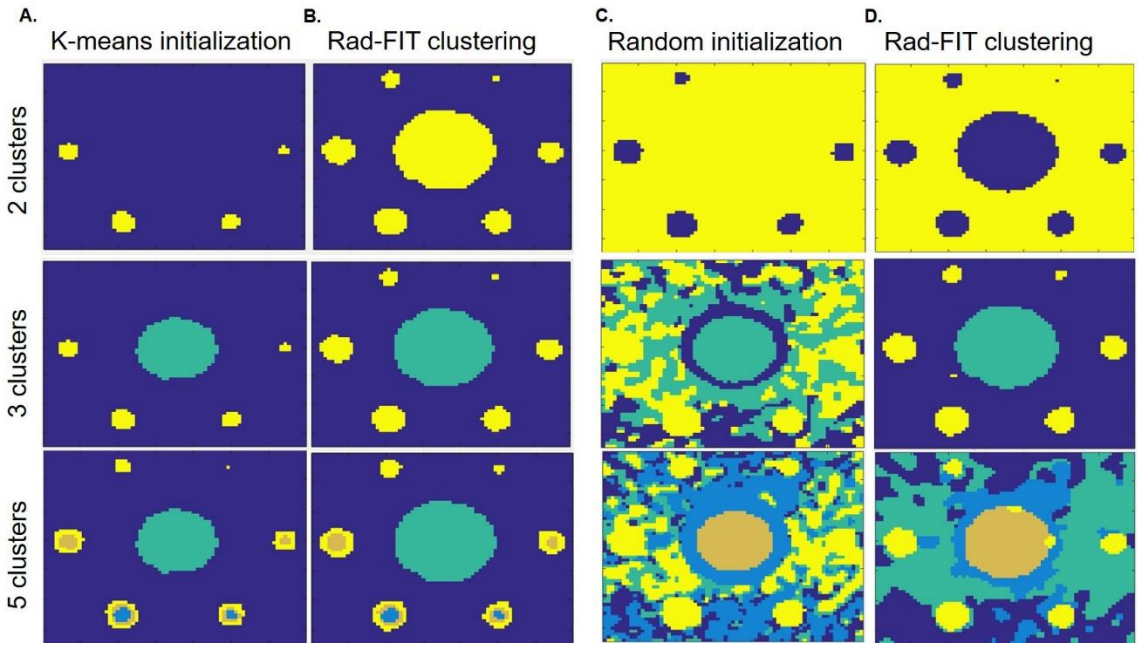


Figure 5.3. Comparison of initialization techniques. Segmentation performance of (A) K-means clustering as an initialization method for (B) Rad-FIT clustering as compared to (C) random initialization as an initialization method for (D) Rad-FIT clustering. Initialization approaches were evaluated for 2-class segmentation, 3-class segmentation, and 5-class segmentation. Varying colors represent how the image has been segmented into discrete classes.

5.6. Value of dynamic imaging versus static imaging

Previous studies have demonstrated the improved prognostic and predictive value of utilizing dynamic PET imaging over static imaging[22]. However, static imaging is more widely implemented in a clinical setting than a more conventional PET imaging protocol[219]. To determine whether temporal sampling over a 60 min dynamic imaging scan offered additional information to that from a static image, or an image with a sub-sampled number of temporal frames when identifying 4-D sub-regions, the performance of Rad-FIT clustering was evaluated on the final frame of the 60-minute scan as well as using the first, middle, and final frames over the 60-minute scan. (Figure 5.4). The

resulting segmentation performances were qualitatively assessed and demonstrated the additional information provided by leveraging the entire dynamic scan.

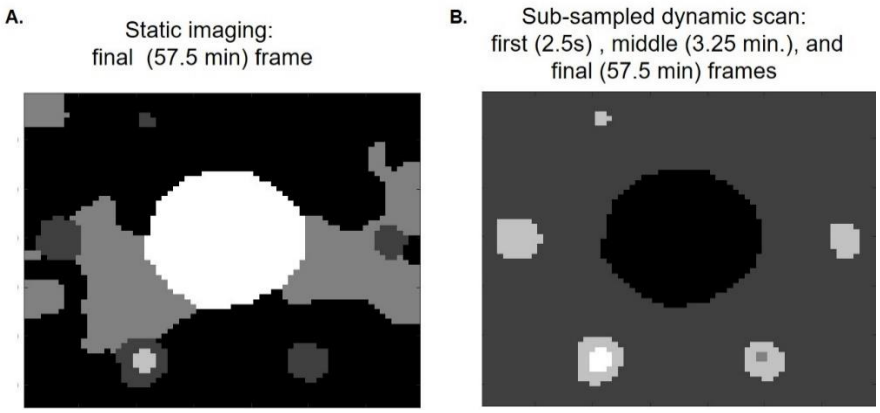


Figure 5.4. Rad-FIT clustering performance when only a static scan or a sub-sampled dynamic scan is utilized. Qualitative assessment suggests the full dynamic scan offers information beyond that of a static or sub-sampled scan required for more accurate segmentation performance. Varying gray scale colors represent different classes each image has been segmented into.

5.7. Signal to noise ratio experiments

The floating parameter β used in the MRF-based prior probability in Rad-FIT clustering controls the degree of influence a voxel's spatial neighborhood has on its own label assignment. In order to select the optimal value of β, we assessed an array of values ranging from 0 to 3 to determine which resulted in the best segmentation performance when used in Rad-FIT clustering. These values were selected empirically based on the high degree of variation in clustering performance seen when using this range of β and also based on prior studies demonstrating use of a spatial weight of similar value[220]. To model the effects of spatial noise often seen in PET images due to variations in image reconstruction techniques, we utilized images of the same simulated image phantom reconstructed with one-half and one-quarter of the counts derived from simulated

positron annihilation events to effectively model one-half or one-quarter of the radiotracer signal, resulting in images with increased signal to noise ratio (SNR). For these simulated images, the original 45 frames (full dose of 4.0 mCi (148 MBq), were subsampled to roughly correspond to emulated doses of 2.0 mCi (74 MBq) for the one-half simulation and 1.0 mCi (18.5 MBq) for the one-quarter simulation[213]. Use of the varying SNR images also allowed us to evaluate the optimal value of β that would allow for neighboring voxels of shared functionality to be grouped together independent of image noise. The value of β (β =2.6) that resulted in the highest Dice score with minimal variation across the three simulated image SNR conditions was selected as the optimal value (Figure 5.5).

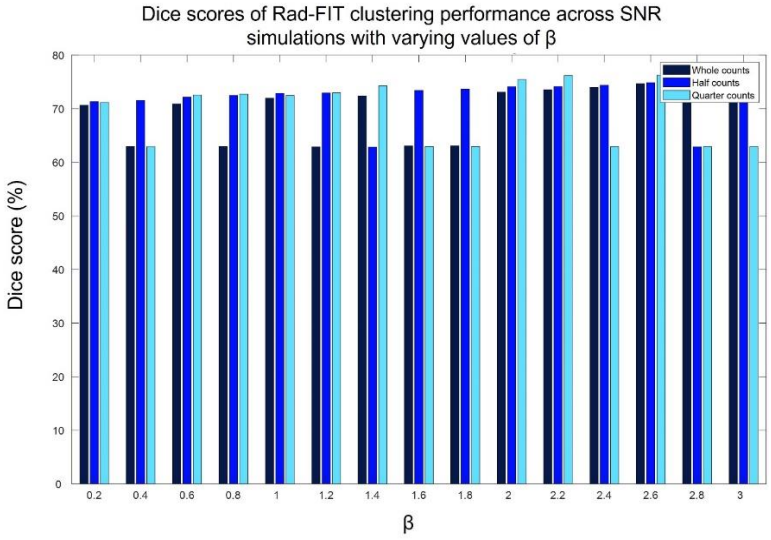


Figure 5.5. Optimal value of β across simulations of varying signal to noise ratio.

5.8. Evaluating method performance on simulated image phantoms

To evaluate the performance of segmenting spatially contiguous regions of heterogeneous tracer activity with Rad-FIT, the following segmentation assessments were performed.

First, the improved value of summarizing temporal information from dynamic voxel behavior using FPCs was evaluated. The segmentation performance of Rad-FIT when simulation voxels were represented by their original time activity curves was compared to the performance when simulation voxels were represented by their FPC scores.

The segmentation performance of Rad-FIT was then evaluated for its ability to segment the high, medium, and low uptake 13mm spheres from each sphere's surrounding background region, respectively. All segmentation performances were compared against the performance of established unsupervised segmentation algorithms including K-means clustering, hierarchical clustering, and spectral clustering[216]. K-means clustering and hierarchical clustering are well-established unsupervised clustering techniques used to find intrinsic groupings within data structures. As spectral clustering is a graph-based segmentation method relying on the functional similarity and adjacency between voxels to identify image partitions, it allowed for more equal comparisons in segmentation performance.

Average signal to background ratios for the low, medium, and high uptake spheres were 1.78, 6.95, and 11.89, respectively. When segmenting low, medium, and

high uptake spheres from their surrounding background region, segmentation performance for all segmentation methods improved when simulation voxels were represented by the FPCs capturing greater than 85% of TAC variability as opposed to using the TACs. Additionally, Rad-FIT clustering demonstrated the highest segmentation performance when segmenting low, medium, and high uptake sphere regions from its surrounding backgrounds when evaluated using the Dice score[221] (Table 5.2) and Jaccard index (Table 5.3). Both the Dice score and Jaccard index are established statistical metrics used to determine the degree of overlap between the true regions and resulting regions from the segmentation algorithm. Both Dice scores and Jaccard indices include values ranging from 0 to 1, with a value of 1 indicating perfect similarity between true and segmented regions. Segmentation performance of each algorithm on a representative center slice (2-D) for the high sphere region (Figure 5.6), medium sphere region (Figure 5.7), and low sphere region (Figure 5.8) can be seen below.

Table 5.2. Average segmentation performance over ten replicates evaluated using the Dice scores when segmenting low, medium, and high uptake simulated sphere regions from surrounding backgrounds. Standard deviation in parentheses.

		Dice scores			
Segmentation region	**Voxel representation**	**Hierarchical clustering**	**Spectral clustering**	**K-means clustering**	**Rad-FIT clustering**
Low uptake sphere	Time activity curves	0.15 (0)	0.04 (0.001)	0.13 (0.06)	0.24 (0.07)
	FPC	0.67 (0)	0.16 (0.03)	0.65 (0.02)	0.70 (0.01)
Medium uptake sphere	Time activity curves	0.77 (<0.001)	0.06 (0.01)	0.44 (0.4)	0.73(< .001)
	FPC	0.78 (0)	0.20 (0.18)	0.84 (0.03)	0.85 (0)
High uptake sphere	Time activity curves	0.72 (0)	0.18 (0.07)	0.52 (0.40)	0.66 (0)
	FPC	0.83 (0)	0.17 (0.03)	0.84 (0)	0.86 (0)

Table 5.3. Average segmentation performance over ten replicates evaluated using the Jaccard index when segmenting low, medium, and high uptake simulated sphere regions from surrounding backgrounds. Standard deviation in parentheses.

		Jaccard indices			
Segmentation region	**Voxel representation**	**Hierarchical clustering**	**Spectral clustering**	**K-means clustering**	**Rad-FIT clustering**
Low uptake sphere	Time activity curves	0.08 (0)	0.01 (0.01)	0.07 (0.03)	0.14(0.05)
	FPC	0.50 (0)	0.09 (0.03)	0.55 (0.2)	0.54 (0)
Medium uptake sphere	Time activity curves	0.64 (0)	0.015(0.02)	0.30 (0.37)	0.58 (0)
	FPC	0.65 (0)	0.12 (0.17)	0.72 (0.04)	0.74 (< .01)
High uptake sphere	Time activity curves	0.56 (0)	0.52 (0.04)	0.44 (0.36)	0.52 (0)
	FPC	0.71 (0)	0.08 (0.01)	0.73 (0)	0.75 (0)

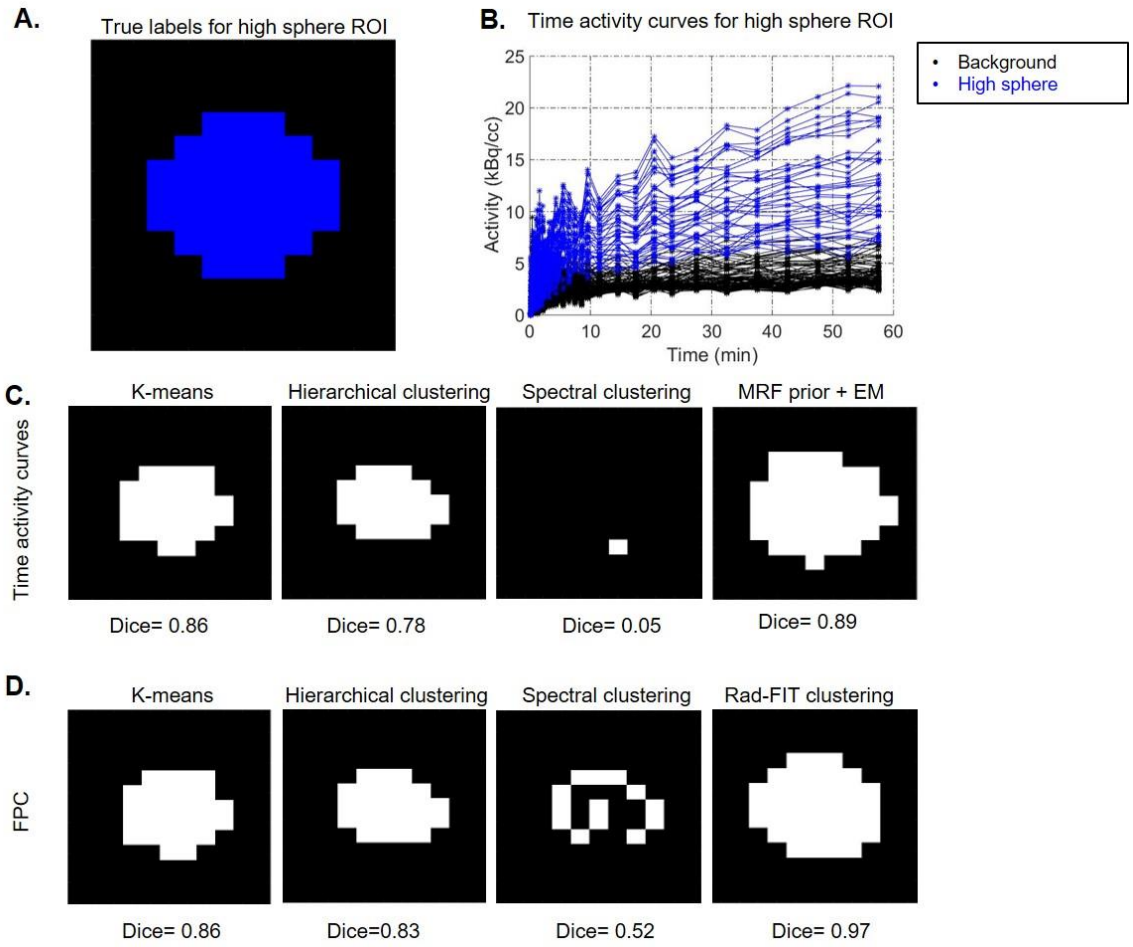


Figure 5.6. Segmentation performance of Rad-FIT clustering and established unsupervised clustering algorithms when segmenting representative center-slice (2-D) of high uptake sphere ROI from background. (A) True labels for high uptake sphere ROI. (B) Time activity curves for voxels in ROI and background. (C) Segmentation performance of algorithms when voxels are represented as time activity curves. (D) Segmentation performance when voxels are represented as FPCs.

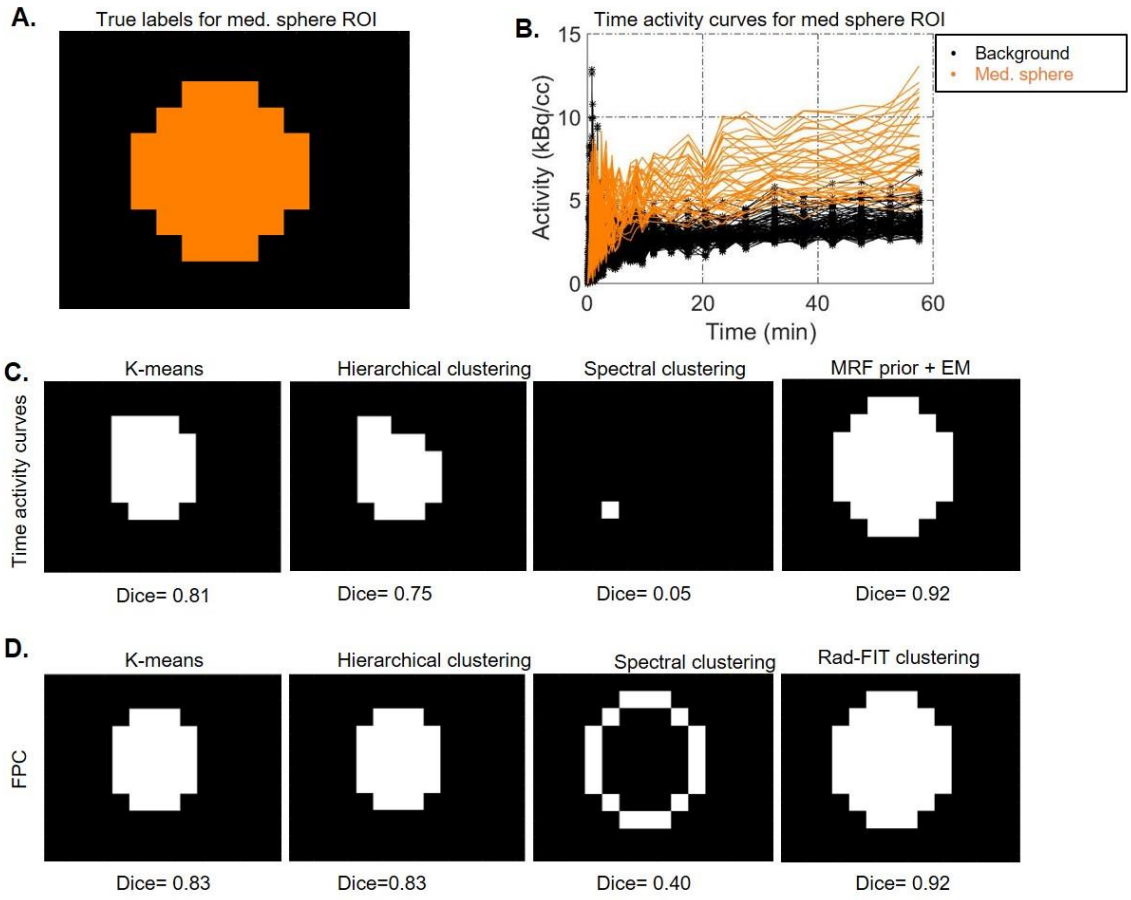


Figure 5.7. Segmentation performance of Rad-FIT clustering and established unsupervised clustering algorithms when segmenting representative center-slice (2-D) of medium uptake sphere ROI from background. (A) True labels for medium uptake sphere ROI. (B) Time activity curves for voxels in ROI and background. (C) Segmentation performance of algorithms when voxels are represented as time activity curves. (D) Segmentation performance when voxels are represented as FPCs.

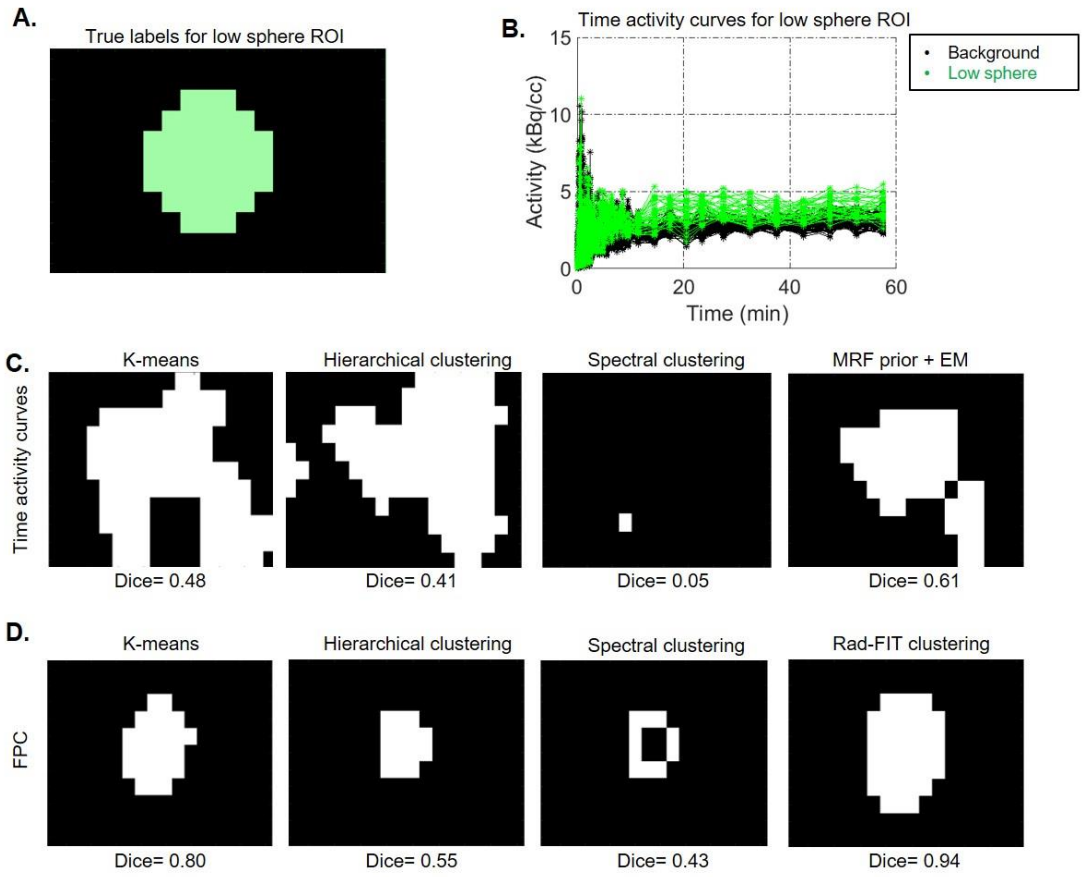
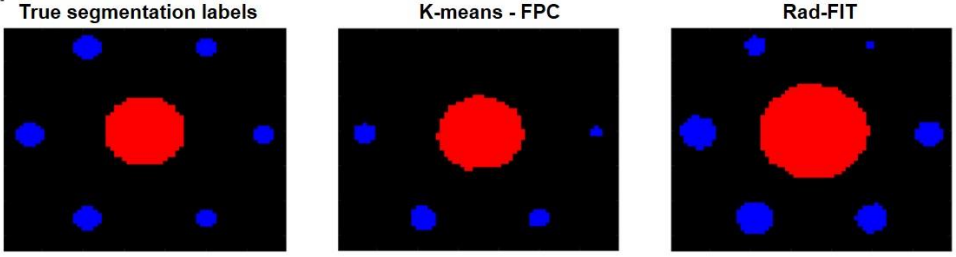


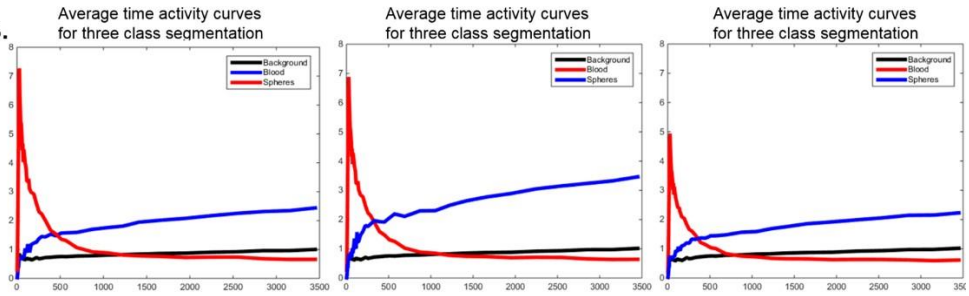
Figure 5.8. Segmentation performance of Rad-FIT clustering and established unsupervised clustering algorithms when segmenting representative center-slice (2-D) of low uptake sphere ROI from background.(A) True labels for low uptake sphere ROI. (B) Time activity curves for voxels in ROI and background. (C) Segmentation performance of algorithms when voxels are represented as time activity curves. (D) Segmentation performance when voxels are represented as FPCs.

The k-means clustering algorithm had the second highest segmentation performance across the evaluated unsupervised clustering algorithms when segmenting individual sphere regions from its surrounding background. This algorithm was subsequently used to compare performance against the Rad-FIT clustering algorithm to segment the simulated image into three regions: background, blood, and all spheres. The Rad-FIT clustering algorithm was able to segment the low uptake spheres from its surrounding background region and resulted in lower mean percent error in the average sphere time activity curve (Figure 5.9).

A.



B.



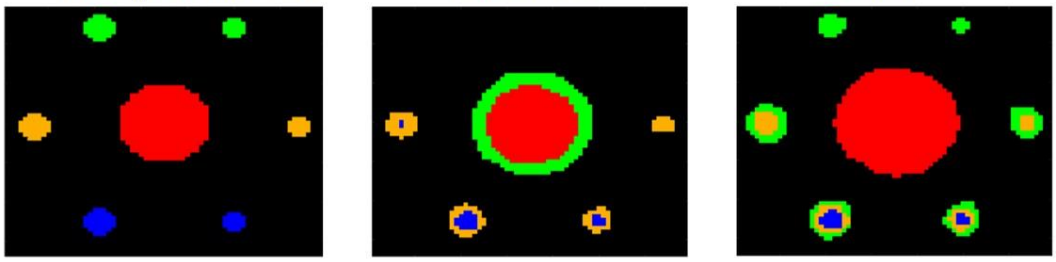
C.

	K-means clustering (FPC)		Rad-FIT clustering	
	Mean Percent error	Dice score	Mean percent error	Dice score
Background	2.9%	0.85	5.8%	0.95
Blood	4.1%	0.92	25.4%	0.70
Spheres	27.1%	0.65	7.1%	0.70

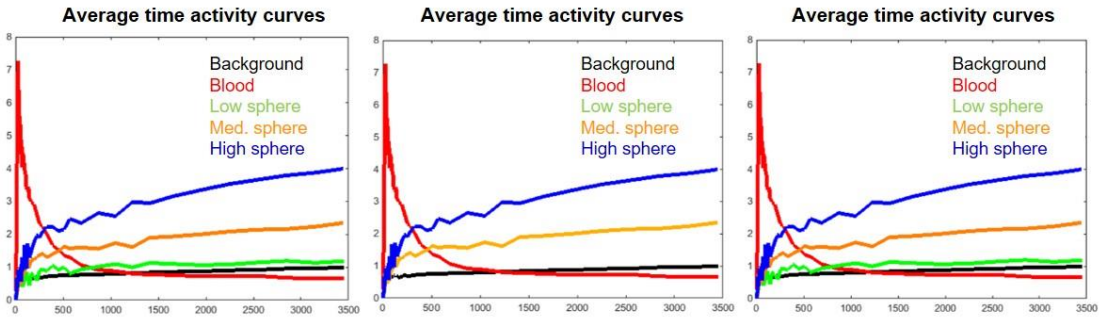
Figure 5.9. Comparing segmentation performances-three class labels. (A) Representative center slice image of true segmentation labels (left), segmentation results using K-means clustering (middle), and segmentation results using Rad-FIT clustering (right) when segmenting simulation images into three classes: background, blood, and spheres. (B) Average time activity curves for true simulation regions (left), K-means clustering identified regions (middle), and Rad-FIT clustering (right) identified regions. (C) Mean percent error and dice scores for segmentation results. The K-means clustering algorithm fails to segment the low uptake spheres from the background region.

The k-means clustering algorithm was subsequently used to compare performance against the Rad-FIT clustering algorithm to segment the simulated image into five regions: background, blood, low uptake spheres, medium uptake spheres, and high uptake spheres. The Rad-FIT clustering algorithm was able to segment the low uptake spheres from its surrounding background region and identify the five regions of distinct tracer uptake (Figure 5.10). In comparison, the K-means clustering algorithm was unable to identify and segment the low uptake spheres as a distinct region from its surrounding background.

A.



B.



C.

	K-means clustering (FPC)			Rad- FIT clustering		
	Mean percent error	Dice score	Jaccard Index	Mean percent error	Dice score	Jaccard Index
Background	6.0%	0.90	0.81	5.9%	0.95	0.91
Blood	24.0%	0.57	0.40	27.2%	0.69	0.53
Low spheres	100%	0	0	17.9%	0.32	0.20
Med. spheres	11.4%	0.61	0.44	12.2%	0.77	0.63
High spheres	18%	0.64	0.47	15%	0.66	0.50

Figure 5.10. Comparing segmentation performances-five class labels. (A) Representative center slice image of true segmentation labels (left), segmentation results using K-means clustering (middle), and segmentation results using Rad-FIT clustering (right) when segmenting simulation images into five classes: background, blood, and low, medium, and high uptake spheres. (B) Average time activity curves for true simulation regions (left), K-means clustering identified regions (middle), and Rad-FIT clustering (right) identified regions. (C) Mean percent error, generalized Dice score, and Jaccard indices for segmentation results. The K-means clustering algorithm fails to segment the low uptake spheres from the background region, and instead segments the blood region to two separate classes.

Overall, the Rad-FIT clustering algorithm demonstrated improved segmentation over the K-means clustering algorithm, as demonstrated by decreased mean percent error and increased Dice score and Jaccard index for each individual region's segmentation as well as segmenting the overall simulated image, measured over ten replicates. Due to this improved segmentation performance, the Rad-FIT clustering algorithm was utilized as a 4-D unsupervised clustering method towards characterizing functional tumor heterogeneity as a prognostic biomarker for women diagnosed with locally advanced breast cancer.

Chapter 6 : Evaluating the prognostic value of characterizing 4-D pharmacokinetic functional tumor heterogeneity

This chapter has been adapted from the following:

Chitalia, R., Viswanath, V., Pantel, A.R., Peterson, L.M., Gastounioti, A., Cohen, E.A., Muzi, M., Karp, J., Mankoff, D.A. and Kontos, D., 2021. Functional 4-D clustering for characterizing intratumor heterogeneity in dynamic imaging: evaluation in FDG PET as a prognostic biomarker for breast cancer. *European Journal of Nuclear Medicine and Molecular Imaging*, pp.1-12.

6.1. Introduction

Molecular and functional imaging modalities permit 4-D sampling of disease burden, capturing both spatial and temporal information that could illuminate various physiologic behaviors. Dynamic positron emission tomography (PET) imaging can quantify specific facets of tumor molecular biology [7,19,222] and can provide information beyond that of static imaging [22,23]. In particular, dynamic PET imaging of the glucose analog, ^{18}F-fluorodeoxyglucose (FDG), can provide simultaneous information on substrate delivery and metabolism [22]. Current clinical characterization of malignant lesions using PET imaging largely utilizes qualitative descriptors [8] and quantitative measures based on static radiotracer uptake (e.g. SUVmax) [223].

The emerging field of radiomics has introduced multi-parametric imaging features extracted with high-throughput computational analysis [33,101,196,224]. Previous work by Eary et al. quantified spatial heterogeneity of radiotracer uptake in static PET imaging and

demonstrated improved prognostic performance over established clinical markers [225]. Stoyanova et al. identified sub-regions within pre-clinical dynamic contrast enhanced (DCE)-MRI images of prostate tumors [226]. Similarly, Cherezov et al. identified tumor habitats using established radiomic texture features [227]. While such studies demonstrate the prognostic potential of characterizing 3-D spatial heterogeneity, and are in line with studies showing differential physiologic functionality across the whole tumor [228,229], these studies do not fully utilize the combined spatial and kinetic (e.g., 4-D) heterogeneity information available using imaging probe kinetics from modalities with high temporal imaging resolution.

The advantages of utilizing kinetic information from dynamic PET imaging have been demonstrated in breast cancer. Previous studies have demonstrated predictive improvement when FDG delivery (K_1) and FDG flux (Ki), in combination with [^{15}O]-water imaging, were utilized with static SUV measures as markers for neoadjuvant chemotherapy response in patients with locally advanced breast cancer [22,230,231]. However, these conventional kinetic parameters derived from FDG PET imaging at baseline alone were unable to show association to disease free survival, likely limited by being derived from the most metabolically active portion of the tumor, and therefore not fully capturing intratumor heterogeneity [22].

We have developed a method to characterize 4-D functional tumor heterogeneity (FTH) by capturing aspects of both spatial and kinetic tumor heterogeneity seen in dynamic imaging. The improved dynamic sampling and molecular specificity available in dynamic PET as compared to other imaging modalities may allow for non-invasive,

novel prognostic and predictive markers to characterize tumor molecular biology. The developed method is agnostic to the specific radiotracer utilized and does not depend on complex kinetic modeling assumptions. Instead, the approach is data driven in terms of identifying intrinsic 4-D patterns of molecular tumor heterogeneity. We present a clinical proof-of-principle by applying our algorithm on dynamic FDG PET imaging scans of primary locally advanced breast cancer. We also investigate the role of imaging signatures as a prognostic biomarker for locally advanced breast cancer and the improved predictive value of FTH characterization compared to standard dynamic and static analytic methods for FDG PET.

6.2. Study Cohort

To investigate the role of intratumor segmentation when characterizing functional heterogeneity, the prognostic value of functional tumor heterogeneity imaging signatures was explored on a previously published data set where serial dynamic FDG PET was shown to be predictive of response and recurrence using standard static uptake and kinetic analysis [22,230,231]. The goal was to test functional tumor heterogeneity imaging signatures extracted from dynamic FDG PET scans of women with locally advanced breast cancer imaging prior to treatment and compare their predictive value to standard approaches.

We used an anonymized data set consisting of women presenting at the University of Washington Breast Cancer Specialty Center with histologically confirmed breast carcinoma who underwent dynamic FDG PET imaging prior to neoadjuvant

chemotherapy and were followed for disease recurrence. The research protocol was approved by the institutional review board and patients studied provided informed consent prior to imaging and follow-up. The data set for this analysis was taken from a study first reported for 35 patients [230]. An additional 30 patients were later studied and added to a follow-up report of the data [22,231]. From this pooled data set of 65 women with complete baseline dynamic FDG PET scans who also completed neo-adjuvant chemotherapy and post-therapy surgery, two women were excluded for electing not to receive chemotherapy, three women were excluded for electing for medical care elsewhere, four patients were excluded for being unwilling to undergo mid-therapy imaging, two patients were excluded due to distant disease, and one patient was excluded due to little or no tracer uptake upon pre-therapy examination resulting in a total of 53 women. Of these, two women were excluded due to image artifacts and one woman was excluded due to incomplete survival information, resulting in a total of 50 women included in this study. Dynamic FDG PET images from these 50 women comprised our study sample reported here. Details of the patient population have been previously described[231].

Of the 50 women included in the data set, 17 women (34%) had recurrence events. A total of 47 (94%) women were diagnosed with infiltrating ductal carcinoma and 3 (6%) women were diagnosed with infiltrating lobular carcinoma.

Of the non-recurrent cases, 58% were ER positive, 52% were PR positive, and 18% were HER2 positive. Of the recurrent cases, 59% were ER positive, 59% were PR positive, and 35% were HER2 positive (Table 6.1).

Table 6.1. Selected study cohort characteristics.

	Non-recurrent cases (n=33,66%)	**Recurrent cases** (n=17, 34%)
Age		
30-39	5 (15%)	6 (35%)
40-49	14 (43%)	6 (35%)
50-59	13 (39%)	1 (6%)
60-69	1 (3%)	3 (18%)
70-79	0 (0%)	1 (6%)
Histologic subtype		
Infiltrating ductal	31 (94%)	16 (94%)
Infiltrating lobular	2 (6%)	1 (6%)
Receptor subgroup		
ER + /HER2 +	2 (6%)	5 (29%)
ER+/ HER2 -	16 (48%)	4 (24%)
ER- /HER2 +	2 (6%)	3 (18%)
ER-/HER2- (Triple negative)	11 (33%)	5 (29%)
pCR		
Complete response	8 (24%)	3 (17%)
ALN positivity		
Max, min, average	18, 0 , 2.26	18, 0 , 5.52
Baseline tumor size		
Max, min, average (cm)	11, 1.1, 5.2 cm	10, 1.9, 4.9 cm
Ki67 status	27 cases	14 cases
Low	2 (7%)	4 (27%)
Intermediate	2 (7%)	5 (33%)
High	23 (85%)	7 (47%)

Each woman had undergone 60-minute dynamic FDG PET centered over the breast prior to neoadjuvant chemotherapy and breast surgery. All women were imaged in the supine position and no positioning devices for immobilization were utilized. Women were infused with 218-396 MBq of FDG over 2 minutes in a 7-10 mL volume, with an intended injected dose of 370 MBq. Images for all women were acquired on an Advanced Tomograph (General Electric Medical Systems, Waukesha, WI) using the same image acquisition protocol. Dynamic images were acquired (25 image frames: 1- min pre-injection frame, 4 x 20 s, 4 x 40 s, 4 x 40 s, 4 x 1 min, 4 x 3 min, 8 x 5 min). Images were reconstructed into 35 x 128 x 128 voxel matrices with a spatial resolution of 10-12 mm[230]. Clinical information collected as part of the study included hormone receptor (HR) status consisting of estrogen receptor (ER) and progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2), clinical stage, tumor size, proliferation (Ki67), pathologic complete response (pCR), axillary lymph node (ALN) positivity, and age at diagnosis (Table 6.1). In the study, recurrence free survival (RFS) was tracked for each patient, defined as date of known recurrence, date of death, or date of most recent clinical follow-up with no evidence of disease, following the patient's date of surgery. Patients received standard of care follow-up including routine period imaging of CT scans, blood marker analysis (CA2729), and follow-up visits to check for symptoms.

Established ROI-based measures of uptake and kinetics for dynamic FDG PET – summed imaged standardized uptake value (SUV), and the kinetic parameters of FDG blood-to-tissue transport (K_1) and FDG trapping flux (Ki) - were calculated based on kinetic modeling of dynamic data for each woman and have been previously

reported[22,230]. These kinetic parameters were measured for each tumor from a 1.5 cm-diameter circle VOI surrounding the area of maximal tumor FDG uptake seen on the 30-60m summed image.

6.3. Functional tumor heterogeneity (FTH) signature extraction

A 3-D bounding region surrounding each unifocal lesion was manually identified by a radiologist blinded to the outcome of each patient using the final of the summed FDG images for the 25 imaging frames (30-60 minutes post-injection) and guided by ROIs previously used for extraction of SUV, K_1, and Ki for consistency. An established segmentation approach was applied to the TACs generated from the 25 imaging frames of each voxel within the bounding region to segment the tumor from its surrounding background[217].

Within the segmented 3-D tumor region, Rad-FIT clustering was applied to segment each tumor region into three, spatially constrained sub-regions with distinct functional behavior. Three sub-regions were selected based on the rationale that there are currently three major subtypes of breast tumors broadly recognized: hormone receptor positive, HER2 positive, and triple negative[4]. The three sub-regions within each tumor were ranked in order of descending mean value of the first FPC to allow for consistent comparisons across tumors.

The resulting Rad-FIT clustering within each tumor was summarized using metrics describing sub-region compactness and separation. These metrics were chosen to summarize how well the functional behavior of each tumor will cluster into three groups

and to allow for comparisons of intratumor heterogeneity across women. Compactness was measured using the between cluster sum of squares (BCSS) scaled by the total sum of squares (TSS) :

$$\frac{\sum_{k=1}^{K}(\overline{x_k}-\bar{X})^2}{\sum_{i=1}^{N}(x_i-\bar{X})^2} \quad \text{Eq. 6.1}$$

where K represents the 3 sub-regions and N is the total number of voxels within each tumor. The separation between sub-regions was determined using the Bhattacharya distance [232] to calculate the distance between FPC distributions of two sub-regions, represented as ϕ and defined as:

$$\phi(p,q) = -ln\left(\sum_{x \in X}\sqrt{p(x)\,q(x)}\right) \quad \text{Eq. 6.2}$$

Use of this distance allows for a similarity measure between the distributions of FPC values within two sub-regions.

Based on the definitions above, a total of four features summarizing intratumor heterogeneity from Rad-FIT clustering results were extracted to form an FTH signature (Figure 6.1): (1) BCSS/TSS, (2) distance between sub-region 1 and 2 ($\phi(1,2)$), (3) distance between sub-region 2 and 3 ($\phi(2,3)$), and (4) distance between sub-region 1 and 3 ($\phi(1,3)$).

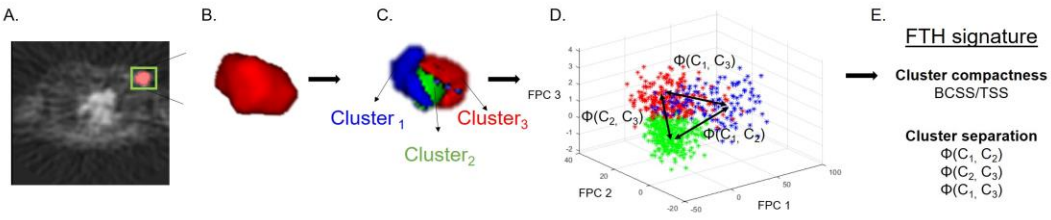


Figure 6.1. FTH signature extraction. (A) 3-D tumor region identified by a radiologist shown in green. (B) Automated 3-D segmentation of tumor from background with pixels represented using time activity curves. (C) Rad-FIT clustering performed, identifying three, spatially contiguous sub-regions. (D) Cluster compactness and cluster separation distances are calculated to form features in FTH signature. (E) Intratumor heterogeneity summarized using the FTH signature.

BCSS/TSS provides a measure of how compact the resulting clusters are; the more compact each cluster is, the greater heterogeneity between the identified sub-regions. Calculating the distance between the distributions of FPC values of two sub-regions provides a metric for how separated the clusters are; a greater distance between sub-regions indicates greater heterogeneity within the whole tumor region. This FTH signature can be used to interpret how distinct the three identified sub-regions are within each tumor. As such, the average value of the FTH signature, or FTH signature index, can provide a metric for intratumor heterogeneity across tumors.

6.4. Statistical Analysis: Evaluation of FTH signatures as a prognostic biomarker

Our goals in this proof of principal study were to test the prognostic value of FTH signatures from dynamic breast cancer FDG PET and to assess for incremental value compared to standard clinical parameters and conventional FDG PET static and dynamic analysis measures used in prior published analyses. FTH signatures were first z-score normalized across all women. Time-to-event analysis was then used to assess the prognostic value of the FTH signatures in predicting recurrence-free survival (RFS). To

this end, a three-fold cross validated (CV) Cox proportional hazards model was used to compare improved prognostic discriminatory capacity over baseline models of established prognostic factors consisting of ER status, PR status, tumor size, pCR, and ALN positivity and kinetic parameters consisting of the SUV, K_1, and Ki. These prognostic factors were chosen based on the available data as well as the intent to compare analysis results to prior published data [22,230,231].

Model performance was evaluated using an averaged C-statistic over the test sets for all three folds and the log-likelihood statistical test.

The prognostic value of the FTH signature was evaluated via Kaplan-Meier survival analysis using each patient's risk core, dichotomizing patients into high and low risk groups. The risk score for each patient was defined as the patient's FTH signature weighted by the corresponding coefficients from each of the three test sets from a 3-fold cross validated model for each covariate in the FTH signature [184,224]. Risk scores generated from baseline features of ER status, PR status, tumor size, pCR, and ALN positivity and from baseline and kinetic features were also assessed. Statistical significance of Kaplan-Meier stratification was evaluated using the Log Rank Test.

Lastly, an exploratory unsupervised hierarchical clustering was performed on the extracted FTH signatures from each tumor. The resulting *c* clusters obtained from the hierarchical clustering algorithm were interpreted as *c* intrinsic FTH phenotypes seen in this study population.

The optimal number of stable FTH phenotypes was determined using consensus clustering[166]. Statistical significance of the identified, stable FTH phenotypes was evaluated using the SigClust method[167].

The distribution of histopathologic and kinetic prognostic covariate values across tumors assigned to each of the FTH phenotypes was assessed using chi-square tests for categorical biomarkers and one-way analysis of variance test for continuous biomarkers.

6.5. Results Evaluation of the FTH signature as a prognostic biomarker

Representative tumor images after Rad-FIT clustering demonstrate intratumor heterogeneity within breast tumors (Figure 6.2). Tumors with increased intratumor heterogeneity can be identified as having sub-regions with distinct time activity curve behaviors, while tumors with decreased intratumor heterogeneity display little distinction between the time activity curve behavior of the identified sub-regions.

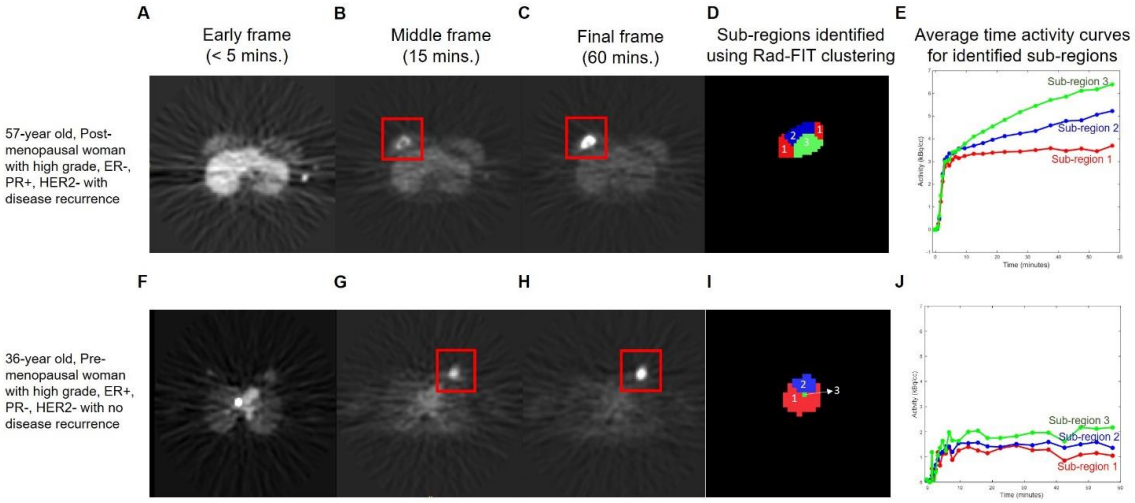


Figure 6.2. Representative images for the primary tumors of two women diagnosed with locally advanced breast cancer with future tumor recurrence (top) or tumor non-recurrence (bottom). (A) representative slice from an early frame at less than 5 minutes after tracer injection, (B) representative slice from middle frame at 15 minutes after tracer injection, and (C) representative slice from final frame taken at 60 minutes after tracer injection of a 57-year old, post-menopausal woman with a high grade, ER-, PR+, HER2-, tumor who had disease recurrence upon follow-up (top). (D) Three sub-regions identified using Rad-FIT clustering labeled as region 1 (red), 2 (blue), and 3 (green), and (E) average TACs for each identified sub-region. (F) A representative slice from an early frame at less than 5 minutes after tracer injection, (G) slice from middle frame at 15 minutes after tracer injection, (H) and representative slice from final frame taken at 60 minutes after tracer injection of a 36-year old, pre-menopausal woman with a high grade, ER+, PR-, HER2-, tumor with no disease recurrence (bottom). (I) Three sub-regions with distinct 4-D behavior identified using Rad-FIT clustering labeled as region 1 (red), 2 (blue), and 3 (green), and (J) average time activity curves for each identified sub-region.

As expected, in a full multivariate Cox proportional hazards model after adjusting for ER status, PR status, tumor size, pCR, and ALN positivity (Table 6.2), $\phi(1,2)$ and $\phi(2,3)$ were associated with disease free survival (Table 6.3).

Table 6.2. Risk of breast cancer recurrence in Hazzard Ratios (HR) associated with baseline model.

Covariate	**HR**	**95% CI**	**p-value**
ER status	1.17	0.24-9.98	0.87
PR status	0.51	0.05-4.8	0.55
Tumor size	0.99	0.82-1.22	0.99
pCR	1.12	1.01-1.21	0.01 *
ALN positivity	0.85	0.23- 3.15	0.81

Table 6.3. Risk of breast cancer recurrence associated with FTH imaging signature adjusting for baseline and kinetic features.

Covariate	**HR**	**95% CI**	**p-value**
BCSS/TSS	1.08	0.62-1.89	0.77
$\phi(1,2)$	0.04	0.002-0.66	0.02*
$\phi(1,3)$	0.82	0.48-1.39	0.46
$\phi(2,3)$	14.08	2.41-21.18	0.003*

A baseline, three-fold CV Cox proportional hazards model consisting of ER status, PR status, tumor size, pCR, and ALN positivity resulted in a mean C-statistic of 0.51 when predicting RFS. Adding SUV, K_1, and Ki parameters to the baseline model resulted in a mean CV C-statistic of 0.54. Adding the FTH signature to the baseline model improved the mean CV C-statistic to 0.74 ($p < 0.01$) (Figure 6.3A).

Dichotomizing patients into low- and high-risk groups based on the baseline model risk scores (Figure 6.3B) and baseline plus the kinetic model risk scores (Figure 6.3C) demonstrated no statistically significant separation between Kaplan-Meier curves. Patient dichotomization into low- and high-risk groups based on the baseline plus FTH signature risk scores (Figure 6.3D) resulted in a statistically significant separation between Kaplan- Meier curves ($p < 0.05$) for RFS probability.

A

Cox proportional hazard's model	3-fold CV C- statistic
Baseline	0.51
Baseline + kinetic	0.54
Baseline + FTH signature	0.74

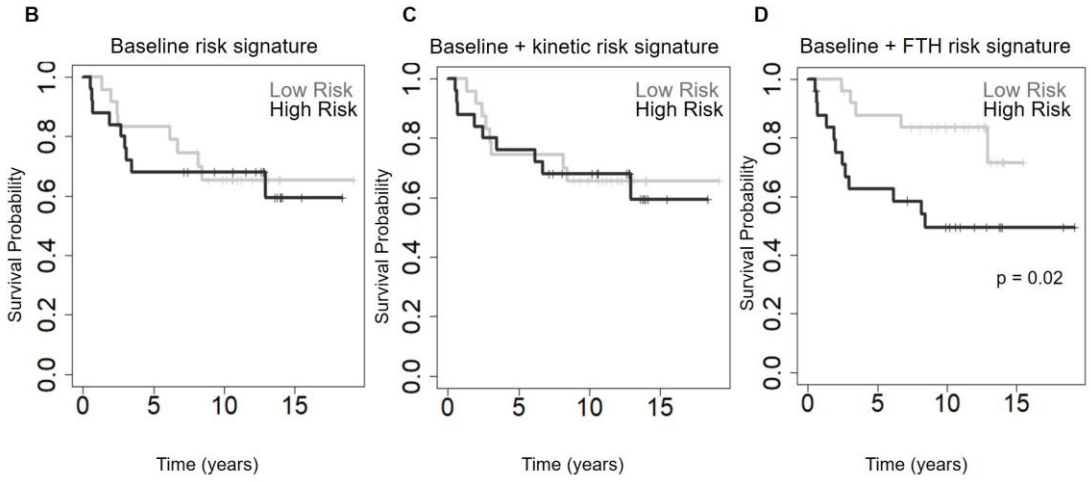


Figure 6.3. FTH survival analysis. (A) Cross validated c-scores and Kaplan-Meier survival curves for baseline (ER status, PR status, tumor size at baseline, pCR, and ALN positivity), baseline plus kinetic (ER status, PR status, tumor size at baseline, pCR, ALN positivity, SUV, K_1, Ki) and baseline plus FTH signature models. Kaplan-Meier curves generated when patients are stratified by risk scores generated from (B) the baseline model, (C) baseline plus kinetic features model, and (D) baseline plus FTH signature model.

Unsupervised hierarchical clustering of women based on the extracted FTH signatures from each tumor identified two clusters which were interpreted as FTH

phenotypes seen in the study population and found to be statistically significant via the SigClust method (p=0.04) (Figure 6.4A).

As a higher FTH signature index suggests greater separation between the three sub-regions' FPC values and can therefore be interpreted as greater intratumor heterogeneity, the identified FTH phenotypes were ranked based on the mean FTH signature index found across all women assigned to each phenotype. The resulting FTH phenotypes 1 and 2 were interpreted as a low FTH versus high FTH phenotypes, respectively, with tumors in phenotype 1 having lower mean FTH signature indices (blue color in Figure 6.4A), versus tumors in phenotype 2 which had on average higher FTH signature indices (red color in Figure 6.4A).

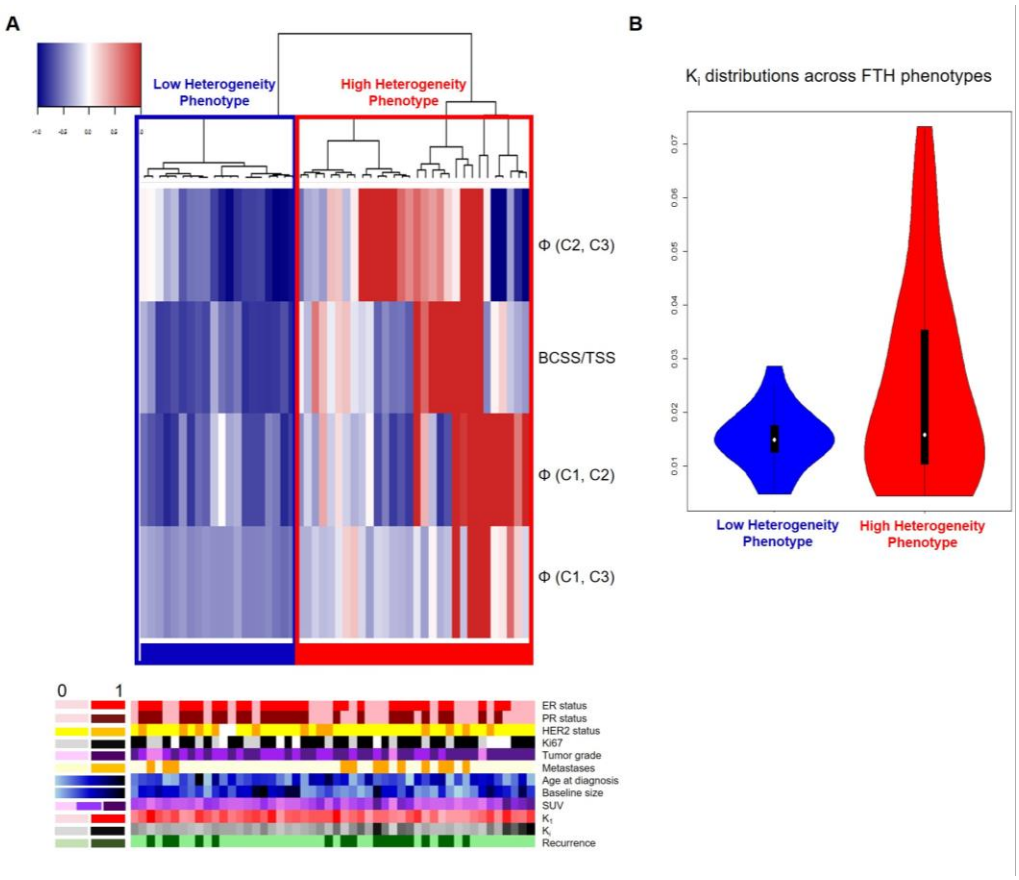


Figure 6.4. FTH phenotypes. (A) Unsupervised hierarchical clustering of FTH signatures identifies 2 significant phenotypes of FTH, with clinical covariate distribution across identified phenotypes displayed in the bottom legend. The resulting cluster dendrogram can be seen above a heatmap in which each row represents a feature within the FTH signature, and each column represents a tumor (B) Distributions of Ki across the identified phenotypes were found to be statistically significant ($p<0.05$).

Tumor PR status was found to be statistically significantly different across the two FTH phenotypes ($p<0.05$), with tumors in the low FTH phenotype having a higher proportion of PR positive tumors. Other clinical covariates including ER status, HER2 status, Ki67 status, pCR, ALN positivity, and tumor grade were not statistically significant across identified phenotypes. From the FDG PET covariates, Ki was found to be statistically significantly different across the two phenotypes ($p<0.05$), with tumors in the high FTH phenotype having a greater interquartile range (0.025) and greater variance

(0.0003) than tumors in the low FTH phenotype (interquartile range: 0.004, variance: 3.27e-5) (Figure 6.4B). K_1 and SUV values were not found to be statistically significant across the identified phenotypes.

6.6. Discussion

Our results suggest that incorporation of both spatial and kinetic information in a 4-D dynamic activity curve clustering paradigm allows for improved segmentation of dynamic PET imaging data over established unsupervised clustering techniques utilizing kinetic information alone. Established unsupervised voxel parcellation techniques largely assume voxel independence[8], and as such, may be inadequate for identifying spatially constrained, functionally similar sub-regions under the hypothesis that sub-clonal populations can occupy spatially contiguous regions with common biologic properties [39]. Along these lines, partial volume effects seen in imaging modalities suggest that neighboring voxels may share information regarding underlying tissue structure due to the spatial limitations of the imaging device[204]. Therefore, analyzing imaging presentations of intratumor heterogeneity requires a fully 4-D approach.

We had the goal of developing methodology broadly applicable to PET tracers with similar kinetic features. With this in mind, dynamic simulations utilized for Rad-FIT development and validation (Chapter 5) were based on FLT simulated data and used to select an approach which was then applied to a previously collected FDG patient dataset to examine the role of FTH as a prognostic biomarker. The rationale was that while simulation curves were generated using kinetic parameters specific to FLT (flux between 0.03-0.1 mL/min/g), these kinetics curves have parameters similar to FDG PET curves

(flux between 0.02 - 0.09 mL/min/g), and they can be generalized to all tracers fit to a two-compartment model, including FDG, as long as the model and range of parameters has overlap with these tracers. Additionally, and as our algorithm is in principle agnostic to the type of tracer used or any related kinetic modeling parameters, utilizing simulated images of a different two-compartment radiotracer during Rad-FIT development allowed for a more generalizable algorithm that was not biased towards a single specific tracer in subsequent analyses.

Extending the Rad-FIT clustering algorithm to characterize intratumor heterogeneity has the potential to identify intratumor sub-regions with discrete functional behavior. This is supported by the average TACs from the identified sub-regions in representative tumors (Figure 6.2), where the tumor with disease recurrence clustered into three sub-regions with distinct curve patterns. The tumor with no disease recurrence and characterized as ER+, demonstrated mostly low uptake and non-rising curves in the identified sub-regions.

Quantifying intratumor heterogeneity using the FTH imaging signature demonstrates prognostic value when predicting RFS. Cox-regression models incorporating FTH signatures added to a baseline model demonstrated a statistically significant improvement in C-statistic. While the dichotomization of baseline risk scores based on known prognostic features did not demonstrate significant Kaplan-Meier curve separation in this relatively small sample size, consistent with previous analyses of this study cohort[22,231], improvement over models combining baseline and kinetic features emphasizes the added prognostic value of utilizing quantitative features summarizing

dynamic tumor behavior over the entire volume. Additionally, risk scores generated using the baseline features and FTH signature resulted in statistically significant patient dichotomization into low- and high-risk groups for RFS using Kaplan-Meier survival analysis as compared to risk signatures generated from the baseline model and baseline plus kinetic model.

While prior studies have shown prognostic value for measures obtained from serial dynamic FDG PET using standard kinetic analysis methods[231], pre-therapy FDG dynamic data were not significantly predictive. Similarly, prior studies demonstrated the predictive value of pre-therapy measures of FDG flux and tumor blood flow obtained from combined ^{15}O-water PET and FDG PET studies [230,231], but pre-therapy FDG kinetic measures alone were not significantly predictive of RFS. In this preliminary analysis, use of the Rad-FIT clustering algorithm extracted significantly prognostic 4-D signatures from pre-therapy dynamic FDG PET data that did predict RFS, a notable incremental improvement on standard approaches to dynamic PET analysis of considerable potential significance.

Additionally, our results suggest that intrinsic imaging phenotypes may exist within locally advanced breast tumors corresponding to FTH. In particular, statistically significant differences in the FDG flux constant, Ki, were seen across the two phenotypes with tumors corresponding to higher degrees of FTH having higher values of Ki. This finding suggests that the tumor characteristic of increased metabolic rate may be captured within the FTH imaging signature generated from the 4-D clustering performed using Rad-FIT and may have prognostic significance when expanded to a larger study cohort.

Interestingly, compared to tumor clinical and histopathologic features we found significant differences across the FTH phenotypes in PR expression, a marker shown to be an indicator of tumor ER functionality and a more differentiated breast cancer biologic phenotype [233].

Limitations of our study should be noted. First the Rad-FIT clustering algorithm utilizes K-means clustering as an initialization to the method, which can allow for sensitivity to cluster initialization due to K-means clustering identifying local optima. Future studies will be conducted to evaluate segmentation performance when random cluster initializations are selected. Additionally, our study utilized a relatively small sample size of patients. To account for potential model overfitting, we utilized three-fold CV in our time-to-event analysis, to ensure model robustness. The identification of FTH phenotypes within the study population is limited by a lack of independent validation and instead was conducted as an exploratory analysis. Future work will include expanding this initial, exploratory analysis to a larger cohort as well as validating the identified FTH phenotypes. While the Rad-FIT clustering paradigm identified three clusters within each tumor, the optimal number of functionally discrete sub-regions may vary across tumors. In this exploratory study, the selection of three for the number of subtypes was chosen empirically, guided by the three major subtypes of breast cancer (ER+/PR+, Her2+, Triple negative). Future work will also include optimization of the Rad-FIT clustering algorithm such that an optimal number of clusters is identified within each tumor. In addition, we have evaluated only the pre-therapy time point in this initial analysis. All women included in our study underwent neoadjuvant chemotherapy and repeat mid-

therapy imaging. To account for the effect of treatment on intratumor heterogeneity and corresponding FTH signatures, we plan to expand our analysis to dynamic FDG PET images taken also during the midpoint of each woman's therapy in a future study. Lastly, alternative approaches exploring a linear analysis of 4-D dynamic PET using a mixture-based approach have been previously reported [234,235]. Future work will include expanding our analysis to compare methodologies and potentially include a mixture-based component. Lastly, we developed and applied this method on simulated dynamic images and clinical dynamic scans of breast cancer patients, with a larger goal of extending this method towards analyzing other solid tumors and different PET tracers in future work.

6.7. Conclusion

In conclusion, we have developed a 4-D clustering and segmentation algorithm to identify functionally discrete, spatially constrained sub-regions within breast tumors that was able to generate prognostic measures from pre-therapy dynamic FDG PET of locally advanced breast cancer not previously identified by ROI-based kinetic analysis. Our results demonstrate that quantifying functional tumor heterogeneity can provide independent and additional prognostic value and may provide a non-invasive- 4-D characterization of breast tumors towards personalized decision making.

Chapter 7 : Extended functionalities of Rad-FIT clustering: exploring the feasibility of an unsupervised method to summarize 4-D pharmacokinetic spatial heterogeneity

7.1 Introduction

We have explored extending the functionalities of a novel 4-D clustering approach to model spatially constrained, functionally discrete sub-regions within breast tumors and extract factors that summarize this heterogeneity. We have previously assessed the initial clinical applications of this approach in developing personalized prognostic imaging signatures characterizing breast tumor heterogeneity by using a fixed number of sub-regions[236]. In this preliminary work, we remove this constraint by using a multivariate Kullback-Leibler divergence-based minimization approach to automatically determine the optimal number of sub-regions for each tumor. We then perform an evaluation of this unsupervised 4-D segmentation algorithm using simulated image phantoms modeling heterogeneous breast lesions. We compare the segmentation performance of the developed algorithm against established unsupervised clustering algorithms.

7.2 Rationale for unsupervised method

Initial development and applications of using Rad-FIT clustering to summarize intratumor heterogeneity involved a pre-determined number of sub-regions for the algorithm to identify (Chapters 5 and 6)[236]. Preliminary exploration of FTH as a prognostic biomarker resulted in predefining the number of intratumor sub-regions to

three, assuming the three broad receptor subtypes for breast cancer (Chapter 6)[236]. However, the degree of intratumor heterogeneity across breast tumors may vary and as such, we wanted to extend Rad-FIT clustering to determine the optimal number of sub-regions within a single tumor in an unsupervised manner. Furthermore, molecular heterogeneity may go above and beyond the three broad histopathologic subtypes (ER+/PR+, Her2+, Triple negative), and we aimed to develop an approach where this determination could be more data driven. Identifying the number of functionally discrete sub-regions, unique to a tumor, may also allow for precise longitudinal monitoring of changes in intratumor heterogeneity, including identifying potentially emerging heterogeneity due to de-novo treatment resistance.

7.2 Heterogeneity phantom images

Four sets of large heterogeneous lesions were simulated using GATE v8.1[211] and cropped into four separate 4-D images with dimensions of 71 x 71 x 68 x 70 frames. Each image consisted of a set of three 10-cm-diameter spheres, with centers placed in the same axial location 5 cm apart from each other to form an equilateral triangle. This allowed for overlap of the edges of the spheres to emulate a large heterogeneous lesion, where overlapping areas had the combined activity of each sphere. The four sets of spheres had the following activity to background ratios at 60-min post-injection: Simulation 1) 2:1, 2:1, 2:1; Simulation 2) 2:1, 3:1, 4:1; Simulation 3) 4:1, 6:1, 8:1; and Simulation 4) 10:1, 20:1, 30:1 (Figure 7.1).

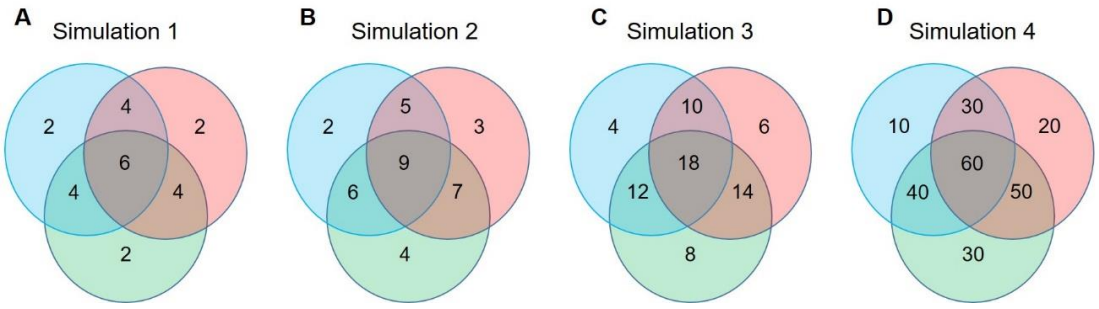


Figure 7.1. Diagrams of simulated heterogeneity phantoms demonstrating sphere overlap and the resulting flux scales for overlapping regions in simulation 1 (A), simulation 2 (B), simulation 3 (C), simulation 4 (D).

Time activity curves (TACs) within each of the simulated images were based on previous FLT dynamic simulations where the background TAC is based on muscle data and the 2:1, 4:1, and 8:1 contrast spheres were based on previously shown[213] low, medium, and high flux spheres, respectively. Higher contrast spheres were simulated based on scaled TACs of the high flux sphere. Simulated data were binned into 70 frames (17 x 1.2 s, 10 x 2.4 s, 12 x 4.8 s, 4 x 12 s, 11 x 30s, 4 x 60s, 6 x 180s, 6 x 300s), reconstructed using list-mode TOF OSEM[237] and binned into 2 mm isotropic voxels. The true number of discrete functional regions was determined based on the spatial location and average activity value of the region, resulting in: Heterogeneity simulation 1: three clusters, Heterogeneity simulation 2: five clusters, Heterogeneity simulation 3: seven clusters, Heterogeneity simulation 4: six clusters (Figure 7.2).

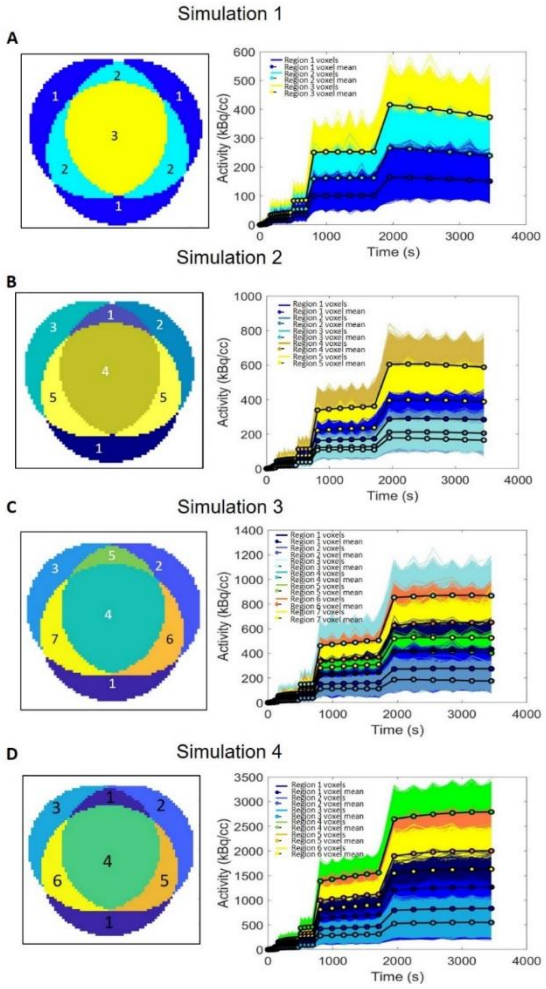


Figure 7.2. The discrete sub-regions within each heterogeneity simulation. The number of regions within the heterogeneity image based on the flux scaling of overlapping sphere regions and corresponding voxel time activity curves and mean time activity curves for each individual region in simulation 1 (A), simulation 2 (B), simulation 3 (C), and simulation 4 (D).

7.3 Determining optimal number of clusters

In order to select the optimal number of clusters within each tumor, we amended the methodology initially presented by Eloyan et al.[220] to incorporate multivariate Gaussian distributions of voxel representations and Markov Random Field (MRF)-based prior probabilities for each voxel.

The goal for determining an empirical estimation of true voxel labeling assignment is to minimize the distance between the estimated probability function of a distribution with its true underlying distribution. To that end, when the optimal number of cluster sub-regions, unique to a specific tumor, is determined, the Kullback-Leibler (KL) divergence between true $p(\boldsymbol{v}_{\mathbf{x,y,z}})$ and estimated final $p(\boldsymbol{v}_{x,y,z,f})|z_{x,y,z} = k\ ,\phi)$ probability density will be 0.

However, as the true probability density for the tumor image is unknown, the difference in this KL divergence moving from *k* to *k*+1 clusters is leveraged such that:

Eq. 7.1

$$\frac{1}{V}\sum_{1}^{V}\log\left(\frac{p(\boldsymbol{v}_{\mathbf{x,y,z}})}{p(\boldsymbol{v}_{x,y,z,f})|z_{x,y,z}=k\ ,\phi}\right)-\frac{\mathbf{1}}{\boldsymbol{V}}\sum_{1}^{V}\log\left(\frac{p(\boldsymbol{v}_{\mathbf{x,y,z}})}{p(\boldsymbol{v}_{x,y,z,f})|z_{x,y,z}=k+1\ ,\phi}\right)$$

$$=\frac{1}{V}\sum_{1}^{V}\log\left(\frac{p(\boldsymbol{v}_{x,y,z,f})|z_{x,y,z}=k+1\ ,\phi}{p(\boldsymbol{v}_{x,y,z,f})|z_{x,y,z}=k\ ,\phi}\right)$$

This ratio can be empirically estimated for each voxel moving from *k* to *k*+1 clusters as:

Eq. 7.2

$$\Delta_{\mathbf{v}}=\log\left(\frac{p(\boldsymbol{v}_{x,y,z,f})|z_{x,y,z}=k+1\ ,\phi}{p(\boldsymbol{v}_{x,y,z,f})|z_{x,y,z}=k\ ,\phi}\right)$$

and summarized over all voxels as:

Eq. 7.3

$$\sqrt{V}\left(\frac{\bar{\Delta}_{v}}{var\ (\Delta_{v})}\right)$$

This metric is calculated across an array of values for *k*, and the value *K* at which a minimum is reached is selected as the optimal number of sub-regions.

7.4 Beta parameter selection

While previously implemented Rad-FIT clustering utilized a set value for the floating parameter, β, based on signal to noise variation across the simulated images, it is likely that the optimal value of β will vary across tumors. This is due to variations in image acquisition and reconstruction protocols resulting in varying image signal to noise ratios and varying partial volume related effects seen across clinical images. As such, the dependency of a voxel's sub-region classification on its surrounding neighborhood may vary.

A previous study by Eloyan et al[220]. assessed parameter values of 4, 6, and 12 and found no significant difference in the resulting image clustering. Ultimately, the authors selected the smallest value of beta that prevented losing image sharpness. When evaluating the final number of sub-regions within tumors from a previously analyzed cohort of dynamic FDG PET images of invasive breast cancer[236] when beta parameter values ranged from 0.75 to 3, we saw large variations in the resulting number of sub-regions as well as their spatial distribution within a tumor (Figure 7.3).

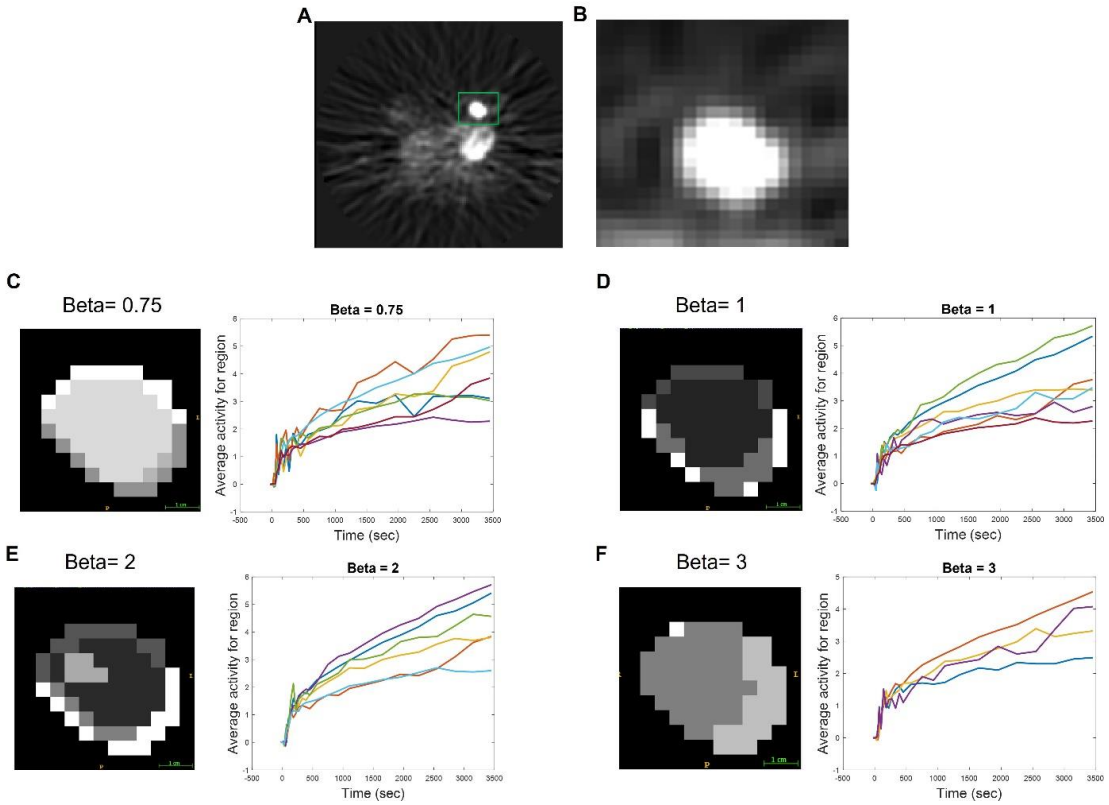


Figure 7.3. Representative breast lesion segmented using U-Rad-FIT clustering at varying values of beta. Representative frame of dynamic PET image (A) and lesion ROI (B). Resulting cluster segmentations and each region's average time activity curves when using beta = 0.75 (C), beta= 1(D), beta =2 (E), and beta =3 (F). Smaller values of beta result in varying numbers of sub-regions, likely picking up image noise.

This suggested that not only is selection of an optimal value required for each tumor, but also that the value of the beta parameter should be larger to avoid characterizing image noise as unique sub-regions. When larger values of beta are selected (e.g. beta= 6) resulting average time activity curves from the U-Rad-FIT derived sub-regions demonstrate discrete curve behavior and result in more anatomically realistic sub-regions in relation to the lesions' size (Figure 7.4).

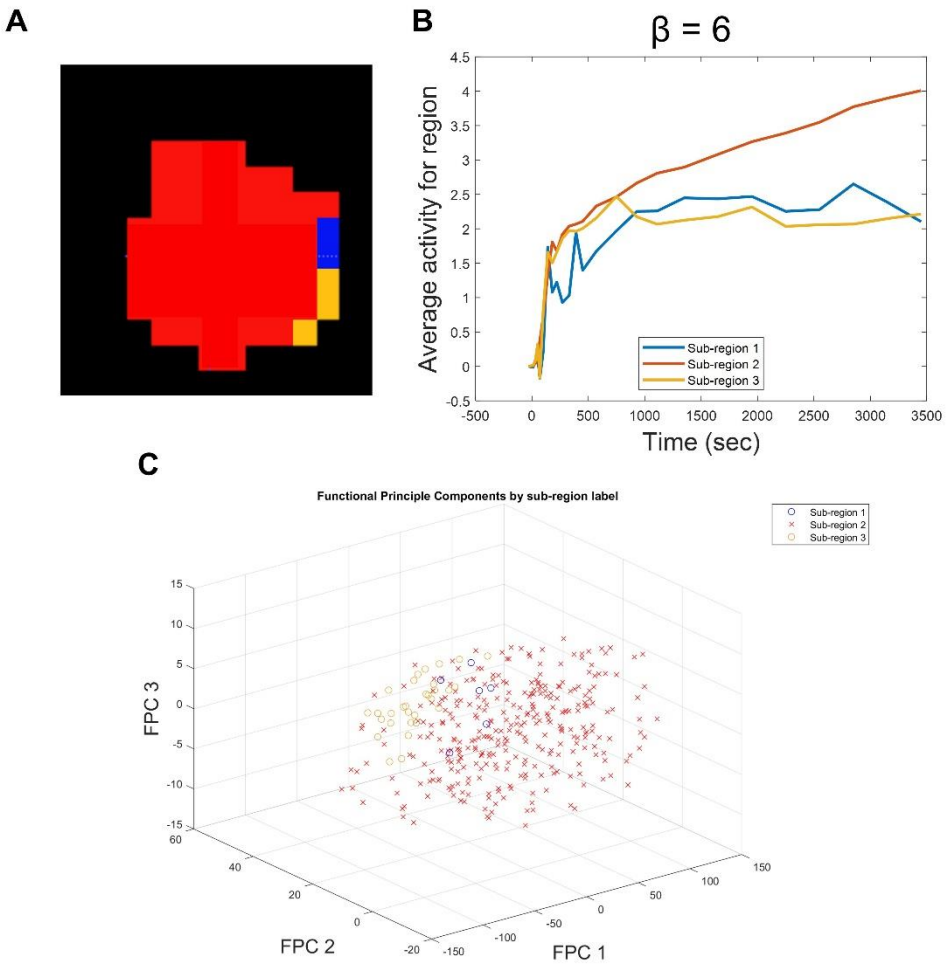


Figure 7.4. Representative lesion segmented using U-Rad-FIT clustering using a beta value of 6 (A). Each identified sub-region's average time activity curves (B), and all voxels represented as FPCs (C). Using a larger value of beta allows for a more physiologically likely number of identified sub-regions, as larger spatial weights on prior probabilities may result in less sensitivity to image noise.

Therefore, to account for variations in signal to noise across images and to select the optimal value of the floating beta parameter unique to each lesion, we implemented a selection criterion for choosing a value of beta which maximizes inter-cluster separation and minimizes intra-cluster variance in time (functional principal component) and space domains. To implement this, we leveraged the Calinski-Harabasz (CH) index[238] calculated for both FPCs and spatial coordinates:

$$CH_{FPC}\frac{\sum_{k=1}^{K} n_k \left\| \bar{v}_{k_{fpc}} - \bar{V}_{fpc} \right\|^2}{K-1} * \frac{n-K}{\sum_{k=1}^{K}\sum_{i=1}^{n_k} \left\| v_{i_{fpc}} - \bar{v}_{k_{fpc}} \right\|^2} \quad \text{Eq. 7.4}$$

$$CH_{spatial}\frac{\sum_{k=1}^{K} n_k \left\| \bar{v}_{k_{xyz}} - \bar{V}_{xyz} \right\|^2}{K-1} * \frac{n-K}{\sum_{k=1}^{K}\sum_{i=1}^{n_k} \left\| v_{i_{xyz}} - \bar{v}_{k_{xyz}} \right\|^2} \quad \text{Eq. 7.5}$$

$$CH_{4D} = CH_{FPC} + CH_{spatial} \quad \text{Eq. 7.6}$$

where n_k is the number of points in cluster *k*, $\bar{V}_{fpc}$ is the FPC centroid of the entire data set, $\bar{v}_{k_{fpc}}$ is the fpc centroid of cluster *k*, and $v_{i_{fpc}}$ is the fpc value for voxel, v_i, in cluster *k*. Similarly, $\bar{V}_{xyz}$ is the spatial coordinates of the centroid of the entire data set, $\bar{v}_{k_{xyz}}$ is the spatial coordinates of the centroid of cluster *k*, and $v_{i_{xyz}}$ is the spatial coordinates for voxel, v_i, in cluster *k*. The total number of voxels is defined as *n* and *K* is the total number of clusters.

The value of β resulting in the highest modified CH index (CH_{4D}), was selected as the optimal value for that tumor. This would result in selecting the value of β that resulted in the most compact and well-separated clusters in time and space.

7.5 Unsupervised Radiomic Functional Intratumor (U-Rad-FIT) clustering

As a result of determining the optimal number of sub-regions using the optimized value for the floating parameter, β, the algorithm for Rad-FIT clustering has been extended to Unsupervised Radiomic Functional Intratumor (U-Rad-FIT) clustering, and can be defined as follows in below in Algorithm 1:

Algorithm 7.1. Unsupervised Radiomic functional intratumor (U-Rad-FIT) clustering

Input: $\mathbf{V_i} \in R^{3\times t}$, $\beta = 4 \ldots B$ (floating parameter), $\kappa = 1 \ldots K$ (number of sub-regions)

Output: $\mathbf{z} \in [0,1]^K$ (Clustering Assignment)

Step 1: Functional principal component analysis (retain components with 85% of variance: $\mathbf{V_{(x,y,z,t)}} \gg \mathbf{V_{(x,y,z,f)}}$

for k = *1 …K*

for $\beta = 4 \ldots B$

Rad-FIT Clustering:

Initialization: Initialize **z,** $\boldsymbol{\mu}, \boldsymbol{\sigma}$ by K-means clustering algorithm

Loop- Repeat (t) until convergence

Expectation step: Fix **z,** $\boldsymbol{\mu}, \boldsymbol{\sigma}$

$$p\left(z_{x,y,z} = k \,|z^l, l \in N_{x,y,z}\right) = \frac{e^{\beta\, \delta_{x,y,z(k)}}}{\sum_{k=1}^{K} e^{\beta\, \delta_{x,y,z(k)}}}$$

$$p\left(v_{x,y,z,f}\right)|z_{x,y,z} = k\,, \phi) = \frac{1}{2\pi^{m\backslash 2}\,|\Sigma_k|^{\frac{1}{2}}} e^{-\frac{1}{2}\left(v_{x,y,z,f}-\mu_k\right)^T \Sigma_k^{-1}\left(v_{x,y,z,f}-\mu_k\right)}$$

$$p\left(z^t_{x,y,z} = k \middle| v_{x,y,z,f}\right) = \frac{p\left(v_{x,y,z,f}\middle|z^t_{x,y,z} = k\right)p\left(z^t_{x,y,z}=k\right)}{\Sigma^n p\left(v_{x,y,z,f}\middle|z^t_{x,y,z} = k\right)p\left(z^t_{x,y,z}=k\right)}$$

Solve for $z^t = \underset{z\,\in Z}{argmax}\left[P\left(v_{x,y,z,f}\middle|z_{x,y,z},\ \boldsymbol{\mu}, \boldsymbol{\sigma}^t\right)P(Z)\right]$

Maximization step:

$$\mu_k^{t+1} = \frac{\Sigma^n p\left(z^t_{x,y,z} = k\ |v_{x,y,z,f}\right)v_{x,y,z,f}}{\Sigma^n p\left(z^t_{x,y,z}=k\ |x_{i)}\right.}$$

$$\left(\sigma^2_{kf}\right)^{t+1} = \frac{1}{m}\frac{\Sigma^n p\left(z^t_{x,y,z} = k\ |v_{x,y,z,f}\right)\left(v_{x,y,z,f}-\mu^{t+1}_{kf}\right)^2}{\Sigma^n p\left(z^t_{x,y,z} = k|v_{x,y,z,f}\right)}, f = 1,2,..m$$

end

$\underset{\beta\,\in B}{\text{argmax}}\, CH_{4D} = CH_{FPC} + CH_{spatial}$ **(**see Eq. 4 and 5)

Rad-FIT Clustering (β)

Calculate KL divergence metric from *k* and *k+1*:$\Delta_{k,k+1} = \sqrt{V}\left(\frac{\bar{\Delta}_v}{var\,(\Delta_v)}\right)$

end

while $\phi > 0.01$

calculate $\phi_i = \Delta_{i,i+1} - \Delta_{i+1,i+2}$

for $i \in 1 \ldots K-2$

7.6 Evaluating performance of U-Rad-FIT clustering

The performance of U-Rad-FIT clustering in identifying and segmenting the number of discrete functional regions was evaluated in each heterogeneity image and compared against the performance of K-means clustering with voxels represented both as time activity curves and as functional principal components (FPCs). Segmentation performance was evaluated using the Dice score, with generalized and region-based Dice scores evaluated for each image. K-means clustering was applied using the optimal number of clusters identified by U-Rad-FIT clustering to ensure robust comparisons.

Using the methodology outlined above for U-Rad-FIT clustering, the optimal number of functional clusters, in heterogeneity images 1 through 4 was determined to be three, five, six, and six, respectively. The overlap of spheres in each of the simulated images, with varying activity to background ratios, results in heterogeneous lesions with varying difficulties of separability.

When comparing the generalized segmentation performance of U-Rad-FIT clustering against K-means clustering with voxels represented as time activity curves and FPCs and segmenting the tumor into the same number of clusters, U-Rad-FIT clustering demonstrated the highest performance overall (Figure 7.5). All three segmentation approaches performed similarly on heterogeneity simulation 1. In heterogeneity simulations 2-4, which represented more challenging spatial and functional heterogeneity, U-Rad-FIT clustering demonstrated better performance.

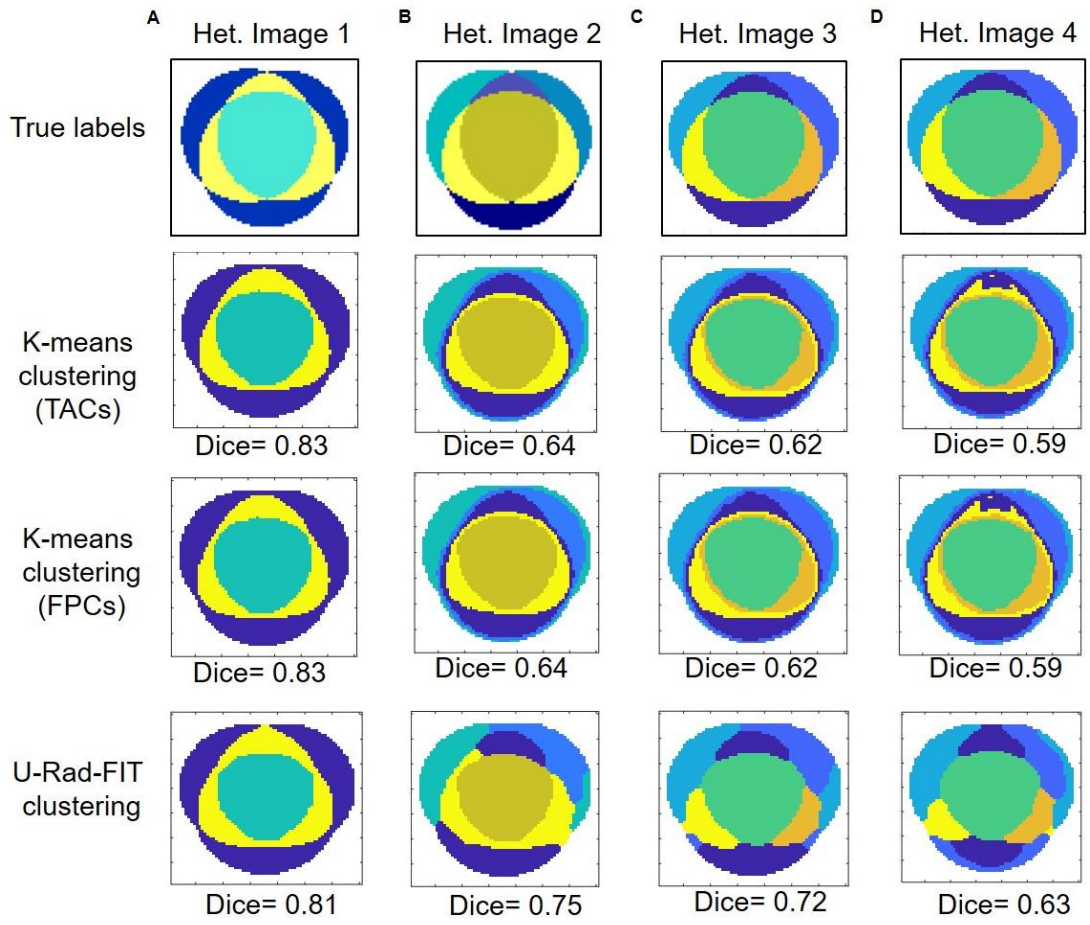
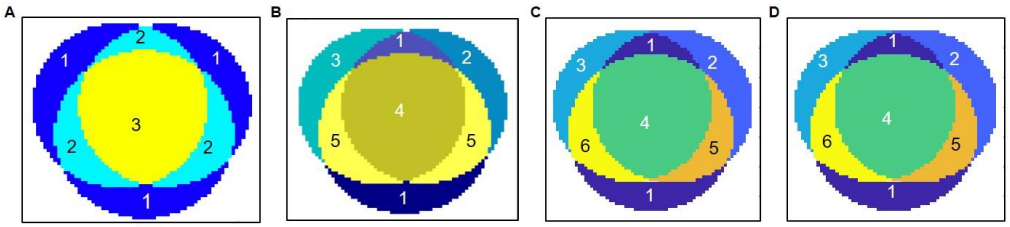


Figure 7.5. Generalized Dice score performance for each simulated heterogeneity image. True segmentation labels of 3 cluster-regions, K-means clustering results with voxels represented as TACs, K-means clustering results with voxels represented as FPCs, and U-Rad-FIT clustering results for Heterogeneity simulation 1 (A). True segmentation labels of 5 cluster-regions, K-means clustering results with voxels represented as TACs, K-means clustering results with voxels represented as FPCs, and U-Rad-FIT clustering results for Heterogeneity simulation 2 (B). True segmentation labels of 6 cluster-regions, K-means clustering results with voxels represented as TACs, K-means clustering results with voxels represented as FPCs, and U-Rad-FIT clustering results for Heterogeneity simulation 3 (C). True segmentation labels of 6 cluster-regions, K-means clustering results with voxels represented as TACs, K-means clustering results with voxels represented as FPCs, and U-Rad-FIT clustering results for Heterogeneity simulation 4 (D).

Comparing region-based segmentation performances on each the four heterogeneity simulations, against K-means clustering with voxels represented as FPCs, which was the higher performing control segmentation, U-Rad-FIT clustering demonstrated improved Dice scores for each region (Figure 7.6).



	Het. Image 1		Het. Image 2		Het. Image 3		Het. Image 4	
K-means clustering	Region 1	0.91	Region 1	0.56	Region 1	0.63	Region 1	0.62
	Region 2	0.67	Region 2	0.55	Region 2	0.56	Region 2	0.54
	Region 3	0.84	Region 3	0.69	Region 3	0.72	Region 3	0.68
			Region 4	0.86	Region 4	0.83	Region 4	0.84
			Region 5	0.56	Region 5	0.44	Region 5	0.39
					Region 6	0.53	Region 6	0.47
U-Rad-FIT clustering	Region 1	0.91	Region 1	0.76	Region 1	0.74	Region 1	0.56
	Region 2	0.71	Region 2	0.79	Region 2	0.67	Region 2	0.49
	Region 3	0.89	Region 3	0.73	Region 3	0.67	Region 3	0.62
			Region 4	0.82	Region 4	0.89	Region 4	0.90
			Region 5	0.64	Region 5	0.64	Region 5	0.62
					Region 6	0.68	Region 6	0.57

Figure 7.6. Region based Dice scores for each heterogeneity simulation image using K-means clustering with voxels represented as FPCs and U-Rad-FIT clustering.

7.7 Conclusion

This exploratory work demonstrates an initial extension of Rad-FIT clustering to be unsupervised, without a predefined number of sub-regions to segment. Initial experiments in developing U-Rad-FIT clustering suggests an improved segmentation performance when identifying and segmenting the true number of functionally discrete sub-regions as compared to conventional unsupervised clustering techniques. Also, preliminary experiments in selecting the optimal β parameter controlling the spatial influence in segmentation suggest that varying lesions may have a varying optimal parameter based on image signal to noise ratios. U-Rad-FIT clustering may be used in its originally developed format in which the number of sub-regions to identify is pre-

determined, or further validation of U-Rad-FIT clustering may result in identifying the optimal number of sub-regions in breast lesions. Further work is required to validate the identified sub-regions through ground truth histology/immunohistochemistry mapping for a physiologic interpretation of the kinetic variations displayed across the different sub-regions.

Potential applications of U-Rad-FIT clustering may include longitudinal monitoring of changes in the number of sub-regions or sub-region kinetic properties over the course of chemotherapy for detection early response or non-response to treatment, including the emergence of de-novo treatment resistance. Longitudinal changes in the number of sub-regions within a lesion may allow for non-invasive characterization of an increase or decrease in lesion heterogeneity, or for identification of more aggressive or resistant sub-clonal populations. Lastly, sub-regions identified using U-Rad-FIT clustering may be spatially registered to histological or other radiographic images for a multi-modal understanding of physiological differences across tumor sub-clones.

Chapter 8 : Extended applications of quantifying lesion heterogeneity and future directions

8.1 Future directions: extending applications to pre-cancerous lesions towards personalized medicine

While the scope of this thesis has been largely aimed towards exploring predictive and prognostic imaging biomarkers for invasive breast cancer, these methods to quantify lesion heterogeneity may be extended to invasive lesions in varying anatomic site and to pre-cancerous breast lesions. While the incidence rate of invasive breast cancer remains high for women, diagnoses of pre-invasive intra-ductal proliferations such as ductal in-situ carcinoma (DCIS), have increased as a result of increased breast cancer screening, and comprise approximately 25% of diagnosed breast cancers[239]. Treatment of DCIS largely follows that of invasive breast cancer: radiation treatment and/or endocrine therapy. However, it is estimated that less than 40% of diagnosed DCIS cases will progress to invasive disease[240]. As such, improved stratification of low-risk and high-risk DCIS cases is required to allow for personalized treatment strategies to reduce over treatment by local and systemic therapies. Current histopathologic biomarkers may be limited in stratifying low-risk vs high-risk cases due to inadequate sampling of disease burden and heterogeneity.

8.1.1 Discovery study cohort and feature generation

As a preliminary analysis to explore the potential of imaging phenotypes of DCIS heterogeneity, we retrospectively analyzed of cohort of 163 women from the previously conducted E4112 trial by the American College of Radiology Imaging Network (ACRIN). For each woman, DCE-MRI imaging was performed, with images acquired across 34 different sites following a standardized study protocol. The in-plane resolution of the images ranged from 0.35-1.25 mm/pixel, and were acquired on either 1.5T or 3T scanners. For this preliminary analysis, the pre-treatment, pre-operative images were evaluated for each woman. Clinical covariates including DCIS score, lesion grade, and comedonecrosis were available for each woman. First, all images were preprocessed using N4 bias field normalization[241]. For each case, the 3-D primary lesion was segmented by a trained radiologist from the first-post contrast image. To promote comparability in analysis across images from varying clinical sites, all images were resampled using a linear interpolation to the lowest resolution scan in the dataset. Within each lesion in the first-post contrast image, 95 radiomic features were extracted using the Cancer Imaging Phenomics Toolkit (CaPTk)[163]. All extracted features were z-score normalized across all tumors and features were sign normalized such that increased feature value indicates increased tumor heterogeneity. To avoid bias from non-uniform or heavily skewed features in this exploratory analysis, features were selected with the relatively strict criteria of having IQR >1 and skewness >5. This resulted in a total of 32 radiomic features.

8.1.2 Imaging phenotypes of DICS heterogeneity

To identify intrinsic imaging phenotypes of DCIS, unsupervised hierarchical clustering was performed on the extracted, 32-feature feature vector for tumors in the discovery cohort. The *k* clusters obtained from the unsupervised hierarchical clustering algorithm were interpreted as imaging phenotypes seen in the population. Clustering was performed using the Euclidean distance between tumor feature vectors and Ward's minimum variance method was used as the clustering criterion. Two statistically significant phenotypes were identified in the discovery cohort and found to be stable using Consensus Clustering[166] and significant via SigClust[167] methods ($p<0.001$) (Figure 8.1). Distributions of DCIS score, lesion grade, and necrosis were evaluated across the identified imaging phenotypes with necrosis found to be statistically significant ($p=0.02$) (Figure 8.2).

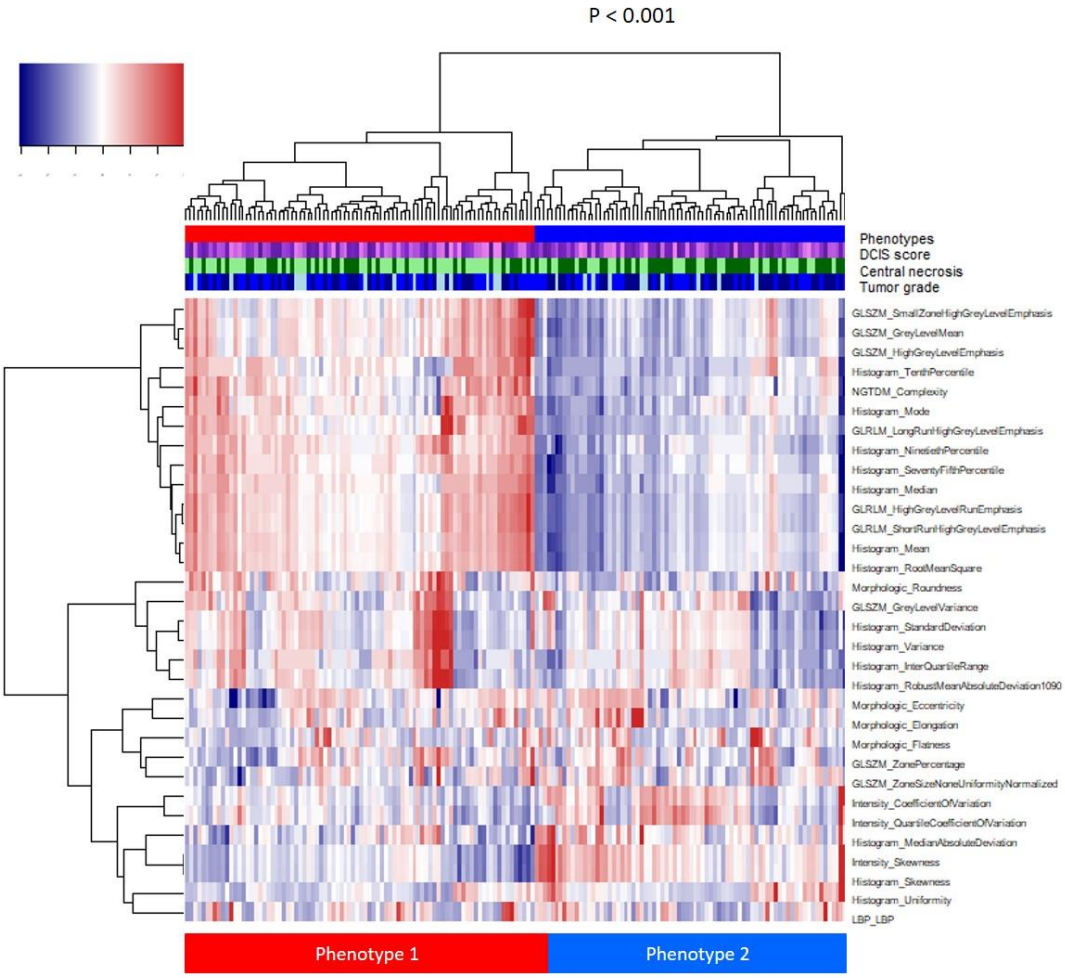


Figure 8.1. Identification of intrinsic imaging phenotypes of DCIS tumor heterogeneity. Unsupervised hierarchical clustering of SER features identifies three intrinsic phenotypes in the discovery cohort.

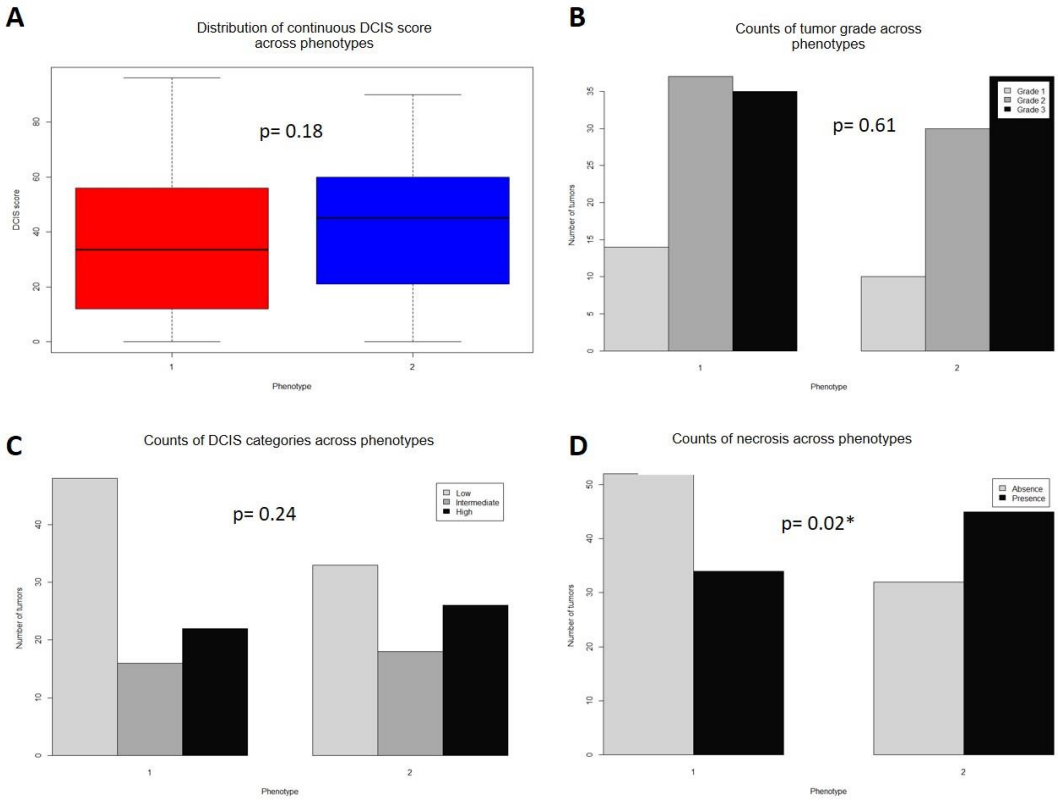


Figure 8.2. Associations between histopathologic prognostic markers and heterogeneity phenotypes identified in the discovery cohort. Distribution of DCIS scores across heterogeneity phenotypes (A). Number of grade 1, grade 2 and grade 3 lesions across heterogeneity phenotypes(B). Number of low, intermediate, and high DCIS scores across heterogeneity phenotypes (C). Number DCIS lesions with and without comedonecrosis across heterogeneity phenotypes (D).

8.1.3 Validation study cohort

A retrospective, exploratory validation cohort[242], provided by the University of Washington, was utilized in this analysis to validate the DCIS imaging phenotypes. The validation cohort consisted of 20 patients with DCIS who underwent preoperative DCE-MRI at the University of Washington between 2004 and 2014 with ipsilateral recurrence more than 6 months after definitive surgical treatment were retrospectively identified.

For each patient, a control subject with DCIS that did not recur was identified and matched on the basis of clinical, histopathologic, and treatment features including, age, menopausal status, presence of a high risk genetic mutation, DCIS Van Nuys Pathologic grade, presence of comedonecrosis, estrogen receptor status, final surgical margins, postsurgical endocrine therapy, and postsurgical radiation therapy. Therefore, of the 20 patients included in the validation cohort, 10 patients had a future event of recurrence and 10 did not. Median time to recurrence was 14 months, and median follow-up for control subjects was 102 months.

8.1.4 Validation of imaging phenotypes of DCIS lesion heterogeneity

In order to validate the imaging phenotypes of DCIS lesion heterogeneity identified in the discovery cohort, the first post-contrast DCE-MRI images from the validation cohort were preprocessed as detailed above. From the resampled first post-contrast lesions, the same 32 radiomic features as selected from the discovery cohort were extracted using CapTK. Features in the validation cohort were first z-score normalized across tumors and sign normalized such that an increasing feature value corresponded to an increase in lesion heterogeneity. Unsupervised hierarchical clustering was then applied to the validation cohort tumors using the 32-feature feature vector. Two significant imaging phenotypes were also identified in the validation cohort through clustering analysis and determined to be stable and significant ($p=0.0001$) using ConsensusClustering and SigClust methods, respectively. The in-group proportion (IGP) statistic[243] was used to determine whether the imaging phenotypes identified in the discovery cohort existed in the validation cohort. Imaging phenotypes were found to have

high consistency between the discovery and validation cohorts with IGP values of 90% and 99% for Phenotype 1 and Phenotype 2, respectively. The IGP for Phenotype 2 was found to be significant (p=0.02) (Figure 8.3).

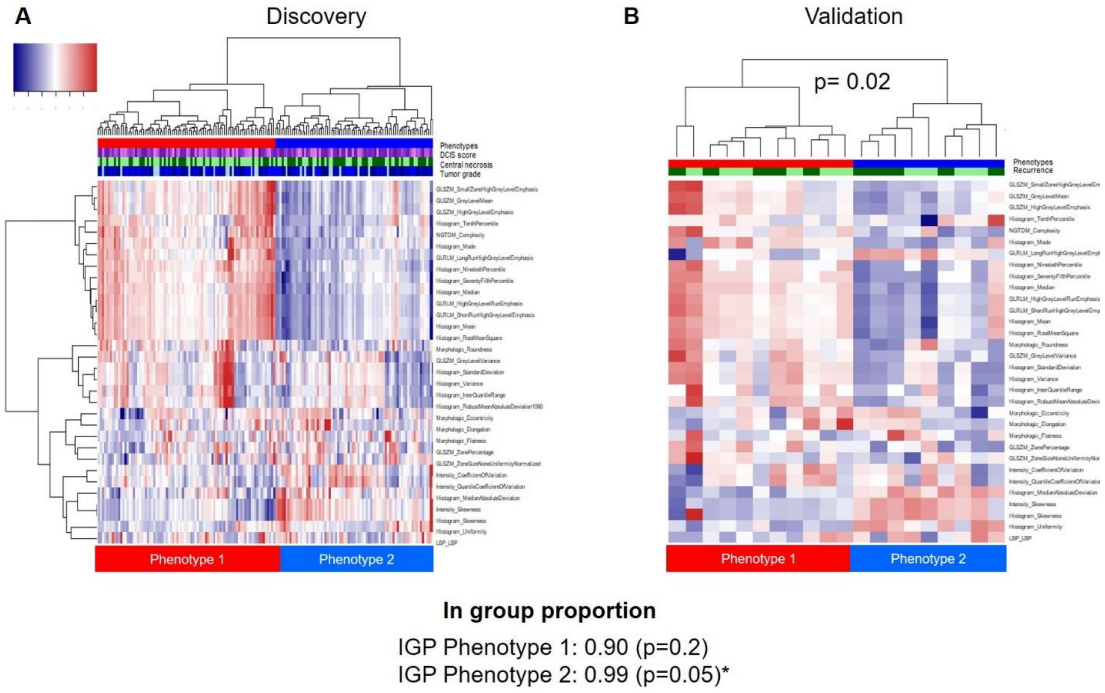


Figure 8.3. Independent validation of intrinsic imaging phenotypes of tumor heterogeneity. Phenotypes identified in the discovery cohort (A) are significantly reproducible in the validation cohort (B).

8.1.5 Associations between radiomic features and recurrence

In order to explore the prognostic value of imaging features in predicting recurrence case/control for the validation set, we performed univariate logistic regression on features found to be significant across the imaging phenotypes using the Significance Analysis of Microarrays (SAM) algorithm[198]. SAM is a nonparametric statistical algorithm designed to identify significant variables associated with a specific trait (e.g. phenotype assignment). This exploratory analysis resulted in three imaging features, histogram variance (AUC = 0.72), histogram uniformity (0.74), and morphologic roundness (AUC= 0.70) demonstrating high discriminatory capacity for predicting case from control in the 20 tumors, case/control matched for clinical and histopathological covariates, in the validation cohort (Figure 8.4).

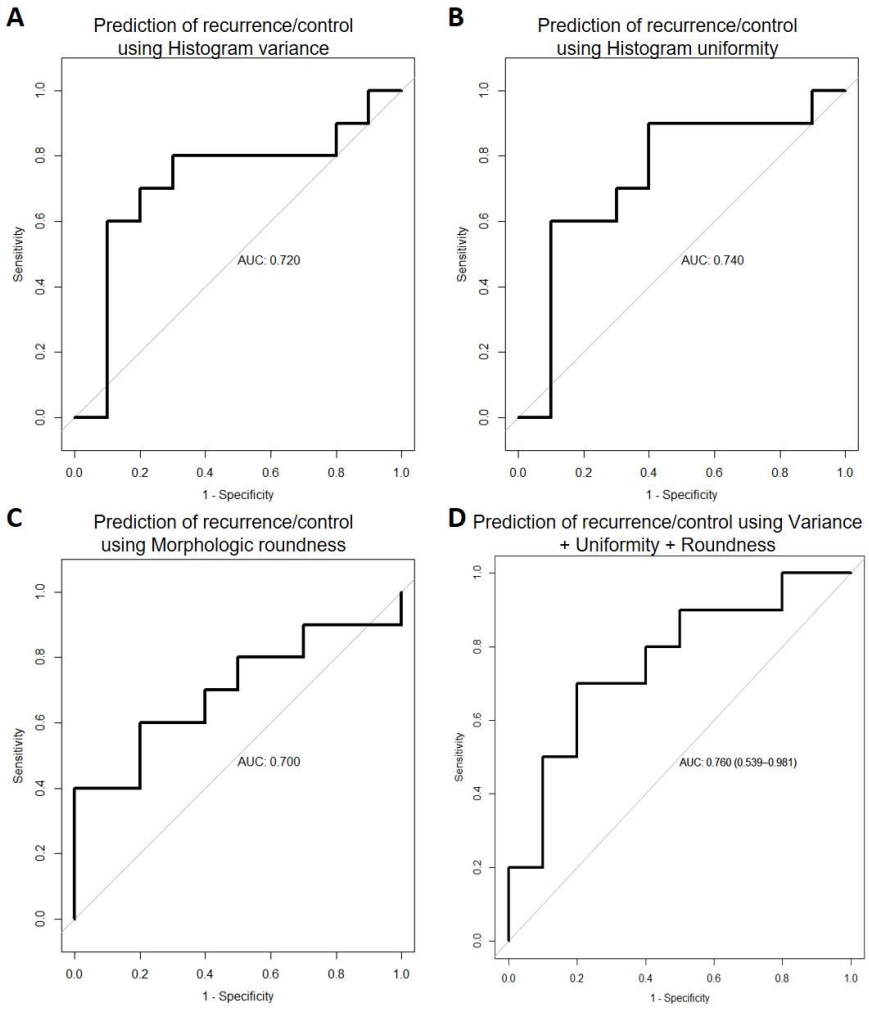


Figure 8.4. Independent prognostic value of radiomic features from the validation cohort when predicting recurrence case from matched controls. ROC curve for Histogram variance (A), ROC curve for histogram uniformity (B), ROC curve for Morphologic roundness (C), and ROC curve for combined prediction using variance, uniformity, and roundness (D).

8.2. Future directions: Feature harmonization

In order to explore the generalizability of radiomic features extracted from images with varying voxel resolutions and from varying scanner modalities (e.g. manufacturer, field strength), ComBat analysis may be subsequently employed to harmonize extracted radiomic features. ComBat[244,245] is a harmonization technique originally developed for

genomic studies that aims to correct for variation in radiomic features due to image parameter related batch effects using empirical Bayes to estimate location and scale parameters to eliminate such variation. The impact of radiomic feature variation introduced by image resolution and acquisition parameters such as field strength may be mitigated using a nested ComBat[246] approach. Future analysis for this work could examine the potential effects of harmonized features when performing unsupervised hierarchical clustering, and further explore the generalizability of harmonized features when extending the imaging phenotypes to independent validation cohorts.

8.2 Future directions: Radiogenomic analysis

Radiogenomic analysis aims to leverage high-throughput genomic, molecular, and sequencing data alongside high-throughput imaging data towards the advancement of cancer precision care. In order to further understand the biological basis of imaging phenotypes of tumor heterogeneity, one could explore potentially differential genomic and proteomic expressions across imaging phenotypes to better understand the physiologic mechanisms driving tumor image heterogeneity. This work[196] has identified associations between imaging phenotypes of spatial heterogeneity, which may capture differential vascularization and angiogenic properties across the tumor lesion, and ER percentage, cell differentiation, and mitotic stage. Future work may aim to further illuminate more mechanistic properties of such spatial heterogeneity through personalized genomic analyses. Additionally, the complementary information provided by radiogenomic analyses may further improve prognostic performance when using an

augmented model that integrates imaging biomarkers with established and emerging molecular prognostic markers in stratifying women at higher risk of tumor recurrence, as demonstrated in Chapter 4 of this work, thus allowing for improved personalization of treatment strategies. Future work towards this aim may utilize a larger prospective trial in which DCE-MRI images are acquired for woman diagnosed with locally advanced breast cancer. Concurrently, FFPE primary tumor blocks may be analyzed for RNA extraction and subsequent RNA-Seq data processing. Such analyses could allow for associations between imaging phenotypes and biomarkers and intrinsic molecular biomarkers including Claudin-low phenotypes, or PAM50 derived subgroups.

8.3 Future directions: Histopathologic spatial mapping

The validation of Rad-FIT clustering using simulated image phantoms allowed for method development and validation, as well as demonstrated the added value of incorporating spatial and temporal information for tumor voxel clustering as compared to conventional methods utilizing spatial or temporal information, alone. Future work to further understand the physiological differences in identified tumor sub-regions could incorporate whole tissue histopathological analysis and leverage multi-modality spatial registration to spatially align dynamic PET images with digital pathology images. Such work could leverage the information afforded by histopathologic analysis to explore the biological differences between Rad-FIT identified sub-regions and such longitudinal analysis may provide insight into how the metabolic demands of tumor sub-populations change in response to treatment. Furthermore, incorporation of spatially mapped-histopathologic images would allow for biological validation of imaging heterogeneity

beyond that of biopsy derived histopathologic clinical covariates. This would have the potential to explore whether spatially corresponding imaging and histopathologic biomarkers capture similar physiology at varying resolutions or capture complementary information. Lastly, such mapping could allow for insight into metabolic heterogeneity within the tumor microenvironment as seen in FDG dynamic PET imaging.

8.4 Future directions: Multi-modality image analysis and deep learning

Medical imaging has the potential to be clinically leveraged as a non-invasive assay during breast tumor diagnosis and treatment monitoring. Leveraging the differential physiologic insights provided through modalities such as MRI and PET can allow for whole-tumor analyses of targeted biology. Registering PET images to DCE-MR images can localize regions of specific biological activation, as seen using targeted radiotracers in PET, from which quantitative image features can be extracted from DCE-MR images. This would provide an *in vivo* validation for heterogeneity features extracted from sub-regions within the tumor that may differ in underlying biology. Additionally, this methodological development could be extended to understand the mechanisms of additional types of cancer formation and development.

Work presented in this thesis largely focused on leveraging "hand-crafted" features, either summarizing spatial heterogeneity through computer vision derived features, or through features developed in order to summarize 4-D heterogeneity through tumor sub-region analysis. Future work in exploring 4-D tumor heterogeneity could leverage deep learning techniques in a myriad of ways. For example, autoencoders may

be utilized to better summarize dynamic PET voxel time activity behavior by projecting information non-linearly into a lower dimensional space, thereby providing additional insight in identifying functionally discrete tumor sub-regions. Additionally, deep learning visualization techniques, including deconvolutional networks[247], may be used following a model trained to predict recurrence free survival using longitudinal imaging, to provide insight into more aggressive tumor regions driving recurrence and highlight what aspects of tumor imaging properties are most predictive of this response. Such work may be used to drive targeted radiation treatments with additional emphasis placed on these spatial regions. Lastly, deep learning models could be utilized for non-linear combinations of multi-modality imaging features or radiogenomic datasets for the development of improved prognostic and predictive imaging biomarkers.

Chapter 9 : Conclusions

Precision medicine aims to tailor disease prognosis and treatment based on specific genotypic and phenotypic characteristics of an individual. Advances in genomics, proteomics, and high-throughput sequencing have allowed for a personalized understanding of breast tumor behaviors. Medical imaging is uniquely suited for characterizing lesion heterogeneity through the ability to non-invasively image the entire disease burden, and is routinely acquired throughout breast cancer care.

Characterizing intratumor heterogeneity may allow for personalized prognostic and predictive biomarkers and provide complementary information to other precision medicine assays, allowing for tailored patient treatment. Identifying imaging phenotypes of tumor heterogeneity may allow for improved patient prognostic and predictive stratification. Characterizing imaging representations of heterogeneous tumor sub-clones may be used to select informative biopsy sites or direct treatment strategies. This work demonstrates that computationally derived imaging biomarkers can provide non-invasive, quantitative insight into breast tumor biology.

First, we identified imaging phenotypes of intratumor spatial heterogeneity in DCE-MRI images of invasive breast cancer using radiomic analysis, as outlined in Chapter 3. We reproduced and validated these imaging phenotypes in an independent, publicly available cohort. We then evaluated the prognostic value of such imaging phenotypes in predicting 10-year recurrence free survival (RFS), and found that tumors

that had increased amounts of spatial heterogeneity prior to treatment or surgery, had an increased likelihood of tumor recurrence. This may support the hypothesis that heterogeneous tumors are associated with more aggressive tumor behavior and treatment resistance. Furthermore, we found that high degrees of imaging heterogeneity were associated with poorly defined tumors and tumors with higher mitotic grades, potentially contributing to genetic diversity and sub-clonal evolution within these primary lesions. Lastly, we determined that imaging phenotypes augmented prognostic models utilizing conventional histopathologic biomarkers.

We then developed imaging biomarkers characterizing longitudinal changes in spatial heterogeneity as presented in DCE-MRI images, as described in Chapter 4. Here we calculated imaging features characterizing changes in the kinetic behaviors of primary invasive breast lesions from baseline to an early-time point during the course of neoadjuvant chemotherapy. These imaging features characterized heterogeneity induced by a tumor's change in vasculature or angiogenic properties in response to treatment. We identified two intrinsic imaging phenotypes of such change in heterogeneity and determined the prognostic value of these phenotypes when adding phenotype assignment to established personalized molecular signatures. Interpretation of the two radiomic phenotypes as capturing an increase and decrease in intratumor heterogeneity from pre-treatment to early-treatment showed that tumors assigned to the phenotype with increasing intratumor heterogeneity had a greater number of future recurrences. Additionally, with the stratification of women within functional tumor volume (FTV), a previously described and broadly established imaging prognostic biomarker, subgroups

by phenotype demonstrated the added value of radiomic analysis of change in intratumor heterogeneity in modeling prognosis. We illuminated that augmenting clinical, histopathologic, and molecular covariates with imaging phenotypes may allow for personalized risk stratification and early adaptation of treatment strategies.

Together, this work demonstrates novel phenotypic differences in clinical imaging representations of primary breast tumors at diagnosis as well as phenotypic differences in how tumor imaging presentations of heterogeneity change in response to treatment. This work also explored prognostic and predictive clinical imaging biomarkers to character this heterogeneity. These imaging biomarkers captured spatial and temporal heterogeneity through dynamic clinical imaging modalities using baseline diagnostic clinical imaging and also explored longitudinal changes in tumor heterogeneity in response to neoadjuvant treatment.

We then further aimed to characterize spatial and temporal heterogeneity in primary breast lesions by identifying intratumor regions of varying functional behaviors. We leveraged the improved dynamic image sampling afforded through dynamic PET imaging to understand the potentially differential kinetic patterns across a single tumor following contrast tracer injection. To do so, we developed a tracer agnostic segmentation algorithm, Radiomic Functional Intratumor (Rad-FIT) clustering, that identifies spatially constrained, functionally discrete sub-regions. This algorithm can be applied to non-rigid, arbitrarily shaped regions without manual initialization or defined control points. Development and validation of this segmentation algorithm in Chapter 5 demonstrated improved performance over established unsupervised clustering algorithms when

segmenting simulated image phantoms. The results from these experiments showed that a segmentation algorithm incorporating both spatial and temporal information had improved 4-D segmentation accuracy over established methods using only temporal dynamics or spatial information, alone.

Using this developed 4-D segmentation algorithm, we identified three sub-regions within locally advanced breast lesions imaged with dynamic FDG-PET in a retrospective analysis. In order to summarize the metabolic heterogeneity within each tumor by characterizing the three sub-regions, we developed a functional tumor heterogeneity (FTH) imaging signature describing how compact and separate the detected sub-regions were. We found that quantifying intratumor metabolic heterogeneity using this FTH imaging signature augmented established histopathologic covariates in Cox-regression models predicting RFS. This work, outlined in Chapter 6, suggested that imaging characteristics of 4-D metabolic functional heterogeneity may be leveraged as a prognostic biomarker by providing information not previously identified by ROI-based kinetic analyses.

In Chapter 7 we explored extending Rad-FIT clustering to be completely unsupervised, or Unsupervised Rad-FIT (U-Rad-FIT) clustering. We explored methods to determine the optimal number of spatially constrained functionally discrete sub-regions within a single lesion towards developing precision medicine-based imaging biomarkers. We evaluated algorithm performance using novel dynamic imaging phantoms simulating a heterogeneous lesion.

Lastly, we explored the extension of the techniques and imaging biomarkers developed throughout this thesis to images of pre-cancerous breast lesions and discussed future directions and ongoing experiments for the continued development and validation of imaging biomarkers of intratumor heterogeneity.

This thesis provides evidence that imaging biomarkers have the potential to be utilized towards precision medicine for cancer care. While this work focused on breast cancer, future work should also seek to evaluate the developed approaches to other cancer sites. Further developments in the standardization, interpretation, and validation of such biomarkers are required for ultimately translating quantitative imaging biomarkers from development in a research setting to implementations in oncologic clinical practice.

Appendix A: Radiomic features

Table A.1. Gray-level histogram texture features

Gray-level histogram features	**Qualitative description**	**Mathematical description**
Mean	Mean gray-level value	$\sum_k \frac{k*g(k)}{\sum_k g(k)}$
Min	Min gray-level value	Min(*k*)
Max	Max gray-level value	Max(*k*)
5th Percentile	Histogram bin that 5% of grey level values are less than or equal to	k: 5% of values ≤ *k*
Mean 5th	Mean value of gray-level values that 5% of grey level values are less than or equal to	$\frac{\sum_k k*g(k)}{\sum_k g(k)}$ for *k* ≤ fifth percentile
95th Percentile	Histogram bin that 95% of grey level values are greater than or equal to	k: 95% of values ≥ *k*
Mean 95th	Mean value of gray-level values that 95% of grey level values are greater than or equal to	$\frac{\sum_k k*g(k)}{\sum_k g(k)}$ for *k* ≥ ninety- fifth percentile
Sum	Sum of gray-level values	$\sum_k k * g(k)$
Sigma	Measure of variation of gray-level values around the mean	$\sqrt{\sum_k (k-\mu)^2 * g(k)}$
Entropy	Measure of histogram nonuniformity	$-\sum_k g(k) * \log(g(k))$
Kurtosis	Measure of histogram flatness	$\sigma^{-4}\sum_k (k-\mu)^4 * g(k) - 3$
Skewness	Measure of histogram symmetry	$\sigma^{-3}\sum_k (k-\mu)^3 * g(k)$

Table A.2. Representative gray-level co-occurrence matrix texture features

Co-occurrence matrix features	**Qualitative description**	**Mathematical description**
Contrast	Intensity contrast between pixel and its neighbor	$\sum_{ij} \lvert i-j \rvert^2 f(i,j)$
Correlation	Linear gray-level dependence	$\sum_{ij} \frac{\left((i-\mu_i) * (j-\mu_j) * f(i,j)\right)}{\sigma_i \sigma_j}$
Homogeneity	Closeness of distribution in co-occurrence matrix to matrix diagonal	$\sum_{ij} \frac{f(i,j)}{1+\lvert i-j \rvert}$
Energy	Certainty of gray-level co-occurrence	$\sum_{ij} f(i,j)^2$
Entropy	Uncertainty of gray-level co-occurrence	$-\sum_{ij} f(i,j) * \log\left(f(i,j)\right)$
Inverse Difference Moment (IDM)	Local homogeneity in gray-level co-occurrence	$\sum_{ij} \frac{f(i,j)}{(1+(i-j)^2)}$
Cluster Shade	Asymmetry in gray-level values	$\sum_{ij} \left(i-\mu_i+j-\mu_j\right)^3 * f(i,j)$

Table A.3. Run-length texture features

Run-length features		
Short run emphasis (SRE)	Emphasis on short runs	$1/n_r \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{(R(i,j))}{j^2}$
(Long run emphasis (LRE)	Emphasis on long runs	$1/n_r \sum_{i=1}^{M} \sum_{j=1}^{N} (R(i,j)) * j^2$
Gray level nonuniformity (GLN)	Degree of gray-level run dissimilarity	$1/n_r \sum_{i=1}^{M} (\sum_{j=1}^{N} R(i,j))^2$
Run length nonuniformity (RLN)	Dissimilarity in length of runs	$1/n_r \sum_{j=1}^{N} (\sum_{i=1}^{M} R(i,j))^2$
Run percentage (RP)	Distribution of runs	$\frac{n_r}{\#pixels}$
Low gray level run emphasis (LGRE)	Emphasis on low gray-level values	$1/n_r \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{(R(i,j))}{(i+j)^2}$
High gray level run emphasis (HGRE)	Emphasis on high-gray-level values	$1/n_r \sum_{i=1}^{M} \sum_{j=1}^{N} (R(i,j)) * i^2$
Short run low gray level emphasis (SRLGE)	Emphasis on short runs with low-gray-level values	$1/n_r \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{(R(i,j))}{j^2 (i+j)^2}$
Short run high gray level emphasis (SRHGE)	Emphasis on short runs with high-gray-level values	$1/n_r \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{(R(i,j))(i+j)^2}{j^2}$

nr is the total number of runs, R(*i*,*j*) represents the number of runs with pixels of gray-level intensity value, *i*, and length of run, *j*. 128 gray-levels were used. Estimated by averaging over 0°, 45°, 90°, and 135° orientations.

Table A.4. Structural texture features

Structural feature	**Qualitative description**
Local Binary Pattern (LBP)	Intensity variation between a pixel and its neighboring pixels.

Mathematical description

$$LBP\,(x_c, y_c) = \sum_{p=0}^{P-1} q(I_p - I_c)2^P,\ (x_p, y_p) = [x_c + Qcos\left(\frac{2\pi p}{P}\right), y_c - Qsin\left(\frac{2\pi p}{P}\right)]$$

I_c and I_p are gray-level intensity values for pixel (x_c,y_c) and pixel (x_p,y_p). q= indicator function, 0 for negative inputs and 1 for non-negative inputs.Q, P= parameters to set pixel neighborhood size, set to 1 and 8, respectively[39,40].

Appendix B: Personalized molecular signatures

Table B.1. Recreated MammaPrint classification results compared to the original Esserman et al. classification results.

	Low risk	**High risk**
Esserman's overall results	11	109
Recreated overall results	10	108
Esserman's PCR (positive) results	0/11	25/105
Recreated PCR (positive) results	2/10	26/108
Esserman's RCB (0 or I class) results	1/9	31/99
Recreated RCB (0 or I class) results	2/10	36/108

Table B.2. Recreated p53 mutation gene signature classification results compared to the original Esserman et al. classification results.

	p53 wildtype	**p53 mutant**
Esserman's overall results	59	61
Recreated overall results	57	61
Esserman's PCR (positive) results	5/58	20/58
Recreated PCR (positive) results	8/57	20/61
Esserman's RCB (0 or I class) results	10/53	22/55
Recreated RCB (0 or I class) results	15/57	23/61

Table B.3. Recreated PAM50 ROR-S gene signature classification results compared to the original Esserman et al. classification results.

	Low risk	**Intermediate risk**	**High risk**
Esserman's overall results	32	42	46
Recreated overall results	31	41	46
Esserman's PCR (positive) results	2/32	7/40	16/44
Recreated PCR (positive) results	3/31	11/41	14/46
Esserman's RCB (0 or I class) results	5/28	9/38	18/42
Recreated RCB (0 or I class) results	6/31	14/41	18/46

B.1. Molecular signatures

Gene expression data were obtained from Gene Expression Omnibus (GEO)[185,186] using the publicly available samples from the Esserman et al. study (accession GSE22226)[187] that match the ACRIN 6657/I-SPY 1 MRI data of the discovery cohort TCIA dataset. Samples have been analyzed using two microarray platforms and can be found in GEO under accessions GPL1708 (n = 130) and GPL4133 (n = 20). Initially, gene signature sets were validated by recreating the Esserman et. al study. Afterwards, the final gene signatures set were used solely and in combination with the MRI information of patients to assess their classifying value.

Recreation of Esserman et al. study

Samples were filtered as in Esserman et al.[187] down to n = 120 (initially 121, but one entry was removed due to incomplete data). Briefly, only patients with both microarray expression data and HR/HER2 status, RCB and negative *trastuzumab* treatment status were kept in the final set. *Trastuzumab* status was taken directly from the GEO phenotype information, while RCB class and HR/HER2 status were extracted using the Clinical and Outcome Data found at:

https://wiki.cancerimagingarchive.net/display/Public/ISPY1. Minimal differences in clinical data may be explained due to different updates in the various data sources, as the ACRIN 6657/I-SPY 1 clinical trial was an extended, long-term prospective study. Microarray intensity values provided in GSE22226 are expressed as log_2 Lowess-normalized mean ratio values.

Molecular profiling was built using three of the four gene signatures that are mentioned in Esserman et al. study i.e. 70-gene signature (MammaPrint)[190,191], p53 mutation signature[194] and PAM50 risk of recurrence (ROR-S)[192,193]. The wound-healing response gene signature[248] recreation proved to be unsuccessful due to unavailability of the original supplementary data and microarray probes that comprise the signature gene set. Annotation of the probes was extracted automatically from the latest GEO platform annotation found in the NCBI GEO repository.

B.2. MammaPrint

The original experiment consisted of 98 primary breast cancer human samples, of which 78 were metastasis-free. 44/78 were recurrence-free for more than 5 years, constituting the "good prognosis group". In short, the authors performed hierarchical clustering using significantly regulated genes (two-fold change, $p < 0.01$), Pearson correlation coefficient calculation between the prognostic category and the log expression ratio across all samples and Monte Carlo randomization of the association between the expression ratio and prognosis category to discern the best candidate genes ($n = 231$) to predict recurrence-free survival (RFS). After leave-one-out cross validation for different subsets of genes, the authors ended up with the 70-gene signature.

Classification is achieved by calculating the cosine similarity between the MammaPrint gene signature expression values of the sample to be classified and the average of the MammaPrint gene signature expression values of the "good prognosis group" in the original study. A value greater than -0.4 (threshold suggested by the MammaPrint authors) is considered to classify the sample as "high risk" with respect to future recurrence[191].

The MammaPrint 70-gene signature consists of 70 microarray probes. Original analysis was performed using 25K human oligonucleotide two-color microarrays developed by Rosetta. The 70 probes correspond to 56 genes and 14 Expressed Sequence Tags (ESTs). Expression values provided in supplementary data are calculated as log_{10} mean ratio with median background intensity subtraction. Specifically:

$$expression\ value = log_{10} \frac{Channel_2 mean\ intensity - median(background\ intensity)}{Channel_1 mean\ intensity - normalized\ median(background\ intensity)}$$

In our study, the cosine similarities were calculated using the genefu R package[249]. Probe matching between Esserman et al. microarray annotation and MammaPrint was done using Entrez IDs, resulting into a shrinkage of the probes number (52/70). A possible explanation for this is due to updated annotation information related to the genes overlapping those probes. Furthermore, in order to make the two platform intensities comparable, the Esserman et al. expression values were recalculated to match the MammaPrint values. $-\infty$ and NA values (introduced due to background subtraction) were converted to 0 prior to cosine similarity estimation.

B.3. p53 mutation signature

The p53 mutation gene signature consists of 52 microarray probes derived from the unsupervised clustering of datasets with known p53 mutation status, which is used to classify samples' status as p53 wildtype or p53 mutant. Tumor suppressor p53 mutations are found more frequently in aggressive breast cancers. In the original study, a SAM[250] derived gene list along with a false discovery rate of less than 5% was given as input to an average linkage hierarchical cluster analysis. The analysis was conducted using Pearson correlation in the Cluster program. This gene list was then refined by comparing the p53-associated gene lists between tumor samples and cell lines, leading to a robust list of 52 genes that were common to both data sets (*in vitro* and *in vivo*).

According to Troester et al.[194], classification was performed by calculating the Spearman's correlation metric between the samples under examination and the training

set centroids, similar to the MammaPrint method. Centroids are represented as vectors of the mean average of each gene expression in the p53 signature, one per p53 status group (wildtype *vs.* mutant). Greater correlation to one of the centroids classifies the sample as such.

The training dataset used to create the centroids consists of 66 microarray samples from Sorlie et al. [251] (2001), which are deposited in GEO under accession GSE3193. In the experiment, 4 different microarray platforms were used i.e. GPL180, GPL2776, GPL2777, GPL2778. Microarray platform/source systematic biases between them were originally corrected using the Distance Weighted Discrimination (DWD) algorithm. Due to lack of an R package that implements DWD, in our study we used the ComBat function found in the SVA R package[252], which utilizes an Empirical Bayes approach. Three groups were designated for batch removal (A: Sorlie et al. (2001) data, B: Esserman's GPL1708 data, C: Esserman's GPL4133 data). Non-finite values were ignored.

47/52 probes were common in all datasets by matching their Entrez IDs, which comprised the final p53 gene signature used for the recreation analysis. The same rationale as in the MammaPrint section applies here: log_{10} and log_2 mean ratio and median background intensity subtraction were used in the different Sorlie et al. (2001) microarray datasets, thus every raw intensity was transformed into log_2 mean ratio values, including the microarray intensity values from Esserman et al..

B.4. PAM50 ROR-S

Risk of recurrence score is used to classify patients into three categories (low, medium and high risk of relapse) according to an estimated risk value, using predefined thresholds. The original study progressively identified a 50-gene set through hierarchical clustering and the single sample predictor algorithm (SSP), which was used to cluster 189 breast cancer and 29 normal samples into 5 intrinsic subtypes (Luminal A, Luminal B, HER2-enriched, Basal-like and Normal-like) by employing the Prediction Analysis of Microarray (PAM) centroid-based clustering algorithm. Distance to each subtype was calculated using Spearman's rank correlation, representing the proximity of each sample to each category. Then, the authors performed univariate and multivariate analyses to determine the significance of those subtypes and trained a multivariable Cox model using the Ridge regression fit to the node-negative, untreated subset of the van de Vijver cohort[253]. In order to classify a sample into one of the risk categories, one needs to calculate the resulting weighted sum of the intrinsic subtype Spearman's rank coefficients using the following equation:

$$ROR - S = 0.05 * basal + 0.12 * HER2 + (-0.34 * LumA) + 0.23 * LumB$$

We utilized the genefu R package (rorS function) to classify our samples. Probe matching between the two gene sets was successfully done using Entrez IDs, which resulted in a complete 50/50 annotation. As in MammaPrint, the two platforms' intensities were incomparable. Thus, the Esserman et al. intensities were recalculated with no background subtraction and no Lowess normalization and expressed using log_2 mean ratios. Predefined ROR-S thresholds were used i.e. low $\leq$ 29, 29 $<$ moderate $<$ 53, high $\geq$ 53.

BIBLIOGRAPHY

1 DeSantis, C. E., Ma, J., Sauer, A. G., Newman, L. A. & Jemal, A. Breast cancer statistics, 2017, racial disparity in mortality by state. *CA: a cancer journal for clinicians* **67**, 439-448 (2017).

2 Kamangar, F., Dores, G. M. & Anderson, W. F. Patterns of cancer incidence, mortality, and prevalence across five continents: defining priorities to reduce cancer disparities in different geographic regions of the world. *Journal of clinical oncology* **24**, 2137-2150 (2006).

3 Oh, J. L. in *Methods of cancer diagnosis, therapy and prognosis* 583-587 (Springer, 2008).

4 Polyak, K. Heterogeneity in breast cancer. *The Journal of clinical investigation* **121**, 3786-3788 (2011).

5 Polyak, K. Breast cancer: origins and evolution. *The Journal of clinical investigation* **117**, 3155-3163 (2007).

6 Marusyk, A. & Polyak, K. Tumor heterogeneity: causes and consequences. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* **1805**, 105-117 (2010).

7 Sala, E. *et al.* Unravelling tumour heterogeneity using next-generation imaging: radiomics, radiogenomics, and habitat imaging. *Clinical radiology* **72**, 3-10 (2017).

8 O'Connor, J. P. *et al.* Imaging intratumor heterogeneity: role in therapy response, resistance, and clinical outcome. *Clinical Cancer Research* **21**, 249-257 (2015).

9 Eskey, C. J., Koretsky, A. P., Domach, M. M. & Jain, R. K. 2H-nuclear magnetic resonance imaging of tumor blood flow: spatial and temporal heterogeneity in a tissue-isolated mammary adenocarcinoma. *Cancer research* **52**, 6010-6019 (1992).

10 Padhani, A. R. & Miles, K. A. Multiparametric imaging of tumor response to therapy. *Radiology* **256**, 348-364 (2010).

11 Andor, N. *et al.* Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nature medicine* **22**, 105 (2016).

12 Morris, L. G. *et al.* Pan-cancer analysis of intratumor heterogeneity as a prognostic determinant of survival. *Oncotarget* **7**, 10051 (2016).

13 Turashvili, G. & Brogi, E. Tumor heterogeneity in breast cancer. *Frontiers in medicine* **4**, 227 (2017).

14 Weigel, M. T. & Dowsett, M. Current and emerging biomarkers in breast cancer: prognosis and prediction. *Endocrine-related cancer* **17**, R245-R262 (2010).

15 Onega, T. *et al.* Breast cancer screening in an era of personalized regimens: A conceptual model and National Cancer Institute initiative for risk-based and preference-based approaches at a population level. *Cancer* **120**, 2955-2964 (2014).

16 Gavenonis, S. C. & Roth, S. O. Role of magnetic resonance imaging in evaluating the extent of disease. *Magnetic Resonance Imaging Clinics* **18**, 199-206 (2010).

17 Weinstein, S. & Rosen, M. Breast MR imaging: current indications and advanced imaging techniques. *Radiologic Clinics* **48**, 1013-1042 (2010).

18 Marinovich, M. L. *et al.* Meta-analysis of magnetic resonance imaging in detecting residual breast cancer after neoadjuvant therapy. *Journal of the National Cancer Institute* **105**, 321-333 (2013).

19 Lebron, L., Greenspan, D. & Pandit-Taskar, N. PET imaging of breast cancer: role in patient management. *PET clinics* **10**, 159-195 (2015).

20 Hylton, N. M. Vascularity assessment of breast lesions with gadolinium-enhanced MR imaging. *Magnetic resonance imaging clinics of North America* **9**, 321-332, vi (2001).
21 McDonald, E. S., Clark, A. S., Tchou, J., Zhang, P. & Freedman, G. M. Clinical diagnosis and management of breast cancer. *J Nucl Med* **57**, 9S-16S (2016).
22 Dunnwald, L. K. *et al.* PET tumor metabolism in locally advanced breast cancer patients undergoing neoadjuvant chemotherapy: value of static versus kinetic measures of fluorodeoxyglucose uptake. *Clinical Cancer Research* **17**, 2400-2409 (2011).
23 Humbert, O. *et al.* Breast cancer blood flow and metabolism on dual-acquisition 18F-FDG PET: correlation with tumor phenotype and neoadjuvant chemotherapy response. *Journal of Nuclear Medicine* **59**, 1035-1041 (2018).
24 McInerney, T. & Terzopoulos, D. Deformable models in medical image analysis: a survey. *Medical image analysis* **1**, 91-108 (1996).
25 Castellano, G., Bonilha, L., Li, L. & Cendes, F. Texture analysis of medical images. *Clinical radiology* **59**, 1061-1069 (2004).
26 Ganeshan, B. & Miles, K. A. Quantifying tumour heterogeneity with CT. *Cancer imaging* **13**, 140 (2013).
27 Davatzikos, C. *et al.* Precision diagnostics based on machine learning-derived imaging signatures. *Magnetic resonance imaging* **64**, 49-61 (2019).
28 Gillies, R. J., Kinahan, P. E. & Hricak, H. Radiomics: images are more than pictures, they are data. *Radiology* **278**, 563-577 (2015).
29 Edmonds, C. E. & Mankoff, D. A. in *Molecular Pathology of Breast Cancer* 187-206 (Springer, 2016).
30 Altman, D. G., Vergouwe, Y., Royston, P. & Moons, K. G. Prognosis and prognostic research: validating a prognostic model. *Bmj* **338**, b605 (2009).
31 Dunn, L. & DeMichele, A. Genomic predictors of outcome and treatment response in breast cancer. *Molecular diagnosis & therapy* **13**, 73-90 (2009).
32 Schmidt, K. *et al.* Errors introduced by tissue heterogeneity in estimation of local cerebral glucose utilization with current kinetic models of the [18F] fluorodeoxyglucose method. *Journal of Cerebral Blood Flow & Metabolism* **12**, 823-834 (1992).
33 Chitalia, R. D. & Kontos, D. Role of texture analysis in breast MRI as a cancer biomarker: A review. *Journal of Magnetic Resonance Imaging* **49**, 927-938 (2019).
34 Chitalia, R. D. & Kontos, D. Role of texture analysis in breast MRI as a cancer biomarker: A review. *Journal of Magnetic Resonance Imaging* (2018).
35 Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *cell* **144**, 646-674 (2011).
36 Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747 (2000).
37 Sorlie, T. *et al.* Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein P Lonning, Borresen-Dale AL Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA* **98**, 10869-10874 (2001).
38 Perez, E. A. Breast cancer management: opportunities and barriers to an individualized approach. *The oncologist* **16**, 20-22 (2011).
39 Dagogo-Jack, I. & Shaw, A. T. Tumour heterogeneity and resistance to cancer therapies. *Nature Reviews Clinical Oncology* (2017).
40 Yates, L. R. *et al.* Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nature medicine* **21**, 751 (2015).
41 Harbst, K. *et al.* Multiregion whole-exome sequencing uncovers the genetic evolution and mutational heterogeneity of early-stage metastatic melanoma. *Cancer research* **76**, 4765-4774 (2016).

42 Piotrowska, Z. *et al.* Heterogeneity underlies the emergence of EGFRT790 wild-type clones following treatment of T790M-positive cancers with a third-generation EGFR inhibitor. *Cancer discovery* **5**, 713-722 (2015).
43 Oxnard, G. R. *et al.* Association between plasma genotyping and outcomes of treatment with osimertinib (AZD9291) in advanced non–small-cell lung cancer. *Journal of Clinical Oncology* **34**, 3375 (2016).
44 Chabon, J. J. *et al.* Circulating tumour DNA profiling reveals heterogeneity of EGFR inhibitor resistance mechanisms in lung cancer patients. *Nature communications* **7**, 1-15 (2016).
45 Suda, K. *et al.* Heterogeneity in resistance mechanisms causes shorter duration of epidermal growth factor receptor kinase inhibitor treatment in lung cancer. *Lung cancer* **91**, 36-40 (2016).
46 Koboldt, D. *et al.* Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61-70 (2012).
47 Beca, F. & Polyak, K. Intratumor heterogeneity in breast cancer. *Novel biomarkers in the continuum of breast cancer*, 169-189 (2016).
48 Ali, H. R. *et al.* Genome-driven integrated classification of breast cancer validated in over 7,500 samples. *Genome biology* **15**, 1-14 (2014).
49 Navin, N. *et al.* Inferring tumor progression from genomic heterogeneity. *Genome research* **20**, 68-80 (2010).
50 Wang, Y. *et al.* Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* **512**, 155-160 (2014).
51 Sengupta, D. & Pratx, G. Imaging metabolic heterogeneity in cancer. *Molecular cancer* **15**, 4 (2016).
52 Paradiso, A. *et al.* Heterogeneity of intratumour proliferative activity in primary breast cancer: biological and clinical aspects. *European Journal of Cancer* **31**, 911-916 (1995).
53 Waks, A. G. & Winer, E. P. Breast cancer treatment: a review. *Jama* **321**, 288-300 (2019).
54 Valabrega, G., Montemurro, F. & Aglietta, M. Trastuzumab: mechanism of action, resistance and future perspectives in HER2-overexpressing breast cancer. *Annals of oncology* **18**, 977-984 (2007).
55 Higgins, M. J. & Baselga, J. Targeted therapies for breast cancer. *The Journal of clinical investigation* **121**, 3797-3803 (2011).
56 Alvarez, R. H., Valero, V. & Hortobagyi, G. N. Emerging targeted therapies for breast cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **28**, 3366-3379 (2010).
57 Tse, G. M. *et al.* Magnetic resonance imaging of breast lesions--a pathologic correlation. *Breast cancer research and treatment* **103**, 1-10, doi:10.1007/s10549-006-9352-3 (2007).
58 Hylton, N. Dynamic contrast-enhanced magnetic resonance imaging as an imaging biomarker. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **24**, 3293-3298 (2006).
59 O'Connor, J. P. *et al.* Quantitative imaging biomarkers in the clinical development of targeted therapeutics: current and future perspectives. *The lancet oncology* **9**, 766-776 (2008).
60 Padhani, A. R. Dynamic contrast-enhanced MRI in clinical oncology: current status and future directions. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine* **16**, 407-422 (2002).
61 Quon, A. & Gambhir, S. S. FDG-PET and beyond: molecular breast cancer imaging. *Journal of clinical oncology* **23**, 1664-1673 (2005).

62 Liberti, M. V. & Locasale, J. W. The Warburg effect: how does it benefit cancer cells? *Trends in biochemical sciences* **41**, 211-218 (2016).
63 Jain, R., Kasturi, R. & Schunck, B. G. *Machine vision*. Vol. 5 (McGraw-Hill New York, 1995).
64 Jain, A., Ross, A. & Prabhakar, S. in *Image Processing, 2001. Proceedings. 2001 International Conference on.* 282-285 (IEEE).
65 Chen, K., Wei, H., Hennebert, J., Ingold, R. & Liwicki, M. in *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on.* 488-493 (IEEE).
66 Gastounioti, A., Conant, E. F. & Kontos, D. Beyond breast density: a review on the advancing role of parenchymal texture analysis in breast cancer risk assessment. *Breast Cancer Research* **18**, 91 (2016).
67 Lambin, P. *et al.* Radiomics: extracting more information from medical images using advanced feature analysis. *European journal of cancer* **48**, 441-446 (2012).
68 Kumar, V. *et al.* Radiomics: the process and the challenges. *Magnetic resonance imaging* **30**, 1234-1248 (2012).
69 Braman, N. M. *et al.* Intratumoral and peritumoral radiomics for the pretreatment prediction of pathological complete response to neoadjuvant chemotherapy based on breast DCE-MRI. *Breast Cancer Research* **19**, 57 (2017).
70 Wang, J. *et al.* Identifying triple-negative breast cancer using background parenchymal enhancement heterogeneity on dynamic contrast-enhanced MRI: a pilot radiomics study. *PLoS One* **10**, e0143308 (2015).
71 Haralick, R. M. & Shanmugam, K. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, 610-621 (1973).
72 Galloway, M. M. Texture analysis using grey level run lengths. *NASA STI/Recon Technical Report N* **75** (1974).
73 Tang, X. Texture information in run-length matrices. *IEEE transactions on image processing* **7**, 1602-1609 (1998).
74 Chu, A., Sehgal, C. M. & Greenleaf, J. F. Use of gray value distribution of run lengths for texture analysis. *Pattern Recognition Letters* **11**, 415-419 (1990).
75 Ojala, T., Pietikainen, M. & Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence* **24**, 971-987 (2002).
76 Porter, R. & Canagarajah, N. Robust rotation-invariant texture classification: wavelet, Gabor filter and GMRF based schemes. *IEE Proceedings-Vision, Image and Signal Processing* **144**, 180-188 (1997).
77 Srinivasan, G. & Shobha, G. in *Proceedings of world academy of science, engineering and technology.* 1264-1269.
78 Livens, S., Scheunders, P., Van de Wouwer, G. & Van Dyck, D. Wavelets for texture analysis, an overview. (1997).
79 Holli, K. K. *et al.* Texture analysis of MR images of patients with mild traumatic brain injury. *BMC medical imaging* **10**, 8 (2010).
80 Hackmack, K. *et al.* Multi-scale classification of disease using structural MRI and wavelet transform. *Neuroimage* **62**, 48-58 (2012).
81 Mayerhoefer, M. E. *et al.* Texture analysis for tissue discrimination on T1-weighted MR images of the knee joint in a multicenter study: transferability of texture features and comparison of feature selection methods and classifiers. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine* **22**, 674-680 (2005).

82 Michopoulou, S., Costaridou, L., Vlychou, M., Speller, R. & Todd-Pokropek, A. Texture-based quantification of lumbar intervertebral disc degeneration from conventional T2-weighted MRI. *Acta Radiologica* **52**, 91-98 (2011).
83 Vignati, A. *et al.* Texture features on T2-weighted magnetic resonance imaging: new potential biomarkers for prostate cancer aggressiveness. *Physics in Medicine & Biology* **60**, 2685 (2015).
84 Assefa, D. *et al.* Robust texture features for response monitoring of glioblastoma multiforme on-weighted and-FLAIR MR images: A preliminary investigation in terms of identification and segmentation. *Medical physics* **37**, 1722-1736 (2010).
85 Torheim, T. *et al.* Classification of dynamic contrast enhanced MR images of cervical cancers using texture analysis and support vector machines. *IEEE transactions on medical imaging* **33**, 1648-1656 (2014).
86 Harrison, L. *et al.* Texture analysis on MRI images of non-Hodgkin lymphoma. *Computers in biology and medicine* **38**, 519-524 (2008).
87 Zhang, Y. MRI texture analysis in multiple sclerosis. *Journal of Biomedical Imaging* **2012**, 2 (2012).
88 Morrow, M., Waters, J. & Morris, E. MRI for breast cancer screening, diagnosis, and treatment. *The Lancet* **378**, 1804-1811 (2011).
89 Cho, G. Y. *et al.* Evaluation of breast cancer using intravoxel incoherent motion (IVIM) histogram analysis: comparison with malignant status, histological subtype, and molecular prognostic factors. *European radiology* **26**, 2547-2558 (2016).
90 Gibbs, P. & Turnbull, L. W. Textural analysis of contrast-enhanced MR images of the breast. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* **50**, 92-98 (2003).
91 Nagarajan, M. B. *et al.* Classification of small lesions in breast MRI: evaluating the role of dynamically extracted texture features through feature selection. *Journal of medical and biological engineering* **33** (2013).
92 Nie, K. *et al.* Quantitative analysis of lesion morphology and texture features for diagnostic prediction in breast MRI. *Academic radiology* **15**, 1513-1525 (2008).
93 Bhooshan, N. *et al.* Cancerous breast lesions on dynamic contrast-enhanced MR images: computerized characterization for image-based prognostic markers. *Radiology* **254**, 680-690 (2010).
94 Bhooshan, N. *et al.* Combined use of T2-weighted MRI and T1-weighted dynamic contrast–enhanced MRI in the automated analysis of breast lesions. *Magnetic resonance in medicine* **66**, 555-564 (2011).
95 Holli, K. *et al.* Characterization of breast cancer types by texture analysis of magnetic resonance images. *Academic radiology* **17**, 135-141 (2010).
96 Chen, W., Giger, M. L., Li, H., Bick, U. & Newstead, G. M. Volumetric texture analysis of breast lesions on contrast-enhanced magnetic resonance images. *Magnetic resonance in medicine* **58**, 562-571 (2007).
97 Issa, B., Buckley, D. L. & Turnbull, L. W. Heterogeneity analysis of Gd-DTPA uptake: improvement in breast lesion differentiation. *Journal of computer assisted tomography* **23**, 615-621 (1999).
98 Waugh, S. *et al.* Magnetic resonance imaging texture analysis classification of primary breast cancer. *European radiology* **26**, 322-330 (2016).
99 Holli-Helenius, K. *et al.* MRI texture analysis in differentiating luminal A and luminal B breast cancer molecular subtypes-a feasibility study. *BMC medical imaging* **17**, 69 (2017).

100 Sutton, E. J. *et al.* Breast cancer subtype intertumor heterogeneity: MRI-based features predict results of a genomic assay. *Journal of Magnetic Resonance Imaging* **42**, 1398-1406 (2015).
101 Li, H. *et al.* MR imaging radiomics signatures for predicting the risk of breast cancer recurrence as given by research versions of MammaPrint, Oncotype DX, and PAM50 gene assays. *Radiology* **281**, 382-391 (2016).
102 Wu, J. *et al.* Identifying relations between imaging phenotypes and molecular subtypes of breast cancer: model discovery and external validation. *Journal of Magnetic Resonance Imaging* **46**, 1017-1027 (2017).
103 King, V. *et al.* Background parenchymal enhancement at breast MR imaging and breast cancer risk. *Radiology* **260**, 50-60 (2011).
104 Chamming's, F. *et al.* Features from computerized texture analysis of breast cancers at pretreatment MR imaging are associated with response to neoadjuvant chemotherapy. *Radiology* **286**, 412-420 (2017).
105 Kim, J.-H. *et al.* Breast cancer heterogeneity: MR imaging texture analysis and survival outcomes. *Radiology* **282**, 665-675 (2016).
106 Pickles, M. D., Lowry, M. & Gibbs, P. Pretreatment prognostic value of dynamic contrast-enhanced magnetic resonance imaging vascular, texture, shape, and size parameters compared with traditional survival indicators obtained from locally advanced breast cancer patients. *Investigative radiology* **51**, 177-185 (2016).
107 Park, H. *et al.* Radiomics Signature on Magnetic Resonance Imaging: Association with Disease-Free Survival in Patients with Invasive Breast Cancer. *Clinical Cancer Research*, clincanres. 3783.2017 (2018).
108 Mahrooghy, M. *et al.* in *International Conference on Medical Image Computing and Computer-Assisted Intervention.* 295-302 (Springer).
109 Mahrooghy, M. *et al.* Pharmacokinetic tumor heterogeneity as a prognostic biomarker for classifying breast cancer recurrence risk. *IEEE Transactions on Biomedical Engineering* **62**, 1585-1594 (2015).
110 Teruel, J. *et al.* Dynamic contrast-enhanced MRI texture analysis for pretreatment prediction of clinical and pathological response to neoadjuvant chemotherapy in patients with locally advanced breast cancer. *NMR in biomedicine* **27**, 887-896 (2014).
111 Thibault, G. *et al.* DCE-MRI Texture Features for Early Prediction of Breast Cancer Therapy Response. *Tomography: a journal for imaging research* **3**, 23 (2017).
112 Golden, D. I., Lipson, J. A., Telli, M. L., Ford, J. M. & Rubin, D. L. Dynamic contrast-enhanced MRI-based biomarkers of therapeutic response in triple-negative breast cancer. *Journal of the American Medical Informatics Association* **20**, 1059-1066 (2013).
113 Parikh, J. *et al.* Changes in primary breast cancer heterogeneity may augment midtreatment MR imaging assessment of response to neoadjuvant chemotherapy. *Radiology* **272**, 100-112 (2014).
114 Michoux, N. *et al.* Texture analysis on MR images helps predicting non-response to NAC in breast cancer. *BMC cancer* **15**, 574 (2015).
115 Wu, J., Gong, G., Cui, Y. & Li, R. Intratumor partitioning and texture analysis of dynamic contrast-enhanced (DCE)-MRI identifies relevant tumor subregions to predict pathological response of breast cancer to neoadjuvant chemotherapy. *Journal of Magnetic Resonance Imaging* **44**, 1107-1115 (2016).
116 Ashraf, A. *et al.* Breast DCE-MRI kinetic heterogeneity tumor markers: preliminary associations with neoadjuvant chemotherapy response. *Translational oncology* **8**, 154-162 (2015).

117 Ahmed, A., Gibbs, P., Pickles, M. & Turnbull, L. Texture analysis in assessment and prediction of chemotherapy response in breast cancer. *Journal of Magnetic Resonance Imaging* **38**, 89-101 (2013).
118 El Adoui, M., Drisis, S., Larhmam, M., Lemort, M. & Benjelloun, M. Breast cancer heterogeneity analysis as index of response to treatment using MRI images: a review. *Imaging Med* **9**, 109-119 (2017).
119 Johansen, R. *et al.* Predicting survival and early clinical response to primary chemotherapy for patients with locally advanced breast cancer using DCE-MRI. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine* **29**, 1300-1307 (2009).
120 Padhani, A. R. *et al.* Prediction of clinicopathologic response of breast cancer to primary chemotherapy at contrast-enhanced MR imaging: initial clinical results. *Radiology* **239**, 361-374 (2006).
121 Burnside, E. S. *et al.* Using computer-extracted image phenotypes from tumors on breast magnetic resonance imaging to predict breast cancer pathologic stage. *Cancer* **122**, 748-757 (2016).
122 Agner, S. C. *et al.* Computerized image analysis for identifying triple-negative breast cancers and differentiating them from other molecular subtypes of breast cancer on dynamic contrast-enhanced MR images: a feasibility study. *Radiology* **272**, 91-99 (2014).
123 Zhu, Y. *et al.* Deciphering genomic underpinnings of quantitative MRI-based radiomic phenotypes of invasive breast carcinoma. *Scientific reports* **5**, 17787 (2015).
124 Sollini, M. *et al.* PET/CT radiomics in breast cancer: mind the step. *Methods* **188**, 122-132 (2021).
125 Ou, X. *et al.* Ability of 18F-FDG PET/CT radiomic features to distinguish breast carcinoma from breast lymphoma. *Contrast media & molecular imaging* **2019** (2019).
126 Lemarignier, C. *et al.* Correlation between tumour characteristics, SUV measurements, metabolic tumour volume, TLG and textural features assessed with F-FDG PET in a large cohort of oestrogen receptor-positive breast cancer patients. *European Journal of Nuclear Medicine & Molecular Imaging* **44** (2017).
127 Groheux, D. *et al.* Do clinical, histological or immunohistochemical primary tumour characteristics translate into different 18 F-FDG PET/CT volumetric and heterogeneity features in stage II/III breast cancer? *European journal of nuclear medicine and molecular imaging* **42**, 1682-1691 (2015).
128 Huang, S.-y. *et al.* Exploration of PET and MRI radiomic features for decoding breast cancer phenotypes and prognosis. *NPJ breast cancer* **4**, 1-13 (2018).
129 Ha, S., Park, S., Bang, J.-I., Kim, E.-K. & Lee, H.-Y. Metabolic radiomics for pretreatment 18 F-FDG PET/CT to characterize locally advanced breast cancer: histopathologic characteristics, response to neoadjuvant chemotherapy, and prognosis. *Scientific reports* **7**, 1-11 (2017).
130 Antunovic, L. *et al.* [18 F] FDG PET/CT features for the molecular characterization of primary breast tumors. *European journal of nuclear medicine and molecular imaging* **44**, 1945-1954 (2017).
131 Moscoso, A. *et al.* Texture analysis of high-resolution dedicated breast 18 F-FDG PET images correlates with immunohistochemical factors and subtype of breast cancer. *European journal of nuclear medicine and molecular imaging* **45**, 196-206 (2018).
132 Molina-García, D. *et al.* Intratumoral heterogeneity in 18 F-FDG PET/CT by textural analysis in breast cancer as a predictive and prognostic subrogate. *Annals of nuclear medicine* **32**, 379-388 (2018).

133 Acar, E., Turgut, B., Yiğit, S. & Kaya, G. Comparison of the volumetric and radiomics findings of 18F-FDG PET/CT images with immunohistochemical prognostic factors in local/locally advanced breast cancer. *Nuclear medicine communications* **40**, 764-772 (2019).
134 Antunovic, L. *et al.* PET/CT radiomics in breast cancer: promising tool for prediction of pathological response to neoadjuvant chemotherapy. *European journal of nuclear medicine and molecular imaging* **46**, 1468-1477 (2019).
135 Lee, H. *et al.* Predicting response to neoadjuvant chemotherapy in patients with breast cancer: combined statistical modeling using clinicopathological factors and FDG PET/CT texture parameters. *Clinical nuclear medicine* **44**, 21-29 (2019).
136 Yoon, H. J., Kim, Y., Chung, J. & Kim, B. S. Predicting neo-adjuvant chemotherapy response and progression-free survival of locally advanced breast cancer using textural features of intratumoral heterogeneity on F-18 FDG PET/CT and diffusion-weighted MR imaging. *The breast journal* **25**, 373-380 (2019).
137 Mendel, K. R., Li, H. & Giger, M. L. in *Medical Imaging 2016: Computer-Aided Diagnosis.* 97851W (International Society for Optics and Photonics).
138 Tagliafico, A. *et al.* Quantitative evaluation of background parenchymal enhancement (BPE) on breast MRI. A feasibility study with a semi-automatic and automatic software compared to observer-based scores. *The British journal of radiology* **88**, 20150417 (2015).
139 Liao, G. J. *et al.* Background parenchymal enhancement on breast MRI: a comprehensive review. *Journal of Magnetic Resonance Imaging* **51**, 43-61 (2020).
140 Mazurowski, M. A., Zhang, J., Grimm, L. J., Yoon, S. C. & Silber, J. I. Radiogenomic analysis of breast cancer: luminal B molecular subtype is associated with enhancement dynamics at MR imaging. *Radiology* **273**, 365-372 (2014).
141 Wu, S. *et al.* Breast MRI contrast enhancement kinetics of normal parenchyma correlate with presence of breast cancer. *Breast Cancer Research* **18**, 76 (2016).
142 Singh, A., Chitalia, R. & Kontos, D. Radiogenomics in brain, breast, and lung cancer: opportunities and challenges. *Journal of Medical Imaging* **8**, 031907 (2021).
143 Mazurowski, M. A. Radiogenomics: what it is and why it is important. *Journal of the American College of Radiology* **12**, 862-866 (2015).
144 Leithner, D. *et al.* Radiomic signatures with contrast-enhanced magnetic resonance imaging for the assessment of breast cancer receptor status and molecular subtypes: initial results. *Breast Cancer Research* **21**, 1-11 (2019).
145 Grimm, L. J., Zhang, J. & Mazurowski, M. A. Computational approach to radiogenomics of breast cancer: luminal A and luminal B molecular subtypes are associated with imaging features on routine breast MRI extracted using computer vision algorithms. *Journal of Magnetic Resonance Imaging* **42**, 902-907 (2015).
146 Saha, A. *et al.* A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features. *British journal of cancer* **119**, 508-516 (2018).
147 Ashraf, A. B. *et al.* Identification of intrinsic imaging phenotypes for breast cancer tumors: preliminary associations with gene expression profiles. *Radiology* **272**, 374-384 (2014).
148 Tamez-Peña, J.-G. *et al.* Radiogenomics analysis identifies correlations of digital mammography with clinical molecular signatures in breast cancer. *PloS one* **13**, e0193871 (2018).
149 Panth, K. M. *et al.* Is there a causal relationship between genetic changes and radiomics-based image features? An in vivo preclinical experiment with doxycycline inducible GADD34 tumor cells. *Radiotherapy and Oncology* **116**, 462-466 (2015).

150 Yamamoto, S. *et al.* Breast cancer: radiogenomic biomarker reveals associations among dynamic contrast-enhanced MR imaging, long noncoding RNA, and metastasis. *Radiology* **275**, 384-392 (2015).
151 Bismeijer, T. *et al.* Radiogenomic analysis of breast cancer by linking MRI phenotypes with tumor gene expression. *Radiology* **296**, 277-287 (2020).
152 Duda, R. (John Wiley, New York, 2000).
153 Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, 289-300 (1995).
154 Waugh, S. A., Lerski, R. A., Bidaut, L. & Thompson, A. M. The influence of field strength and different clinical breast MRI protocols on the outcome of texture analysis using foam phantoms. *Medical physics* **38**, 5058-5066 (2011).
155 Lubner, M. G., Smith, A. D., Sandrasegaran, K., Sahani, D. V. & Pickhardt, P. J. CT texture analysis: definitions, applications, biologic correlates, and challenges. *RadioGraphics* **37**, 1483-1503 (2017).
156 Parekh, V. S. & Jacobs, M. A. Integrated radiomic framework for breast cancer and tumor biology using advanced machine learning and multiparametric MRI. *NPJ breast cancer* **3**, 43 (2017).
157 Newitt D , H. N., on behlaf of the I-SPY 1 Network and ACRIN 6657 Trial Team. (The Cancer Imaging Archive, 2016).
158 Hylton, N. M. *et al.* Neoadjuvant Chemotherapy for Breast Cancer: Functional Tumor Volume by MR Imaging Predicts Recurrence-free Survival-Results from the ACRIN 6657/CALGB 150007 I-SPY 1 TRIAL. *Radiology* **279**, 44-55, doi:10.1148/radiol.2015150013 (2016).
159 Clark, K. *et al.* The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *Journal of digital imaging* **26**, 1045-1057 (2013).
160 Ashraf, A. B. *et al.* A multichannel markov random field framework for tumor segmentation with an application to classification of gene expression-based breast cancer recurrence risk. *IEEE transactions on medical imaging* **32**, 637-648 (2013).
161 Sled, J. G., Zijdenbos, A. P. & Evans, A. C. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE transactions on medical imaging* **17**, 87-97 (1998).
162 Szabo, B. K., Aspelin, P., Kristoffersen Wiberg, M., Tot, T. & Bone, B. Invasive breast cancer: correlation of dynamic MR features with prognostic factors. *European radiology* **13**, 2425-2435, doi:10.1007/s00330-003-2000-y (2003).
163 Davatzikos, C. *et al.* Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. *Journal of medical imaging* **5**, 011018 (2018).
164 Friedman, J., Hastie, T. & Tibshirani, R. *The elements of statistical learning*. Vol. 1 (Springer series in statistics New York, 2001).
165 Ward Jr, J. H. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association* **58**, 236-244 (1963).
166 Monti, S., Tamayo, P., Mesirov, J. & Golub, T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning* **52**, 91-118 (2003).
167 Liu, Y., Hayes, D. N., Nobel, A. & Marron, J. Statistical significance of clustering for high-dimension, low–sample size data. *Journal of the American Statistical Association* **103**, 1281-1293 (2008).

168 Dunnwald, L. K., Rossing, M. A. & Li, C. I. Hormone receptor status, tumor characteristics, and prognosis: a prospective cohort of breast cancer patients. *Breast cancer research* **9**, R6 (2007).
169 Fisher, B., Redmond, C., Fisher, E. R. & Caplan, R. Relative worth of estrogen or progesterone receptor and pathologic characteristics of differentiation as indicators of prognosis in node negative breast cancer patients: findings from National Surgical Adjuvant Breast and Bowel Project Protocol B-06. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **6**, 1076-1087, doi:10.1200/JCO.1988.6.7.1076 (1988).
170 Dagogo-Jack, I. & Shaw, A. T. Tumour heterogeneity and resistance to cancer therapies. *Nature reviews Clinical oncology* **15**, 81 (2018).
171 Fox, E. J. & Loeb, L. A. One cell at a time. *Nature* **512**, 143 (2014).
172 Broglio, K. R. *et al.* Association of pathologic complete response to neoadjuvant therapy in HER2-positive breast cancer with long-term outcomes: a meta-analysis. *JAMA oncology* **2**, 751-760 (2016).
173 LeVasseur, N. *et al.* Impact of pathologic complete response on survival after neoadjuvant chemotherapy in early-stage breast cancer: a population-based analysis. *Journal of Cancer Research and Clinical Oncology* **146**, 529-536 (2020).
174 Thompson, A. & Moulder-Thompson, S. Neoadjuvant treatment of breast cancer. *annals of Oncology* **23**, x231-x236 (2012).
175 Yee, D. *et al.* Association of event-free and distant recurrence–free survival with individual-level pathologic complete response in neoadjuvant treatment of stages 2 and 3 breast cancer: Three-year follow-up analysis for the i-spy2 adaptively randomized clinical trial. *JAMA oncology* **6**, 1355-1362 (2020).
176 Zhou, X. *et al.* Alterations of biomarker profiles after neoadjuvant chemotherapy in breast cancer: tumor heterogeneity should be taken into consideration. *Oncotarget* **6**, 36894 (2015).
177 Caswell-Jin, J. L. *et al.* Clonal replacement and heterogeneity in breast tumors treated with neoadjuvant HER2-targeted therapy. *Nature communications* **10**, 1-12 (2019).
178 Győrffy, B. *et al.* Multigene prognostic tests in breast cancer: past, present, future. *Breast cancer research* **17**, 11 (2015).
179 Dowsett, M. *et al.* Prediction of risk of distant recurrence using the 21-gene recurrence score in node-negative and node-positive postmenopausal patients with breast cancer treated with anastrozole or tamoxifen: a TransATAC study. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **28**, 1829-1834 (2010).
180 Baker, L. *et al.* p53 mutation, deprivation and poor prognosis in primary breast cancer. *Br J Cancer* **102**, 719-726, doi:10.1038/sj.bjc.6605540 (2010).
181 Drukker, K. *et al.* Most-enhancing tumor volume by MRI radiomics predicts recurrence-free survival "early on" in neoadjuvant treatment of breast cancer. *Cancer Imaging* **18**, 12 (2018).
182 Hylton, N. M. *et al.* Locally advanced breast cancer: MR imaging for prediction of response to neoadjuvant chemotherapy--results from ACRIN 6657/I-SPY TRIAL. *Radiology* **263**, 663-672, doi:10.1148/radiol.12110748 (2012).
183 Hylton, N. M. *et al.* Neoadjuvant chemotherapy for breast cancer: functional tumor volume by MR imaging predicts recurrence-free survival—results from the ACRIN 6657/CALGB 150007 I-SPY 1 TRIAL. *Radiology* **279**, 44-55 (2016).
184 Jahani, N. *et al.* prediction of treatment Response to neoadjuvant chemotherapy for Breast cancer via early changes in tumor Heterogeneity captured by Dce-MRi Registration. *Scientific reports* **9**, 1-12 (2019).

185 Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**, 207-210, doi:10.1093/nar/30.1.207 (2002).
186 Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res* **41**, D991-995, doi:10.1093/nar/gks1193 (2013).
187 Esserman, L. J. *et al.* Chemotherapy response and recurrence-free survival in neoadjuvant breast cancer depends on biomarker profiles: results from the I-SPY 1 TRIAL (CALGB 150007/150012; ACRIN 6657). *Breast Cancer Res Treat* **132**, 1049-1062, doi:10.1007/s10549-011-1895-2 (2012).
188 Newitt, D. & Hylton, N. Single site breast DCE-MRI data and segmentations from patients undergoing neoadjuvant chemotherapy. *The Cancer Imaging Archive* **2** (2016).
189 Esserman, L. J. *et al.* Chemotherapy response and recurrence-free survival in neoadjuvant breast cancer depends on biomarker profiles: results from the I-SPY 1 TRIAL (CALGB 150007/150012; ACRIN 6657). *Breast cancer research and treatment* **132**, 1049-1062 (2012).
190 van 't Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530-536, doi:10.1038/415530a (2002).
191 Glas, A. M. *et al.* Converting a breast cancer microarray signature into a high-throughput diagnostic test. *BMC Genomics* **7**, 278, doi:10.1186/1471-2164-7-278 (2006).
192 Hu, Z. *et al.* The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics* **7**, 96, doi:10.1186/1471-2164-7-96 (2006).
193 Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol* **27**, 1160-1167, doi:10.1200/JCO.2008.18.1370 (2009).
194 Troester, M. A. *et al.* Gene expression patterns associated with p53 status in breast cancer. *BMC Cancer* **6**, 276, doi:10.1186/1471-2407-6-276 (2006).
195 Martin, N. & Maes, H. *Multivariate analysis*. (Academic press London, 1979).
196 Chitalia, R. *et al.* Imaging phenotypes of breast cancer heterogeneity in pre-operative breast Dynamic Contrast Enhanced Magnetic Resonance Imaging (DCE-MRI) scans predict 10-year recurrence. *Clinical Cancer Research*, clincanres.4067.2018, doi:10.1158/1078-0432.ccr-18-4067 (2019).
197 Rupert Jr, G. *Simultaneous statistical inference*. (Springer Science & Business Media, 2012).
198 Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences* **98**, 5116-5121 (2001).
199 Verma, S. *et al.* Overview of dynamic contrast-enhanced MRI in prostate cancer diagnosis and management. *American Journal of Roentgenology* **198**, 1277-1288 (2012).
200 Vance, G. H. *et al.* Genetic heterogeneity in HER2 testing in breast cancer: panel summary and guidelines. *Archives of pathology & laboratory medicine* **133**, 611-612 (2009).
201 Bareche, Y. *et al.* Unraveling triple-negative breast cancer tumor microenvironment heterogeneity: towards an optimized treatment approach. *JNCI: Journal of the National Cancer Institute* **112**, 708-719 (2020).
202 Seol, H. *et al.* Intratumoral heterogeneity of HER2 gene amplification in breast cancer: its clinicopathological significance. *Modern pathology* **25**, 938-948 (2012).
203 Radiology, A. C. o. ACR practice parameter for the performance of contrast-enhanced magnetic resonance imaging (MRI) of the breast [homepage on the Internet]. *https://www.acr.org/-/media/ACR/Files/Practice-Parameters/mr-contrast-breast.pdf* (c.2004 [updated 2018; cited 2021 Nov].).

204 Soret, M., Bacharach, S. L. & Buvat, I. Partial-volume effect in PET tumor imaging. *Journal of Nuclear Medicine* **48**, 932-945 (2007).
205 Katiyar, P. *et al.* Spectral clustering predicts tumor tissue heterogeneity using dynamic 18F-FDG PET: a complement to the standard compartmental modeling approach. *Journal of Nuclear Medicine* **58**, 651-657 (2017).
206 Kass, M., Witkin, A. & Terzopoulos, D. Snakes: Active contour models. *International journal of computer vision* **1**, 321-331 (1988).
207 Cootes, T. F., Taylor, C. J., Cooper, D. H. & Graham, J. Active shape models-their training and application. *Computer vision and image understanding* **61**, 38-59 (1995).
208 Tauber, C., Batatia, H. & Ayache, A. Quasi-automatic initialization for parametric active contours. *Pattern Recognition Letters* **31**, 83-90 (2010).
209 Wang, J., Kong, J., Lu, Y., Qi, M. & Zhang, B. A modified FCM algorithm for MRI brain image segmentation using both local and non-local spatial constraints. *Computerized medical imaging and graphics* **32**, 685-698 (2008).
210 Muzi, M. *et al.* Kinetic analysis of 3′-deoxy-3′-fluorothymidine PET studies: validation studies in patients with lung cancer. *Journal of Nuclear Medicine* **46**, 274-282 (2005).
211 Strydhorst, J. & Buvat, I. Redesign of the GATE PET coincidence sorter. *Physics in Medicine & Biology* **61**, N522 (2016).
212 Karp, J. S. *et al.* PennPET Explorer: Design and Preliminary Performance of a Whole-body Imager. *Journal of Nuclear Medicine*, jnumed. 119.229997 (2019).
213 Viswanath, V. *et al.* Quantifying bias and precision of kinetic parameter estimation on the PennPET Explorer, a long axial field-of-view scanner. *IEEE Transactions on Radiation and Plasma Medical Sciences* (2020).
214 Müller, M. Dynamic time warping. *Information retrieval for music and motion*, 69-84 (2007).
215 Senin, P. Dynamic time warping algorithm review. *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA* **855**, 40 (2008).
216 Von Luxburg, U. A tutorial on spectral clustering. *Statistics and computing* **17**, 395-416 (2007).
217 Chen, J. L., Gunn, S. R., Nixon, M. S. & Gunn, R. N. in *Biennial International Conference on Information Processing in Medical Imaging.* 468-474 (Springer).
218 Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **39**, 1-22 (1977).
219 Muzi, M. *et al.* Quantitative assessment of dynamic PET imaging data in cancer imaging. *Magnetic resonance imaging* **30**, 1203-1215 (2012).
220 Eloyan, A., Yue, M. S. & Khachatryan, D. Tumor heterogeneity estimation for radiomics in cancer. *Statistics in medicine* **39**, 4704-4723 (2020).
221 Dice, L. R. Measures of the amount of ecologic association between species. *Ecology* **26**, 297-302 (1945).
222 McDonald, E. S., Mankoff, D. A. & Mach, R. H. Novel strategies for breast cancer imaging: new imaging agents to guide treatment. *J Nucl Med* **57**, 69S-74S (2016).
223 Groheux, D. *et al.* 18F-FDG PET/CT for staging and restaging of breast cancer. *J Nucl Med* **57**, 17S-26S (2016).
224 Aerts, H. J. *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications* **5**, 4006 (2014).
225 Eary, J. F., O'Sullivan, F., O'Sullivan, J. & Conrad, E. U. Spatial heterogeneity in sarcoma 18F-FDG uptake as a predictor of patient outcome. *Journal of Nuclear Medicine* **49**, 1973-1979 (2008).

226 Stoyanova, R. *et al.* Mapping tumor hypoxia in vivo using pattern recognition of dynamic contrast-enhanced MRI data. *Translational oncology* **5**, 437-IN432 (2012).
227 Cherezov, D. *et al.* Revealing tumor habitats from texture heterogeneity analysis for classification of lung cancer malignancy and aggressiveness. *Scientific reports* **9**, 1-9 (2019).
228 Chang, Y.-C. C. *et al.* Delineation of tumor habitats based on dynamic contrast enhanced MRI. *Scientific reports* **7**, 9746 (2017).
229 Jardim-Perassi, B. V., Martinez, G. & Gillies, R. Habitat imaging of tumor evolution by magnetic resonance imaging (MRI). *Radiomics and Radiogenomics: Technical Basis and Clinical Applications*, 115 (2019).
230 Mankoff, D. A. *et al.* Changes in blood flow and metabolism in locally advanced breast cancer treated with neoadjuvant chemotherapy. *Journal of Nuclear Medicine* **44**, 1806-1814 (2003).
231 Dunnwald, L. K. *et al.* Tumor metabolism and blood flow changes by positron emission tomography: relation to survival in patients treated with neoadjuvant chemotherapy for locally advanced breast cancer. *Journal of Clinical Oncology* **26**, 4449 (2008).
232 Bhattacharyya, A. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.* **35**, 99-109 (1943).
233 Chan, S. R. *et al.* Longitudinal noninvasive imaging of progesterone receptor as a predictive biomarker of tumor responsiveness to estrogen deprivation therapy. *Clinical Cancer Research* **21**, 1063-1070 (2015).
234 O'Sullivan, F. Locally constrained mixture representation of dynamic imaging data from PET and MR studies. *Biostatistics* **7**, 318-338 (2006).
235 O'Sullivan, F. *et al.* Voxel-level mapping of tracer kinetics in PET studies: a statistical approach emphasizing tissue life tables. *The annals of applied statistics* **8**, 1065 (2014).
236 Chitalia R., V. V., Pantel AR, Peterson LM, Gastounioti A., Cohen EA, Muzi M., Karp J., Mankoff DA., Kontos D. Functional 4-D Clustering for characterizing intratumor heterogeneity in dynamic imaging: Evaluation in FDG PET as a prognostic biomarker for breast cancer. *Eur J Nucl Med Mol Imaging* (2021).
237 Popescu, L. M., Matej, S. & Lewitt, R. M. in *IEEE Symposium Conference Record Nuclear Science 2004.* 3536-3540 (IEEE).
238 Heckler, C. E. (Taylor & Francis, 2005).
239 van Seijen, M. *et al.* Ductal carcinoma in situ: to treat or not to treat, that is the question. *British journal of cancer* **121**, 285-292 (2019).
240 Cowell, C. F. *et al.* Progression from ductal carcinoma in situ to invasive breast cancer: revisited. *Molecular oncology* **7**, 859-869 (2013).
241 Tustison, N. J. *et al.* N4ITK: improved N3 bias correction. *IEEE transactions on medical imaging* **29**, 1310-1320 (2010).
242 Luo, J. *et al.* Ductal carcinoma in situ: quantitative preoperative breast MR imaging features associated with recurrence after treatment. *Radiology* **285**, 788-797 (2017).
243 Kapp, A. V. & Tibshirani, R. Are clusters found in one dataset present in another dataset? *Biostatistics* **8**, 9-31 (2007).
244 Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118-127 (2007).
245 Fortin, J.-P. *et al.* Harmonization of multi-site diffusion tensor imaging data. *Neuroimage* **161**, 149-170 (2017).
246 Horng, H. e. a. Generalized ComBat harmonization methods for radiomic features with multi-modal distributions and multiple batch effects. *Nature Scientific Reports*

247 Zeiler, M. D. & Fergus, R. in *European conference on computer vision.* 818-833 (Springer).
248 Chang, H. Y. *et al.* Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proc Natl Acad Sci U S A* **102**, 3738-3743, doi:10.1073/pnas.0409462102 (2005).
249 Gendoo, D. M. *et al.* Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer. *Bioinformatics* **32**, 1097-1099, doi:10.1093/bioinformatics/btv693 (2016).
250 Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* **98**, 5116-5121, doi:10.1073/pnas.091062498 (2001).
251 Sørlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A* **98**, 10869-10874, doi:10.1073/pnas.191367098 (2001).
252 Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882-883, doi:10.1093/bioinformatics/bts034 (2012).
253 van de Vijver, M. J. *et al.* A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* **347**, 1999-2009, doi:10.1056/NEJMoa021967 (2002).